Computers and Chemical Engineering 178 (2023) 108357

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/cace

From partial data to out-of-sample parameter and observation estimation with diffusion maps and geometric harmonics

Eleni D. Koronaki [a,b], Nikolaos Evangelou [c], Yorgos M. Psarellis [c], Andreas G. Boudouvis [b], Ioannis G. Kevrekidis [c,*]

[a] Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg, 29 John F. Kennedy Avenue 1855, Luxembourg
[b] School of Chemical Engineering, National Technical University of Athens, 9 Heroon Polytechniou str., Zographos Campus, 15780, Attiki, Greece
[c] Department of Chemical and Biomolecular Engineering, Whiting School of Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA

ARTICLE INFO

Keywords:
Diffusion maps
Nonlinear manifold learning
Chemical vapor deposition
Gappy POD
Geometric harmonics

ABSTRACT

A data-driven framework is presented, that enables the prediction of quantities, either observations or parameters, given sufficient partial data. The framework is illustrated via a computational model of the deposition of Cu in a Chemical Vapor Deposition (CVD) reactor, where the reactor pressure, the deposition temperature and feed mass flow rate are important process parameters that determine the outcome of the process. The sampled observations are high-dimensional vectors containing the outputs of a detailed CFD steady-state model of the process, i.e. the values of velocity, pressure, temperature, and species mass fractions at each point in the discretization. A machine learning workflow is presented, able to predict out-of-sample (a) observations (e.g. mass fraction in the reactor), given process parameters (e.g. inlet temperature); (b) process parameters, given observation data; and (c) partial observations (e.g. temperature in the reactor), given other partial observations (e.g. mass fraction in the reactor). The proposed workflow relies on two manifold learning schemes: Diffusion Maps and the associated Geometric Harmonics. Diffusion Maps are used for discovering a reduced representation of the available data, and Geometric Harmonics for extending functions defined on the discovered manifold. In our work a special use case of Geometric Harmonics is formulated and implemented, which we call Double Diffusion Maps, to map from the reduced representation back to (partial) observations and process parameters. A comparison of our manifold learning scheme to the traditional Gappy-POD approach is provided: ours can be thought of as a "Gappy DMAPs" approach. The presented methodology is easily transferable to application domains beyond reactor engineering.

1. Introduction

Since nonlinear manifold learning methods were introduced (Balasubramanian et al., 2002; Roweis and Saul, 2000; Coifman and Lafon, 2006a; Nadler et al., 2006; Coifman et al., 2008), a new route was carved for the parsimonious description of data derived from models of nonlinear applications.

The main premise of reduced order modeling methodologies is that state observables often live in low-dimensional manifolds, despite their apparent high dimensionality. Nonlinear manifold learning methodologies identify an intrinsic parametrization of the manifold that describes the data (Xing et al., 2016; Dsilva et al., 2018; Holiday et al., 2019; Xue et al., 2013; Bhattacharjee and Matouš, 2016) and, when coupled with appropriate mappings between the reduced description and the high-dimensional ambient space, they can be used for interpolation and regression (Giovanis and Shields, 2020; Evangelou et al., 2023, 2022; Chiavazzo et al., 2014).

Here, we demonstrate how the mapping between the ambient and the reduced space, determined with Diffusion Maps (DMAPs), can be used not only to enable efficient prediction of outputs (in our case, observations) given new inputs (in our case, process parameters), or the inputs that correspond to a new output, but also in the spirit of static and dynamic observers (Kazantzis and Kravaris, 1998; Luenberger, 1964; Cassez and Tripakis, 2008; Park et al., 2002): for the prediction of all or only part of the variables or parameters, given partial information. To achieve that, DMAPs is implemented in conjunction with a special use case of Geometric Harmonics interpolation (Coifman and Lafon, 2006b; Chiavazzo et al., 2014; Evangelou et al., 2023). The latter is implemented not only as a means of mapping between the reduced

* Corresponding author.
E-mail address: yannisk@jhu.edu (I.G. Kevrekidis).

https://doi.org/10.1016/j.compchemeng.2023.108357
Received 5 January 2023; Received in revised form 12 July 2023; Accepted 15 July 2023
Available online 20 July 2023

and ambient space, but also as a regression tool between the input and output space.

The goal is to reconstruct variables that are inaccessible, due to technical considerations pertinent to process stability and product quality. This is particularly important in the context of dynamical systems with process control as the ultimate goal (Kazantzis et al., 2005; Alonso et al., 2004b; Chiu and Christofides, 2000; Duan and Kravaris, 2020; Patel et al., 2021; Alhajeri et al., 2021; Khatibi et al., 2021). In this work, our methodologies are implemented for Chemical Vapor Deposition (CVD), a process where *in situ* sensors are scarce, and therefore critical process variables that influence product quality are inferred by *ex situ* measurements. Akiki et al. (2020), Gleason (2020), Desenfant et al. (2021), Nishinaka et al. (2021), Koronaki et al. (2014, 2016), Psarellis et al. (2018), Papavasileiou et al. (2022, 2023), Gkinis et al. (2019), Koronaki et al. (2019), Deshpande et al. (2022).

The proposed methodology is reminiscent of the so-called Gappy POD method, proposed by Everson and Sirovich (Everson and Sirovich, 1995), as an extension of the Proper Orthogonal Decomposition (POD) method, that accounts for partly known (hence "gappy") data. According to Gappy POD, it is possible to accurately reconstruct a vector that is only partially known, provided that it is spanned by a previously defined basis of POD vectors (Willcox, 2006; Xing et al., 2022; Jo et al., 2019). In the case where the data live in a *curved manifold*, the size of the POD basis required for accurate reconstruction of the data is expected to be high, since several hyperplanes are necessary to describe it; this will be discussed briefly in a subsequent section. This drawback is addressed with DMAPs, which typically require less coordinates than its linear counterpart, to accurately capture the data variance (Martin-Linares et al., 2023).

The remainder of the paper is organized as follows: the main concepts pertaining to DMAPs and Geometric Harmonics are presented, followed by Double Diffusion Maps (Double DMAPs), which is the particular implementation of the latter necessary for interpolation. The Gappy POD method is then summarized for completeness. The CVD reactor used in this work as a case study is then briefly described, as well as some details about the CFD model which generates the data. The results of the proposed workflow are then presented and compared to the Gappy POD algorithm, followed by our conclusions.

2. Diffusion maps

Diffusion Maps (Coifman and Lafon, 2006a; Nadler et al., 2006; Coifman et al., 2008) is a framework based upon diffusion processes for finding meaningful geometric descriptions of data sets, even when the underlying geometry of the data is complex, nonlinear and corrupted by (relatively low level) noise. The method is based on the construction of a Markov transition probability matrix, corresponding to a random walk on a graph, whose vertices are the data points, with transition probabilities being the local similarities between data points. The first few eigenvectors of the sparse Markov matrix are used as data-driven coordinates that provide a reparametrization of the data.

To construct a low-dimensional embedding for a data set $\mathbf{X}$ of N individual points (represented as d-dimensional real vectors $x_1, \dots, x_N$), $\mathbf{X} \in \mathbb{R}^{N\times d}$,a similarity measure between each pair of vectors x_i, x_j is computed. The standard Euclidean distance is typically used for this purpose. By using this similarity measure, an affinity matrix is constructed. A popular choice is the Gaussian kernel

$$w(i,j) = \exp\left[-\left(\frac{\|x_i - x_j\|}{\epsilon}\right)^2\right]$$

where ϵ is a hyperparameter that quantifies the kernel's bandwidth. To recover a parametrization insensitive to the sampling density, the normalization

$$\widetilde{\mathbf{W}} = \mathbf{P}^{-1}\mathbf{W}\mathbf{P}^{-1}$$

is performed, where $P_{ii} = \sum_{j=1}^{N} W_{ij}$ and W_{ij} the elements of the matrix $\mathbf{W}$. A second normalization applied on $\widetilde{\mathbf{W}}$,

$$\mathbf{K} = \mathbf{D}^{-1}\widetilde{\mathbf{W}}$$

gives a $N \times N$ Markov matrix $\mathbf{K}$; here $\mathbf{D}$ is a diagonal matrix, collecting the row sums of matrix $\widetilde{\mathbf{W}}$. The stochastic matrix $\mathbf{K}$ has a set of real eigenvalues $1 = \lambda_1 \geq \dots \geq \lambda_N$ with corresponding eigenvectors ϕ_i.

To check if dimensionality reduction can be achieved, the number of retained eigenvectors has to be appropriately truncated. In practice, it is useful to consider that not all obtained eigenvectors parametrize independent directions, but rather most of them can be considered as *spanning the same directions with different frequencies*. Eigenvectors that parametrize the same directions in this context are called *harmonics* and the ones that parametrize independent directions *non-harmonics*. Therefore, a minimal representation of the DMAPs space is made possible by carefully selecting the leading non-harmonic coordinates, which do not necessarily correspond to the most dominant eigenmodes of the Markov matrix. This is a stark difference between DMAPs and its linear counterpart, POD (also known as Principal Components Analysis, PCA), where the dominant modes are retained for the truncated representation of the data. If the number of the *non-harmonic* eigenvectors is less than the number of the ambient space dimensions then model (variable) reduction is achieved.

