



Discovering deposition process regimes: Leveraging unsupervised learning for process insights, surrogate modeling, and sensitivity analysis[☆]

Geremy Loachamín-Suntaxi^{a,b}, Paris Papavasileiou^{a,b}, Eleni D. Koronaki^a,
Dimitrios G. Giovanis^c, Georgios Gakis^b, Ioannis G. Aviziotis^b, Martin Kathrein^d,
Gabriele Pozzetti^d, Christoph Czettel^e, Stéphane P.A. Bordas^{a,*}, Andreas G. Boudouvis^b

^a Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, L-4364, Luxembourg

^b School of Chemical Engineering, National Technical University of Athens, Zographos Campus, 15780, Attiki, Greece

^c Department of Civil & Systems Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

^d CERATIZIT Luxembourg S.à r.l. Mamer, L-8201, Luxembourg

^e CERATIZIT Austria GmbH Reutte, A-6600, Austria

ARTICLE INFO

Keywords:

Arrhenius plot
Fe CVD reactor
Hierarchical clustering
Polynomial chaos expansion
Uncertainty quantification
Sensitivity analysis

ABSTRACT

This work introduces a comprehensive approach utilizing data-driven methods to elucidate the deposition process regimes in Chemical Vapor Deposition (CVD) reactors and the interplay of physical mechanism that dominate in each one of them. Through this work, we address three key objectives. Firstly, our methodology relies on process outcomes, derived by a detailed CFD model, to identify clusters of “outcomes” corresponding to distinct process regimes, wherein the relative influence of input variables undergoes notable shifts. This phenomenon is experimentally validated through Arrhenius plot analysis, affirming the efficacy of our approach. Secondly, we demonstrate the development of an efficient surrogate model, based on Polynomial Chaos Expansion (PCE), that maintains accuracy, facilitating streamlined computational analyses. Finally, as a result of PCE, sensitivity analysis is made possible by means of Sobol’ indices, that quantify the impact of process inputs across identified regimes.

The insights gained from our analysis contribute to the formulation of hypotheses regarding phenomena occurring beyond the transition regime. Notably, the significance of temperature even in the diffusion-limited regime, as evidenced by the Arrhenius plot, suggests activation of gas phase reactions at elevated temperatures. Importantly, our proposed methods yield insights that align with experimental observations and theoretical principles, aiding decision-making in process design and optimization. By circumventing the need for costly and time-consuming experiments, our approach offers a pragmatic pathway toward enhanced process efficiency. Moreover, this study underscores the potential of data-driven computational methods for innovating reactor design paradigms.

1. Introduction

Vapor deposition processes have been extensively used for the production of a wide range of materials in the form of coatings and thin films [1–3]. During the last decades, the emergence of nanotechnology has extended the use of such techniques for the production of semiconductors [4], optoelectronics [5], high-k dielectrics [6], functional powder coatings [7], 2D materials [8], as well as nanostructures, such as carbon nanotubes [9]. There are several deposition techniques such

as Physical Vapor Deposition (PVD) [10], or Atomic Layer Deposition (ALD) [11], nevertheless, Chemical Vapor Deposition (CVD) is often preferred due to its versatility, high throughput, relatively low cost, and better control over the deposition uniformity and conformality [1,12].

Besides the wide use and advantages of CVD, the process design and scale-up for industrial applications is often challenging, owing to the complex nature of the different mechanisms and phenomena that

[☆] This research was funded by the Luxembourg National Research Fund (FNR), grant reference 16758846.

* Correspondence to: School of Chemical Engineering, National Technical University of Athens, Zographos Campus, 15780, Attiki, Greece.

E-mail addresses: geremy.loachamin@uni.lu (G. Loachamín-Suntaxi), paris.papavasileiou@uni.lu (P. Papavasileiou), eleni.koronaki@uni.lu (E.D. Koronaki), dgiovan1@jhu.edu (D.G. Giovanis), gakisg@chemeng.ntua.gr (G. Gakis), javiziot@chemeng.ntua.gr (I.G. Aviziotis), Martin.Kathrein@ceratizit.com (M. Kathrein), Gabriele.Pozzetti@plansee-group.com (G. Pozzetti), Christoph.Czettel@ceratizit.com (C. Czettel), stephane.bordas@uni.lu (S.P.A. Bordas), boudouvi@chemeng.ntua.gr (A.G. Boudouvis).

<https://doi.org/10.1016/j.cej.2024.100667>

Received 11 September 2024; Received in revised form 20 October 2024; Accepted 21 October 2024

Available online 26 October 2024

2666-8221/2024 © 2024 Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

take place during the process [1,13]. A typical CVD process incorporates precursor gas flows and mass transfer, precursor reactions and decomposition in the gas phase, as well as surface phenomena, such as adsorption, desorption and