An algorithm for identifying the non-harmonic eigenvectors is presented in (Dsilva et al., 2018), based on local linear regression. In a nutshell, a local linear function is used in order to fit the DMAP coordinate ϕ_k as a local linear function, f, of the previous vectors ($\widetilde{\boldsymbol{\Phi}}_{k-1} = [\phi_1, \phi_2, \dots, \phi_{k-1}]$). If ϕ_k can be accurately expressed as function of the other DMAP coordinates over the data, then it does not represent a new direction on the dataset and is omitted for dimensionality reduction. On the contrary if ϕ_k cannot be expressed as a function of the previous eigenvectors, then ϕ_k is a new independent eigendirection that is retained for a parsimonious representation of the data. To quantify the accuracy of the fit, the following metric is used:

$$r_k = \sqrt{\frac{\sum_{i=1}^{n}(\phi_k(i) - f(\widetilde{\boldsymbol{\Phi}}_{k-1}(i)))^2}{\sum_{i=1}^{n}(\phi_k(i))^2}}$$

A small value of r_k is associated with a ϕ_k that is a harmonic function of the previous eigenmodes, whereas a higher value of r_k signifies that ϕ_k is a new independent direction on the data manifold. It has been demonstrated (Dsilva et al., 2018), that selecting only the eigenvectors that correspond to higher values of r_k leads to a parsimonious representation of the data. Eventually, the vector x_i is mapped to a vector whose first component is the ith component of the first selected nontrivial eigenvector, whose second component is the ith component of the second selected nontrivial eigenvector, etc.

To map a new point, x_{new}, from the ambient space to DMAPs space, a mathematically elegant approach known as Nyström extension is used (Nyström, 1929; Fowlkes et al., 2001). The starting point of the Nyström extension is to compute the distances between the new point, x_{new} in ambient space, and the N data points in the original data set; the same normalizations used for DMAP need to be applied also here. The Nyström extension formula reads

$$\phi_j(x_{new}) = \lambda_j^{-1}\sum_{i=1}^{N}\tilde{k}(x_i, x_{new})\phi_j(x_i),$$

where λ_j is the jth eigenvalue, $\phi_j(x_i)$ is the ith component of the j-th eigenvector and $\tilde{k}(\cdot, x_{new})$ is the kernel's value between the new point and each point in the original data set.

3. Geometric harmonics

Geometric Harmonics was introduced in Coifman and Lafon (2006a), inspired by the Nyström Extension, as a scheme for *extending* functions defined on data $\mathbf{X}$, $f(\mathbf{X}) : \mathbf{X} \to \mathbb{R}$, for $x_{new} \notin \mathbf{X}$. This extension is

achieved by using a particular set of basis functions called Geometric Harmonics. These functions are computed as eigenvectors of the symmetric $N \times N$ $\mathbf{W}$ matrix, defined in Section 2. The eigendecomposition of the symmetric and positive semidefinite matrix $\mathbf{W}$ leads to a set of orthonormal eigenvectors $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_N$ and non negative eigenvalues $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N \geq 0$.

From this set of eigenvectors, we consider a truncated subset $S_\delta = (\alpha: \sigma_\alpha \geq \delta\sigma_1)$ where $\delta > 0$. The extension of f for a new point x_{new} is accomplished by firstly projecting the function of interest in the (truncated) computed set of eigenvectors

$$f \to P_\delta f = \sum_{\alpha \in S_\delta} \langle f, \psi_\alpha \rangle \psi_\alpha,$$

where P_δ denotes the projection of the function f on the eigenvectors we retained and $\langle f, \psi_\alpha \rangle$ is the inner product between the function f and the obtained α-th eigenvector ψ_α. This projection step is performed only once.

To obtain the values of function f for $x_{new} \notin \mathbf{X}$, we extend the Geometric Harmonic functions as

$$\Psi_\alpha(x_{new}) = \sigma_\alpha^{-1} \sum_{i=1}^{N} w(x_{new}, x_i)\psi_\alpha(x_i)$$

where σ_α is the α-th eigenvalue, $\psi_\alpha(x_i)$ is the ith component of the α-th eigenvector and $w(x_{new}, x_i)$ denotes the kernel

$$w(x_{new}, x_i) = \exp\left[-\left(\frac{\|x_{new} - x_i\|}{\tilde{\epsilon}}\right)^2\right]$$

The function f at x_{new} is then estimated as a linear combination of the extended Geometric Harmonics

$$(Ef)(x_{new}) = \sum_{\alpha \in S_\delta} \langle f, \psi_\alpha \rangle \Psi_\alpha(x_{new})$$

where Ef denotes the estimated values of f at x_{new}.

4. Double diffusion maps and latent harmonics

A *slight* twist of the Geometric Harmonics is presented in this section. As discussed above, Geometric Harmonics constructs an input–output mapping between the ambient coordinates $\mathbf{X}$ and a function of interest f defined on $\mathbf{X}$. However, it is possible, if the data are lower dimensional to construct a map in terms of only the non-harmonic eigenvectors, $\boldsymbol{\Phi}$. This can be achieved by operating directly on the non-harmonic DMAPs coordinates. Similar to the *traditional* Geometric Harmonics, firstly an affinity matrix is constructed

$$\overline{w}(i,j) = \exp\left[-\left(\frac{\|\phi_i - \phi_j\|}{\epsilon^\star}\right)^2\right].$$

In this case the affinity matrix is constructed in term of the DMAPs coordinates. To distinguish the notation between Geometric Harmonics and Double DMAPs, we will use overlined symbols and ϵ^*. As in the *traditional* Geometric Harmonics, the function f is projected to a truncated set of the obtained eigenvectors

$$f \to P_\delta f = \sum_{\beta \in \overline{S}_\delta} \langle f, \overline{\psi}_\beta \rangle \overline{\psi}_\beta.$$

The extension of f for $\boldsymbol{\phi}_{new}$ is achieved by firstly extending the values of the Geometric Harmonic functions $\overline{\Psi}_\beta$ for $\boldsymbol{\phi}_{new}$,

$$\overline{\Psi}_\beta(\boldsymbol{\phi}_{new}) = \overline{\sigma}_\beta^{-1} \sum_{i=1}^{N} \overline{w}(\phi_{new}, \phi_i)\overline{\psi}_\beta(\boldsymbol{\phi}_i),$$

and then estimating the value of f at $\boldsymbol{\phi}_{new}$

$$(Ef)(\boldsymbol{\phi}_{new}) = \sum_{\beta \in \overline{S}_\delta} \langle f, \overline{\psi}_\beta \rangle \overline{\Psi}_\beta(\boldsymbol{\phi}_{new})$$

5. Gappy POD

In this section the Gappy POD method is summarized, in order to better highlight the differences with the proposed approach. Lets consider the matrix, $\overline{\mathbf{X}} = \mathbf{X}^{\mathrm{T}}$, the transpose of the data set $\mathbf{X}$. A POD basis, $\mathbf{U} \in \mathbb{R}^{d \times N}$, of $\overline{\mathbf{X}}$ is computed. We approximate $\overline{\mathbf{X}}$ using a truncated number, N^p of basis vectors, where $N^p \leq N$ such that:

$$\frac{\|\overline{\mathbf{X}} - \widetilde{\mathbf{X}}\|}{\|\overline{\mathbf{X}}\|} 100 \leq \epsilon^p$$

where ϵ^p is a prescribed tolerance for the truncation; in our case $\epsilon^p = 5\%$ was selected.

Let us consider a vector x_{new} not in the original data set, spanned by the same basis $\mathbf{U}$, for which only μ values of this vector are known (μ partial observations are known) denoted as $x_{new}^{partial}$

$$x_{new}^{partial} = m \odot x_{new}$$

where $\odot$ denotes the pointwise multiplication, between the vector x_{new} and m the *masking vector* that contains ones in the positions of the μ known vector components and zeros in the rest.

The goal is to find the missing values (observations) of $x_{new}^{partial}$. This is achieved by the following step

$$x_{new}^{rec} = \mathbf{U}^{\mathbf{p}} c$$

where x_{new}^{rec} is the recovered vector, $\mathbf{U}^p$ is the truncated POD basis and c are the POD coefficients estimated by solving the following linear system of equations

$$\mathbf{A} \cdot c = (m \odot \mathbf{U}) \cdot x_{new}^{partial}$$

where $\mathbf{A}$ is given by

$$\mathbf{A} = (m \odot \mathbf{U})^{\mathrm{T}}(m \odot \mathbf{U})$$

5.1. The drawback of hyperplanes in parsimoniously capturing nonlinearity

In the case where the data live on a curved manifold, the size of the POD basis required for accurate reconstruction of the data is expected to be higher than the intrinsic dimension of the manifold, since it needs to be spanned by several hyperplanes. In Fig. 1, an illustrative caricature of data sampled from the singularly perturbed system $\dot{x} = 2 - x - y$, $\dot{y} = \frac{1}{\epsilon}(x - y)$ is aiming to convey this shortcoming of POD in contrast to its nonlinear counterpart, DMAPs. In Fig. 1(a) the direction of the first POD basis vector $\mathbf{u}_1$ is shown as a red vector. The direction of $\mathbf{u}_1$ parametrizes the data but does not fully span them because of their nonlinearity. To accurately reproduce this nonlinear curve both POD basis vectors are needed. Please notice that, as can be seen from Fig. 1(b), the coefficient of the second POD basis vector ($\mathbf{u}_2$) is a function of the first POD mode coefficient ($\mathbf{u}_1$). On the contrary, Fig. 1(c) illustrates that DMAPs applied on this data set need only a single coordinate, ϕ_1 to fully parametrize the data.

6. Case study

The case study here is the vertical cylindrical Metal Organic Chemical Vapor Phase Deposition (MOCVD) reactor used for the deposition of Cu from copper amidinate, described in Spencer et al. (2021), shown in Fig. 2. The mixture of gas reactants enters the chamber from the top, then gets evenly distributed by passing through a showerhead and eventually leads to the deposition of Cu on a heated substrate. The quality of the produced film is affected by various process parameters; among them, significant effect have the deposition temperature, T, the mass flow rate of incoming gas, M and the chamber pressure, P.

In this work, the conservation equations for mass, momentum, energy and species are discretized with the finite volume method with 11,500 finite volumes and solved in ANSYS/Fluent, in a two-dimensional computational geometry with axial symmetry (cf. Fig. 2).

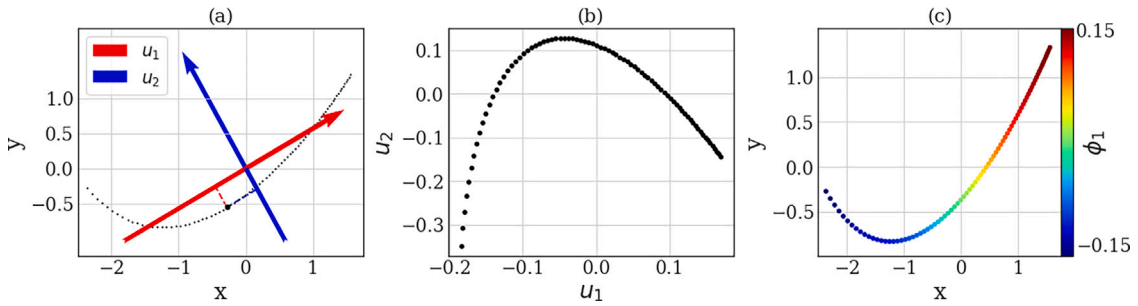

Fig. 1. (a) A data set sampled from a singularly perturbed dynamical system is shown (black dots). The span of the first POD basis vector is shown with a red vector ($\mathbf{u}_1$) and the span of the second POD basis vector is shown as a blue vector ($\mathbf{u}_2$) (b) The components of the first POD basis vector $\mathbf{u}_1$ plotted against the components of the second POD basis vector $\mathbf{u}_2$ indicating that $\mathbf{u}_2$ is a function of $\mathbf{u}_1$ (c) the first non-trivial DMAPs, ϕ_1, eigenvector is plotted as function on the data set, indicating that is able to fully parametrize it.

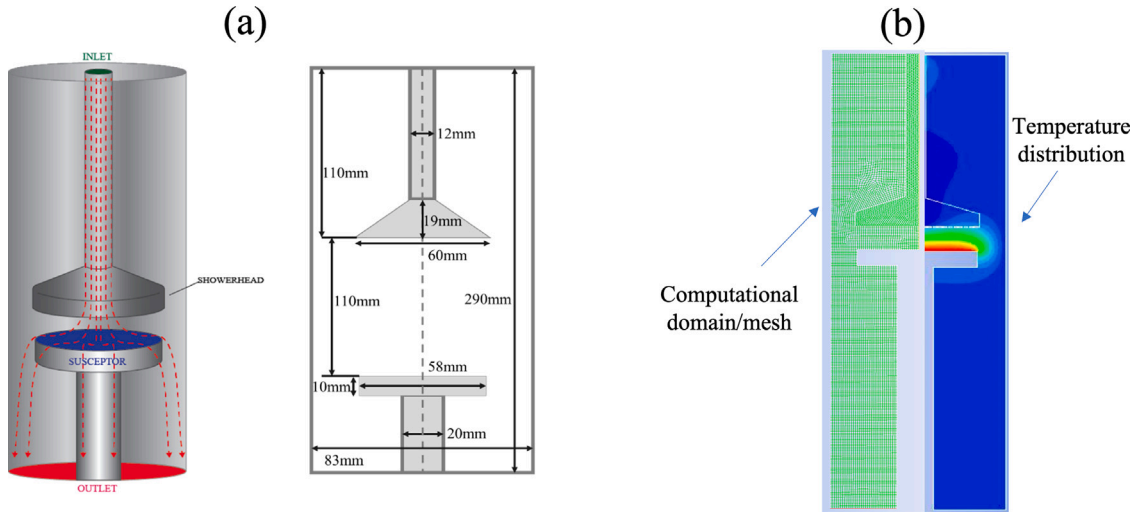

Fig. 2. (a) Schematic illustration of the experimental MOCVD reactor; (b) 2D computational domain with axial symmetry and representative temperature distribution.

The interested reader is referred to Spencer et al. (2021) for details on the setup of the CFD model. Here some relevant details are included for completeness. In this implementation, the temperature of the walls and of the incoming gas mixture is constant at $T_w = 370$ K and $T_g = 370$ K respectively. The composition of the mixture of incoming gas, in terms of mass fractions is $Ar/N_2/H_2/$ Cu amidinate = 73.9%/25.5%/0.4%/0.1%. Steady states are computed for various inputs, i.e. values of three critical, for the process, parameters: the substrate temperature, T, the chamber pressure, P and the mass inflow rate of the mixture of gas reactants, M. Specifically, the input (or parameter) space is uniformly sampled for 487 K $< T <$ 501 K, 7.97×10^{-6} kg/s $< M < 8.87 \times 10^{-6}$ kg/s and 1383 Pa $< P <$ 1463 Pa, as shown in Fig. 3a. This region in parameter space is interesting for the process, as it corresponds to the transition between the kinetics- and transport-limited regime, i.e. the process turns from being limited by the reaction rate to being defined by the diffusion rate of species toward the deposition surface.

The resulting states are collected as an ensemble of high -dimensional vectors containing the values of the two components of velocity, pressure, temperature and precursor mass fraction at each discretization point (cf. Fig. 3b). Eventually, the sample matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$ is assembled, where $d = 58{,}100$ degrees of freedom (dimensions) and $N = 720$ samples, i.e. vectors containing steady states.

7. Results

7.1. Interpolation between ambient and intrinsic space

The DMAPs algorithm is implemented to identify a low dimensional parametrization of the data manifold. In order to establish which are the independent coordinates, it is useful to examine the variation of each eigenvector versus the first non-trivial eigenvector of the Markov matrix, as shown in Fig. 4: the two-dimensional variation of ϕ_3 vs. ϕ_2 (cf. Fig. 4a) signifies that there are two independent directions on the data. Having established that the intrinsic space is at least two-dimensional, the subsequent DMAP eigenvectors are plotted against the first two. The fact that the 3D plot of ϕ_4 vs. ϕ_3 and ϕ_2 reveals a surface (cf. Fig. 4b), suggests that ϕ_4 is a harmonic of the previous two. In contrast, ϕ_5 is a new independent eigenvector and hence its variation versus the first two independent DMAPS reveals a 3D object (cf. Fig. 4c).

These visual observations are verified by the results of the implementation of the local linear regression algorithm (Dsilva et al., 2018), according to which a function $f(\phi_{k-1}, \phi_{k-2}, \dots \phi_1)$ is fitted to ϕ_k. The results suggest that indeed ϕ_2, ϕ_3, ϕ_5 (cf. Fig. 5), represent a parsimonious low dimensional embedding of the available data, since the error, r_k of the local linear regression function is high for ϕ_3 and ϕ_5. On the contrary, r_k for ϕ_4 is small, indicating that it is a harmonic function of ϕ_2.

In an attempt to provide a physical interpretation of the derived low dimensional coordinates, the eigenvectors ϕ_2, ϕ_3, ϕ_5 are plotted and colored by the values of T (Fig. 6a), M (Fig. 6b) and P (Fig. 6c). Each one of the three directions in the reduced space corresponds to the variation of each one of the three input parameters. This is shown with more clarity in (Fig. 6d), (Fig. 6e) and (Fig. 6f), specifically that ϕ_2 corresponds to T, ϕ_3 to M and ϕ_5 to P. Therefore, the DMAP coordinates provide a parametrization that appears to be one-to-one with the actual physical parameters that were varied in order to produce it.

The inverse map, f^{-1}, i.e. from the reduced to the ambient space, is approximated with the double DMAPS Geometric Harmonics interpolation and its accuracy is assessed against a random test sample. For

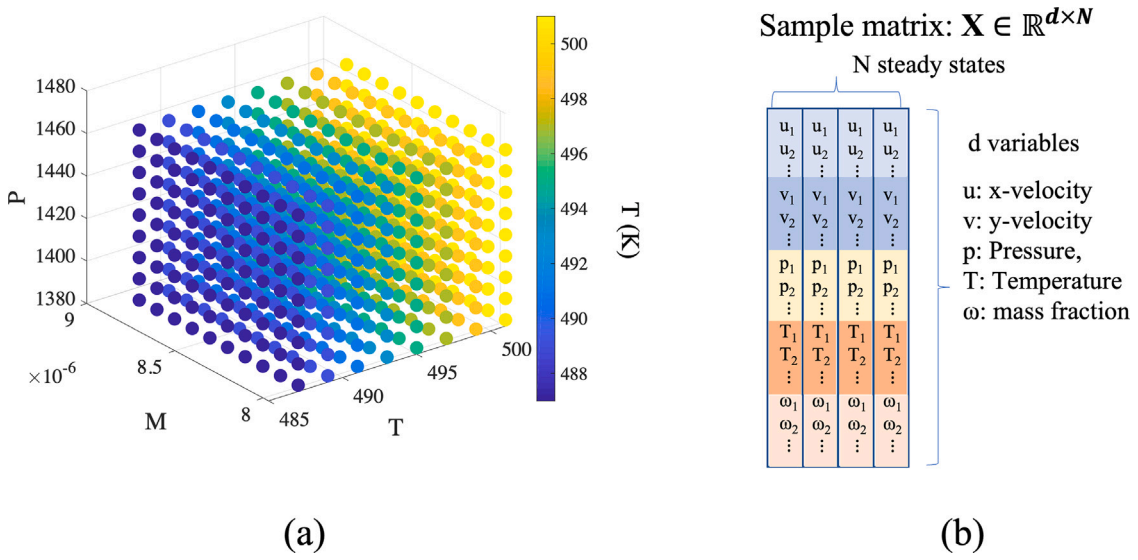

Fig. 3. (a) Input parameters for the collection of data; (b) Sample set $\mathbf{X} \in \mathbb{R}^{d \times N}$, where d is the total number of degrees of freedom of the CFD model and N is the number of collected steady states.

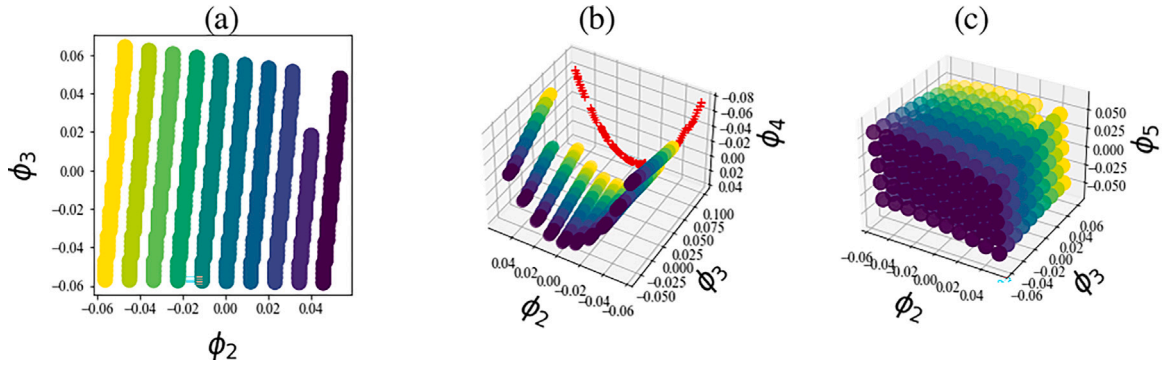

Fig. 4. Diffusion Maps: (a) ϕ_3 vs. ϕ_2; (b) ϕ_4 vs. ϕ_3 and ϕ_2; (c) ϕ_5 vs. ϕ_2 and ϕ_3; the three-dimensional "spread" of ϕ_5 with respect to ϕ_3 and ϕ_2 suggests that these are independent directions on the low dimensional space; the distribution of ϕ_4 with respect to ϕ_2 and ϕ_3 lies on a surface which indicates that ϕ_4 is a harmonic function of ϕ_2 and ϕ_3.

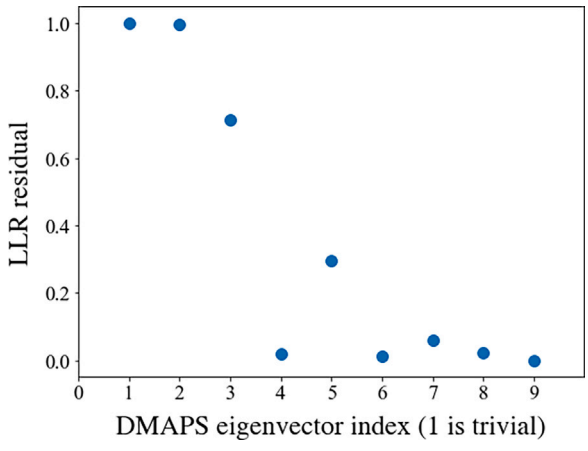

Fig. 5. Residual of the local linear regression algorithm, r_k; The first, second and fourth nontrivial eigenvectors (ϕ_2, ϕ_3, ϕ_5) have the highest r_k values, indicating that they each represent independent directions on the data manifold.

this implementation, the value of the kernel parameter $\epsilon = 5.105$ and 38 eigenvectors are retained as interpolation functions. The % relative error for each vector, shown in Fig. 7, is computed as:

$$\%\text{error} = \left(\frac{X_i^{predicted} - X_i^{actual}}{X_i^{predicted}} \right) 100.$$

The average error for the test samples is 0.01%.

7.2. Prediction of outputs for a new set of inputs

The mapping from the ambient space to the intrinsic space and back provides a means of deriving various useful correlations, such as the prediction for a new output of a set of input parameters. Specifically, Geometric Harmonics interpolation is implemented in order to define functions $g_1 : (\phi_2) = g_1(T, P, M)$, $g_2 : (\phi_3) = g_2(T, P, M)$ and $g_3 : (\phi_5) = g_3(T, P, M)$. The interpolation mean squared error for ϕ_2, ϕ_3 and ϕ_5 is 7.39×10^{-8}, 2.23×10^{-6} and 9.12×10^{-6} respectively. The predicted reduced coordinates versus the actual ones are shown in Fig. 8.

Having established a mapping from the input space to the reduced variables, Double DMAPs interpolation with Geometric Harmonics – which defines the inverse map described in the previous paragraph – is implemented in order to find the corresponding state variables in the ambient space. The average mean squared error for the test sample is (1.2×10^{-5}).

7.3. Prediction of the inputs that correspond to a new output

It is possible to find the values of inputs, (T, P, M) that correspond to a new output, by first using the Nyström extension to obtain the

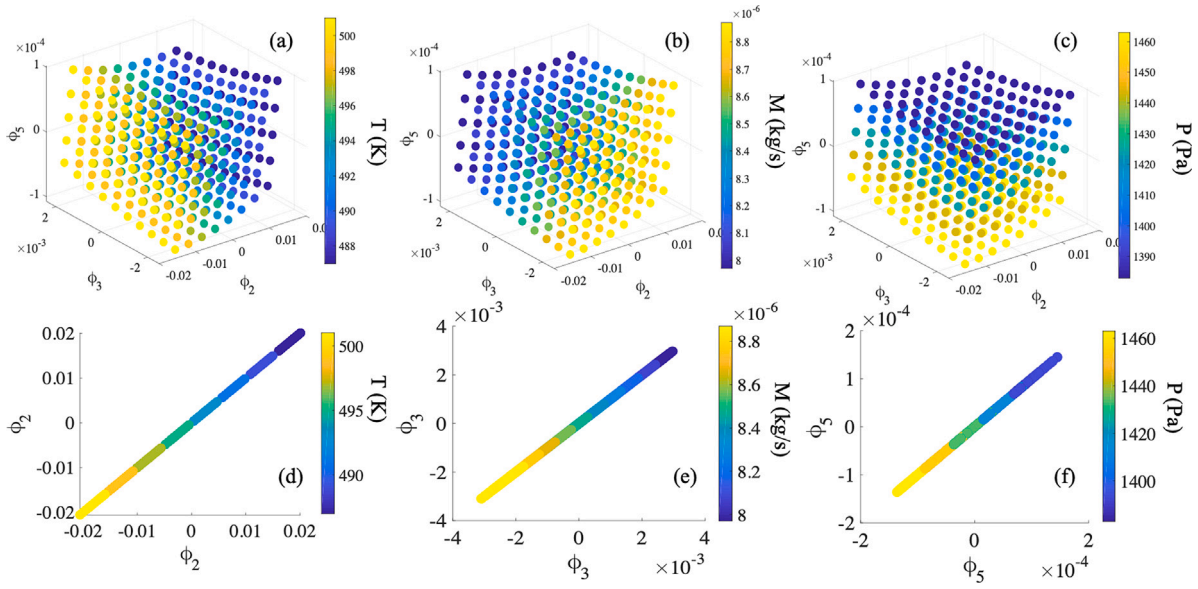

Fig. 6. The identified DMAP coordinates ϕ_2, ϕ_3, ϕ_5, plotted and colored by the process inputs (a) T, (b) M, (c) P; the three-dimensional plots show that the variation of ϕ_2, ϕ_3, ϕ_5, follow the variation of T, M and P respectively. This is further demonstrated in (d), (e) and (f) respectively.

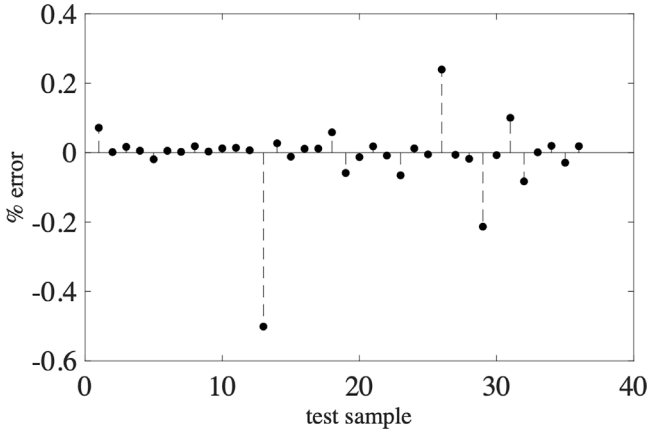

Fig. 7. Prediction error of the Double DMAP implementation of Geometric Harmonics for each one of the vectors x_i in the test sample.

corresponding reduced variables. Then an interpolation function can be constructed from the reduced coordinates to the input parameters, with Double DMAPS: $(T, P, M) = f(\phi_2, \phi_3, \phi_5)$. The interpolation relative error is below 0.5%, as shown in the top row of Fig. 9 and the predicted input parameters versus the actual ones are shown in the bottom row of Fig. 9.

7.4. Prediction of inputs from partial observations

Instead of a full state vector in ambient space, it is also possible to use partial observations, such as, for example, the value of temperature at a few points, in order to predict the corresponding input values. To this end, Geometric Harmonics interpolation is implemented in order to define a function $(\phi_2, \phi_3, \phi_5) = f(X^1_{partial}, X^2_{partial}, X^3_{partial}, \cdots, X^m_{partial})$. In this implementation, $X^i_{partial}, i = 1, \ldots, m$ are values of temperature in seven random positions (cf. Fig. 11a). From the reduced space, it is now possible to map to the input space with the interpolation function from the reduced space to the input space discussed in the previous paragraph.

The number of partial observations required in order to define the function from the partial observations to the input space is dictated by Whitney's embedding theorem (Whitney, 1936). $2n + 1$ independent observations are provably sufficient to create an embedding of the m-dimensional manifold. Here, for the three-dimensional reduced manifold, at least seven partial observations should be considered. Eventually the predicted values of the input parameters for the test sample are in excellent agreement with the actual values, as presented in Fig. 10, where the prediction relative error for the test sample is shown to be on average less than 0.1%.

7.5. Prediction of partial observations from other partial observations

The possibility to predict part of the observations, given a different part of the observations, is discussed in this section. As an illustrative example, the prediction of the value of mass fractions right above the heated susceptor surface (cf. Fig. 11b), given seven temperature measurements in a different part of the geometry (cf. Fig. 11a) will be presented here. This choice is dictated by the fact that, although in this particular process the mass fraction of precursor reaching the deposition surface is crucial for determining both the quality of the product and also the film deposition rate, it is not easily measurable. On the other hand, temperature measurements are generally more accessible, and the idea is to use such measurements in order to make predictions for quantities that are harder to measure, similar to the concept of a nonlinear observer.

To begin with, the function $(\phi_2, \phi_3, \phi_5) = f(X^1_{partial}, X^2_{partial}, \cdots, X^m_{partial})$, discussed in the previous paragraph is used in order to map from the partial observations to the reduced space. The inverse function, f^{-1}, from the reduced to ambient space, is then used in order to predict the desired values (in this case the values of the species mass fractions above the heated substrate at seven points). The average relative error is less than 0.5% (cf. Fig. 12a), whereas the predicted versus the actual values of the mass fraction at a single point above the substrate is shown in Fig. 12b.

7.6. Implementation of Gappy POD and comparison to DMAPs-based predictions

In this section the Gappy POD method is implemented and the results are compared to those delivered by the DMAPs/Geometric Harmonics method presented in the previous sections. The first step is

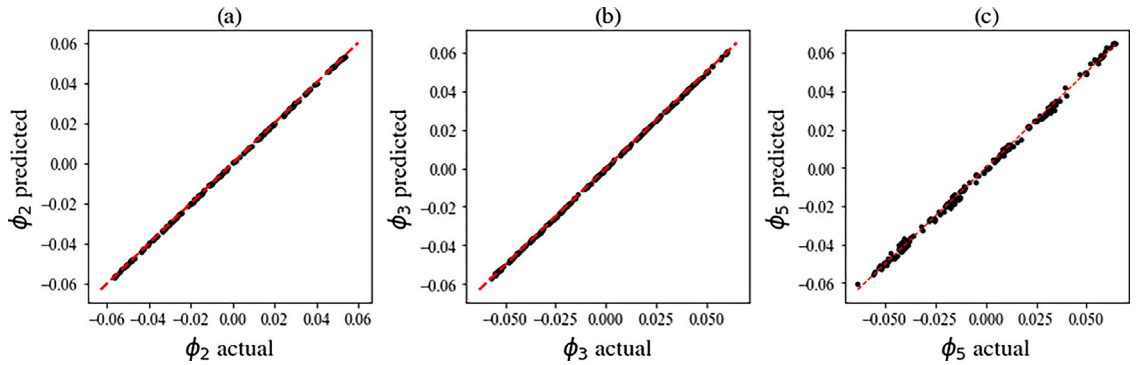

Fig. 8. The predicted versus the actual DMAPs coordinates. The solid lines correspond to $y = x$.

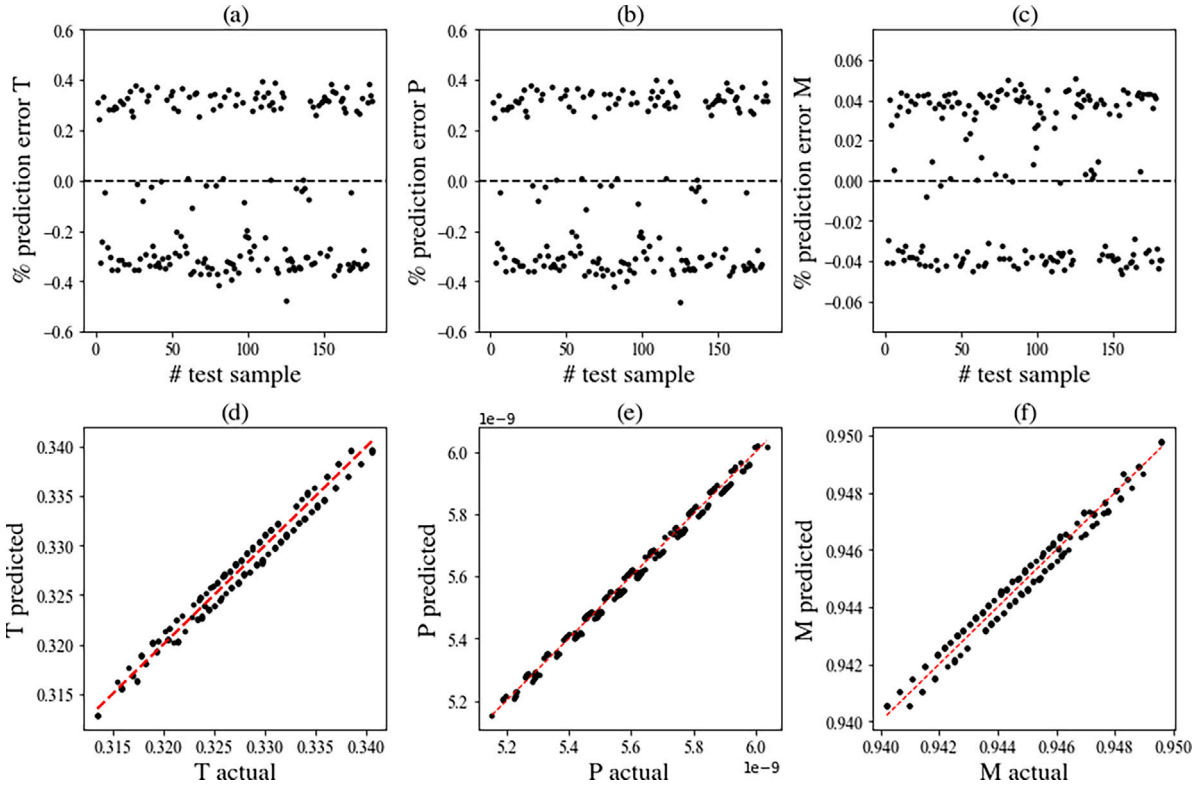

Fig. 9. Prediction % relative error of input parameters (a) T, (b) P, (c) M that correspond to a new state in the ambient space. Predicted versus actual values of (d) T, (e) P, (f) M; the red dashed lines correspond to $y = x$.

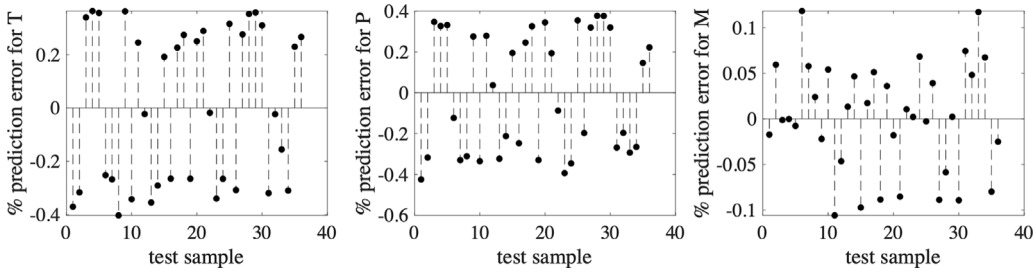

Fig. 10. Prediction error of inputs T, P, M.

to compute a basis of the given data-set using Singular Value Decomposition. The size of the basis is determined based on the cumulative percentage of the energy of the system captured by i modes, defined as

$$E_i\% = \frac{\sum_{n=1}^{i} \xi_n}{\sum_{n=1}^{m} \xi_n} * 100$$

where ξ_n stands for the nth singular value of the diagonal matrix Ξ that results from the singular value decomposition of the transpose of the data matrix $\overline{\mathbf{X}}$. In addition to that, the reconstruction error of the data-set is computed for increasing size of the POD basis. In this implementation, and for the purposes of comparison, the selected POD

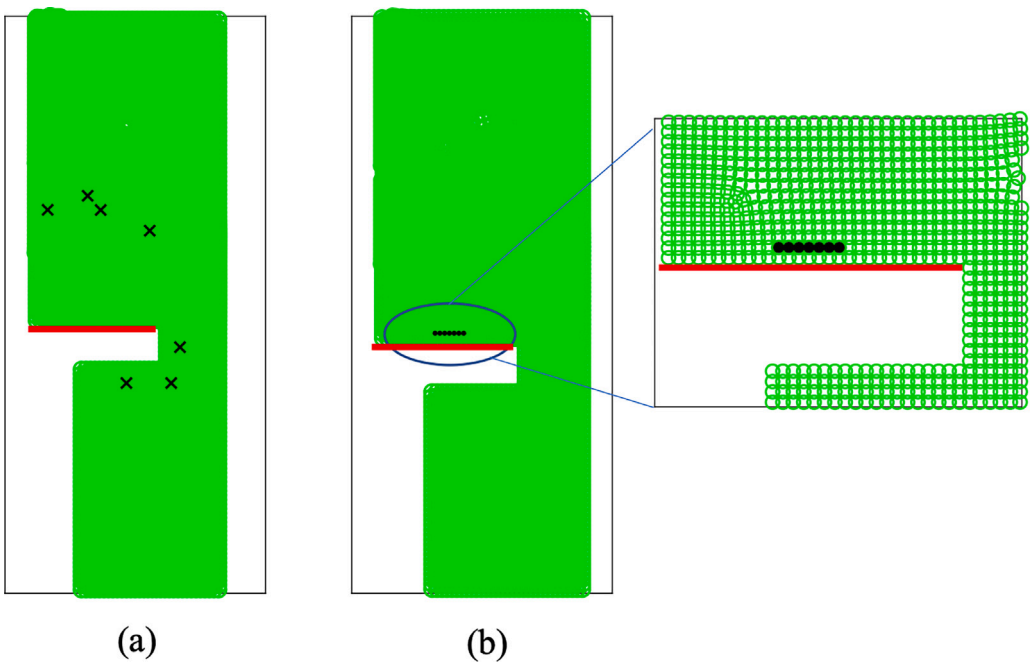

Fig. 11. Partial observations: (a) positions in computational geometry where temperature values are considered; (b) positions in computational geometry where mass fractions are predicted; The red line indicates the heater susceptor surface where deposition of material takes place.

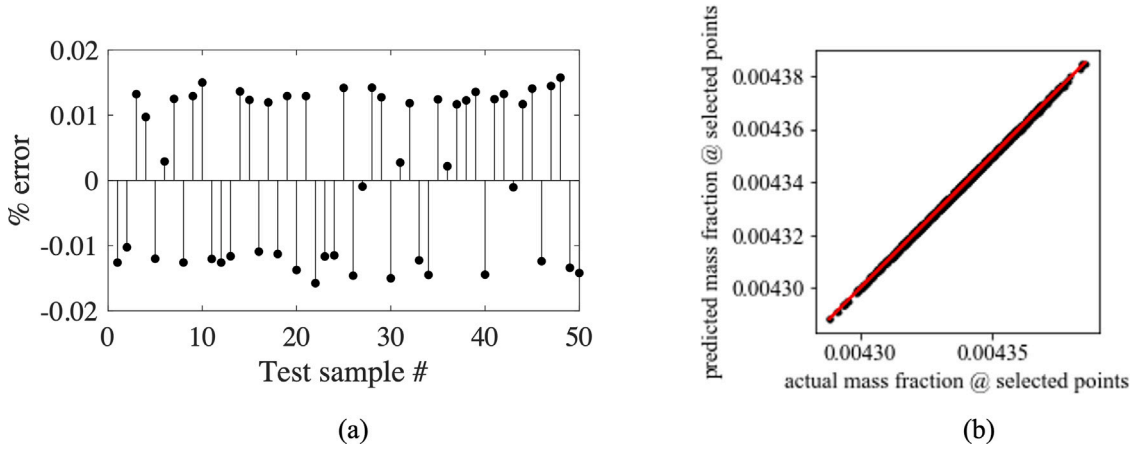

Fig. 12. (a) Approximation error for the mass fraction, ω values in the sample test set (b) predicted vs. actual mass fraction values at one point for all test samples.

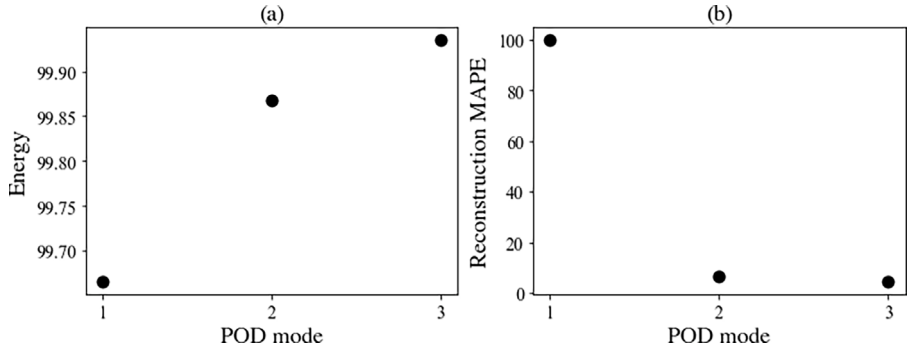

Fig. 13. (a) Cumulative energy captured by POD modes (b) Mean absolute percentage error of the reconstructed data-set **X**.

basis consists of 3 vectors that capture 99.93% of the energy of the system (cf Fig. 13a) and the approximation error is 4.7% (cf Fig. 13b).

Initially, the goal is to predict the output vector x_{new}, containing the distributions of velocity, pressure, temperature and mass fractions, given a new set of process parameters (T, P, M). Therefore, the partial data considered in this comparison correspond to the values of the three process parameters. The predicted values are compared to the projection of the test vector on the POD basis. In this case the

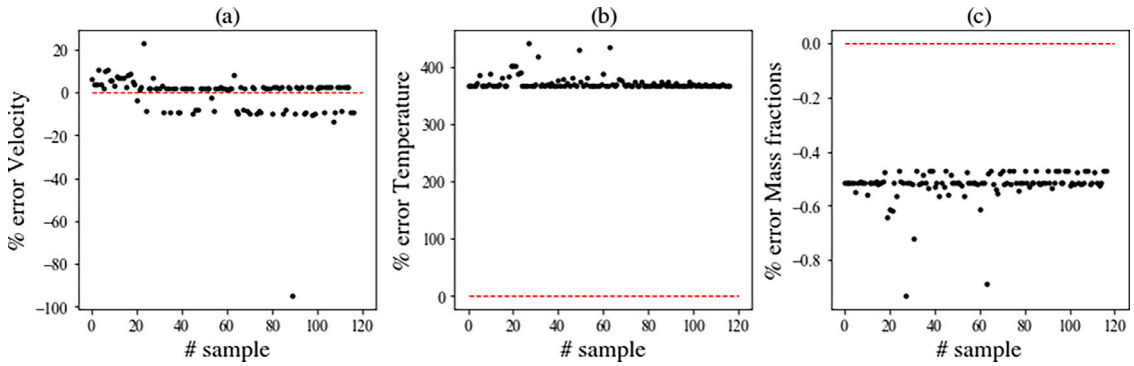

Fig. 14. Prediction error (a) x-velocity (b) temperature (c) precursor mass fraction.

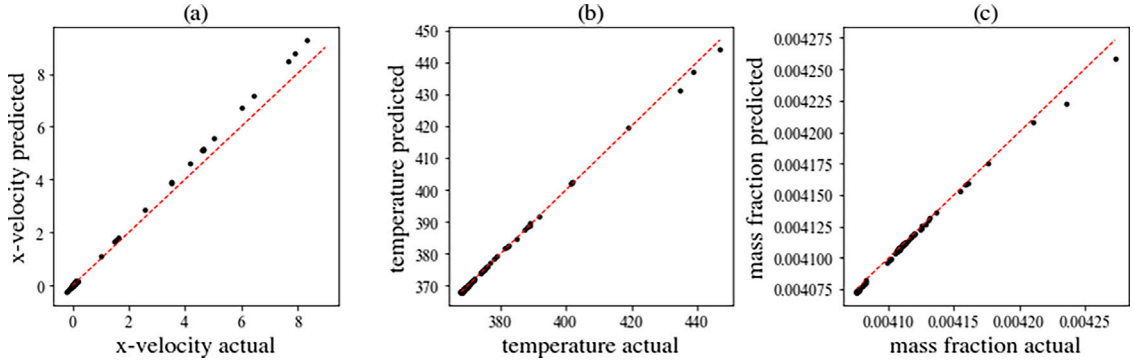

Fig. 15. Predicted vs. actual values of (a) x-velocity (b) temperature (c) precursor mass fraction.

predictions are inaccurate, especially for the temperature distributions (cf. Fig. 14b) and even unphysical as negative values for the mass fractions are produced (cf. Fig. 14c).

Nevertheless, the results of the Gappy POD method are heavily influenced by the choice of the "known" values, i.e. the partial data considered. To illustrate this, instead of the three process parameters, three values of temperature are considered known, a subset of the values mentioned in Section 7.4 and shown in Fig. 11a. In this case the approximation error drops to 8%, while the predicted versus actual velocity, temperature and mass fractions are shown in Fig. 15a, 15b and 15c respectively.

This finding is directly related to the condition number of the matrix **A**, defined in Section 5. Specifically, the elements of this matrix result from the inner products of the "Gappy" POD vectors, i.e. the elements of the original POD vectors that correspond to the known elements of x_{new}. These are no longer orthogonal, and therefore the matrix **A** is fully populated. In general, the positions of the known elements, and hence the non-zero elements of **A**, must be such that orthogonality is preserved. Furthermore, the diagonal entries of **A** must not be very small, which means that the POD basis element at that point must not be small. These two requirements are reflected in the condition number of the matrix **A**: specifically, the smaller the condition number, the more they are satisfied. This analysis is presented in Willcox (2006), in the context of optimal sensor placement, and in Alonso et al. (2004a), where the angle between the measurement subspace and the low dimensional space that spans the data is taken into account. The condition number of **A** drops from 12.4×10^{12} to 24.5 when the known components correspond to the input parameters and the temperature measurements respectively.

When all the temperature measurements at points shown in Fig. 10a are known, the prediction error drops further to 0.1% and it is possible to reproduce accurately the distributions of velocity (cf Fig. 16a), temperature (cf Fig. 16b) and pressure (cf Fig. 16c), as well as the corresponding process parameters (cf Fig. 17 a, b and for T, P and M respectively). In this case the condition number of the matrix **A** is 12.45.

The results above, point directly to the apparent disadvantage of Gappy POD, when compared to the proposed methodology, based on DMAPs: given the same number of POD vectors as DMAP coordinates, the accuracy of Gappy POD is inherently linked to which elements of the partial vector are known. This consideration is much less prevalent when DMAPs/Geometric Harmonics are implemented, which enables the accurate prediction of entire vectors of outputs for various new combinations of input parameters, as well as from partial measurements.

Apart from the pathology related to the known values, one expected drawback of Gappy POD is directly linked to the inability of hyperplanes to accurately parametrize a curved manifold. In this implementation this is reflected in the size of the POD basis required to reconstruct the data set with an error of less than 1%: here 5 POD modes are required to achieve 0.7% reconstruction error, more than the only 3 DMAPs coordinates that are sufficient to parametrize the manifold. By selecting 5 POD modes, it is no longer possible to reconstruct the state vector, given the values of the *three* input parameters, T, P and M. In this case, the values of at least 5 measurements are necessary to achieve accurate reconstruction of the state vector. This is illustrated in Fig. 18, where the predicted input parameters and distributions of velocity, temperature and mass fractions are plotted against the actual values. In this implementation, the value of temperature at five positions are considered known.

8. Conclusions

This work presents a data-driven workflow, based on nonlinear manifold learning, specifically DMAPs, that enables the parsimonious description of high-dimensional data, but also interpolation and regression for out-of-sample predictions with remarkable accuracy.

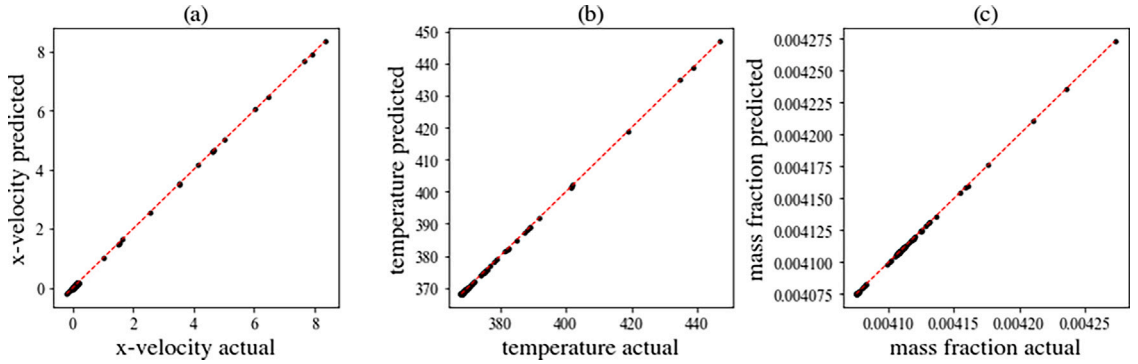

Fig. 16. Predicted vs. actual values of (a) x-velocity (b) temperature (c) precursor mass fractions.

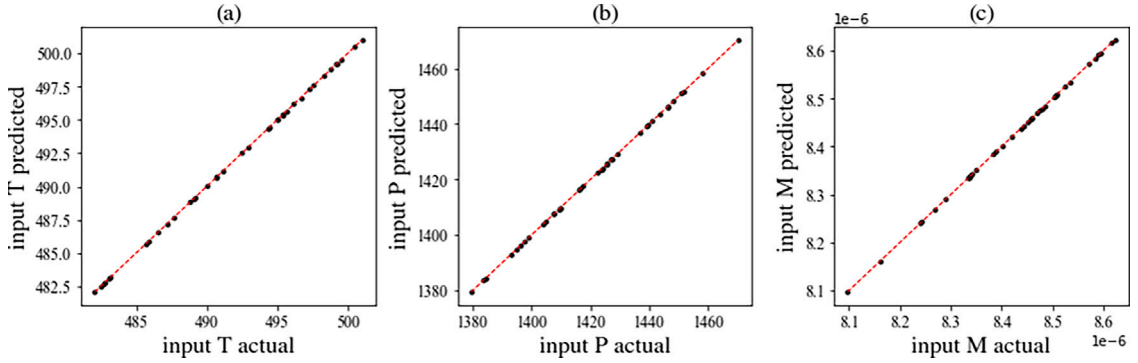

Fig. 17. Predicted vs. actual values of (a) temperature, T (b) mass flow rate, M (c) pressure, P.

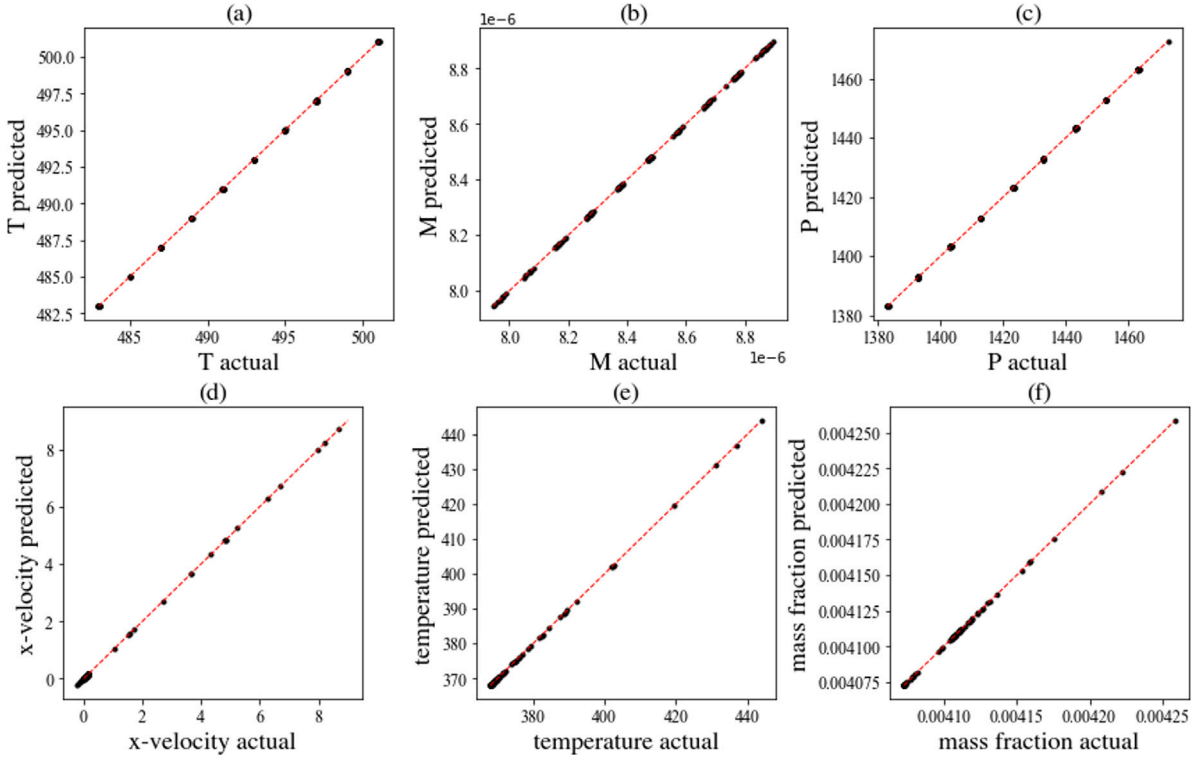

Fig. 18. Gappy POD performance: 5 POD vectors and 5 known Temperature values; predicted vs. actual values of (a) temperature, T (b) mass flow rate, M (c) pressure, P, (d) x-velocity (e) temperature (f) mass fractions.

The case study here is a Chemical Vapour Deposition Reactor, although the proposed approach is not restricted to this particular application. Mapping between the reduced (or DMAPs) and the ambient space is achieved with Geometric Harmonics, amended with a special "twist" that implements a second round of DMAPs in order to define functions for accurate interpolation.

Having defined the reduced description of the data-set and the means to map back and forth between ambient and reduced DMAPs space, we proceed to show the implementation of out-of-sample predictions: We first demonstrate the possibility to predict the high-dimensional output of a new set of inputs, here process parameters, namely temperature, pressure and the mass inflow rate, without additional expensive CFD simulations. The opposite, i.e. accurately predicting the inputs that correspond to a new output, is also possible.

Based on the reduced description of the data-set, provided by DMAPs, and the computational means of transitioning between the ambient and the reduced DMAPs space of the data, we show how to predict not only outputs but also inputs, i.e. process parameters, when only a handful of measurements, temperature in this case, are known we also demonstrate the superiority of interpolating on nonlinear manifolds (our "Gappy DMAPs" approach) rather than on linear hyper-planes, as in the Gappy POD approach.

CRediT authorship contribution statement

Eleni D. Koronaki: Conceived and designed the analysis, Collected the data, Performed the analysis, Wrote the paper. **Nikolaos Evangelou:** Contributed data or analysis tools, Wrote the paper. **Yorgos M. Psarellis:** Contributed data or analysis tools, Wrote the paper. **Andreas G. Boudouvis:** Wrote the paper, Supervision. **Ioannis G. Kevrekidis:** Conceived and designed the analysis, Contributed data or analysis tools, Wrote the paper, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The work of YGK, NE and YMP was partially supported by the US AFOSR FA9550-21-0317 and the US DOE SA22-0052-S001. E.D.K. has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 890676 - DataProMat.

References

Akiki, G., Suchet, D., Daineka, D., Filonovich, S., Bulkin, P., Johnson, E.V., 2020. Area selective deposition of silicon by plasma enhanced chemical vapor deposition using a fluorinated precursor. Appl. Surf. Sci. 531, 147305.

Alhajeri, M.S., Wu, Z., Rincon, D., Albalawi, F., Christofides, P.D., 2021. Machine-learning-based state estimation and predictive control of nonlinear processes. Chem. Eng. Res. Des. 167, 268–280.

Alonso, A.A., Frouzakis, C.E., Kevrekidis, I.G., 2004a. Optimal sensor placement for state reconstruction of distributed process systems. AIChE J. 50 (7), 1438–1452.

Alonso, A.A., Kevrekidis, I.G., Banga, J.R., Frouzakis, C.E., 2004b. Optimal sensor location and reduced order observer design for distributed process systems. Comput. Chem. Eng. 28 (1–2), 27–35.

Balasubramanian, M., Schwartz, E.L., Tenenbaum, J.B., de Silva, V., Langford, J.C., 2002. The isomap algorithm and topological stability. Science 295 (5552), 7.

Bhattacharjee, S., Matouš, K., 2016. A nonlinear manifold-based reduced order model for multiscale analysis of heterogeneous hyperelastic materials. J. Comput. Phys. 313, 635–653.

Cassez, F., Tripakis, S., 2008. Fault diagnosis with static and dynamic observers. Fund. Inform. 88 (4), 497–540.

Chiavazzo, E., Gear, C.W., Dsilva, C.J., Rabin, N., Kevrekidis, I.G., 2014. Reduced models in chemical kinetics via nonlinear data-mining. Processes 2 (1), 112–140.

Chiu, T.Y., Christofides, P.D., 2000. Robust control of particulate processes using uncertain population balances. AIChE J. 46 (2), 266–280.

Coifman, R.R., Kevrekidis, I.G., Lafon, S., Maggioni, M., Nadler, B., 2008. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. Multiscale Model. Simul. 7 (2), 842–864.

Coifman, R.R., Lafon, S., 2006a. Diffusion maps. Appl. Comput. Harmon. Anal. 21 (1), 5–30.

Coifman, R.R., Lafon, S., 2006b. Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions. Appl. Comput. Harmon. Anal. 21 (1), 31–52.

Desenfant, A., Laduye, G., Vignoles, G., Chollon, G., 2021. Kinetic and gas-phase study of the chemical vapor deposition of silicon carbide from C2H3SiCl3/H2. J. Ind. Eng. Chem. 94, 145–158.

Deshpande, S., Lengiewicz, J., Bordas, S.P., 2022. Probabilistic deep learning for real-time large deformation simulations. Comput. Methods Appl. Mech. Engrg. 398, 115307.

Dsilva, C.J., Talmon, R., Coifman, R.R., Kevrekidis, I.G., 2018. Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study. Appl. Comput. Harmon. Anal. 44 (3), 759–773.

Duan, Z., Kravaris, C., 2020. Reduced-order nonlinear observer design for two-time-scale systems. IFAC-PapersOnLine 53 (2), 5922–5927.

Evangelou, N., Dietrich, F., Chiavazzo, E., Lehmberg, D., Meila, M., Kevrekidis, I.G., 2023. Double diffusion maps and their latent harmonics for scientific computations in latent space. J. Comput. Phys. 485, 112072.

Evangelou, N., Wichrowski, N.J., Kevrekidis, G.A., Dietrich, F., Kooshkbaghi, M., McFann, S., Kevrekidis, I.G., 2022. On the parameter combinations that matter and on those that do not: data-driven studies of parameter (non) identifiability. PNAS Nexus 1 (4), pgac154.

Everson, R., Sirovich, L., 1995. Karhunen–Loeve procedure for gappy data. J. Opt. Soc. Amer. A 12 (8), 1657–1664.

Fowlkes, C., Belongie, S., Malik, J., 2001. Efficient spatiotemporal grouping using the nystrom method. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1. CVPR 2001, IEEE, p. I.

Giovanis, D.G., Shields, M.D., 2020. Data-driven surrogates for high dimensional models using Gaussian process regression on the grassmann manifold. Comput. Methods Appl. Mech. Engrg. 370, 113269.

Gkinis, P., Koronaki, E., Skouteris, A., Aviziotis, I., Boudouvis, A., 2019. Building a data-driven reduced order model of a chemical vapor deposition process from low-fidelity CFD simulations. Chem. Eng. Sci. 199, 371–380.

Gleason, K.K., 2020. Chapter 5 - Fluoropolymers by initiated chemical vapor deposition (iCVD). In: Ameduri, B., Fomin, S. (Eds.), Opportunities for Fluoropolymers. In: Progress in Fluorine Science, Elsevier, pp. 113–135. http://dx.doi.org/10.1016/B978-0-12-821966-9.00005-5.

Holiday, A., Kooshkbaghi, M., Bello-Rivas, J.M., Gear, C.W., Zagaris, A., Kevrekidis, I.G., 2019. Manifold learning for parameter reduction. J. Comput. Phys. 392, 419–431.

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

Kazantzis, N., Huynh, N., Wright, R.A., 2005. Nonlinear observer design for the slow states of a singularly perturbed system. Comput. Chem. Eng. 29 (4), 797–806.

Kazantzis, N., Kravaris, C., 1998. Nonlinear observer design using Lyapunov's auxiliary theorem. Systems Control Lett. 34 (5), 241–247.

Khatibi, S., Cassol, G.O., Dubljevic, S., 2021. Model predictive control of a non-isothermal axial dispersion tubular reactor with recycle. Comput. Chem. Eng. 145, 107159.

Koronaki, E., Cheimarios, N., Laux, H., Boudouvis, A., 2014. Non-axisymmetric flow fields in axisymmetric CVD reactor setups revisited: influence on the film's non-uniformity. ECS Solid State Lett. 3 (4), P37.

Koronaki, E., Gakis, G., Cheimarios, N., Boudouvis, A., 2016. Efficient tracing and stability analysis of multiple stationary and periodic states with exploitation of commercial CFD software. Chem. Eng. Sci. 150, 26–34.

Koronaki, E., Gkinis, P., Beex, L., Bordas, S., Theodoropoulos, C., Boudouvis, A., 2019. Classification of states and model order reduction of large scale Chemical Vapor Deposition processes with solution multiplicity. Comput. Chem. Eng. 121, 148–157. http://dx.doi.org/10.1016/j.compchemeng.2018.08.023.

Luenberger, D.G., 1964. Observing the state of a linear system. IEEE Trans. Mil. Electron. 8 (2), 74–80.

Martin-Linares, C.P., Psarellis, Y.M., Karapetsas, G., Koronaki, E.D., Kevrekidis, I.G., 2023. Physics-agnostic and physics-infused machine learning for thin films flows: modeling, and predictions from small data. arXiv:2301.12508.

Nadler, B., Lafon, S., Coifman, R.R., Kevrekidis, I.G., 2006. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. Appl. Comput. Harmon. Anal. 21 (1), 113–127.

Nishinaka, H., Nagaoka, T., Kajita, Y., Yoshimoto, M., 2021. Rapid homoepitaxial growth of (010) β-Ga2O3 thin films via mist chemical vapor deposition. Mater. Sci. Semicond. Process. 128, 105732.

Nyström, E.J., 1929. Über die praktische auflösung von linearen integralgleichungen mit anwendungen auf randwertaufgaben der potentialtheorie. Akademische Buchhandlung.

Papavasileiou, P., Koronaki, E.D., Pozzetti, G., Kathrein, M., Czettl, C., Boudouvis, A.G., Bordas, S.P., 2023. Equation-based and data-driven modeling strategies for industrial coating processes. Comput. Ind. 149, 103938.

Papavasileiou, P., Koronaki, E.D., Pozzetti, G., Kathrein, M., Czettl, C., Boudouvis, A.G., Mountziaris, T., Bordas, S.P., 2022. An efficient chemistry-enhanced CFD model for the investigation of the rate-limiting mechanisms in industrial chemical vapor deposition reactors. Chem. Eng. Res. Des. 186, 314–325.

Park, J.-K., Shin, D.-R., Chung, T.M., 2002. Dynamic observers for linear time-invariant systems. Automatica 38 (6), 1083–1087.

Patel, N., Corbett, B., Mhaskar, P., 2021. Model predictive control using subspace model identification. Comput. Chem. Eng. 149, 107276.

Psarellis, G.M., Aviziotis, I.G., Duguet, T., Vahlas, C., Koronaki, E.D., Boudouvis, A.G., 2018. Investigation of reaction mechanisms in the chemical vapor deposition of Al from DMEAA. Chem. Eng. Sci. 177, 464–470.

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Spencer, R., Gkinis, P., Koronaki, E., Gerogiorgis, D., Bordas, S., Boudouvis, A., 2021. Investigation of the chemical vapor deposition of Cu from copper amidinate through data driven efficient CFD modelling. Comput. Chem. Eng. 149, 107289.

Whitney, H., 1936. Differentiable manifolds. Ann. of Math. 645–680.

Willcox, K., 2006. Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition. Comput. & Fluids 35 (2), 208–226.

Xing, X., Dao, M.H., Zhang, B., Lou, J., Tan, W.S., Cui, Y., Khoo, B.C., 2022. Fusing sensor data with CFD results using gappy POD. Ocean Eng. 246, 110549. http://dx.doi.org/10.1016/j.oceaneng.2022.110549.

Xing, W., Triantafyllidis, V., Shah, A.A., Nair, P., Zabaras, N., 2016. Manifold learning for the emulation of spatial fields from computational models. J. Comput. Phys. 326, 666–690.

Xue, Y., Ludovice, P.J., Grover, M.A., Nedialkova, L.V., Dsilva, C.J., Kevrekidis, I.G., 2013. State reduction in molecular simulations. Comput. Chem. Eng. 51, 102–110.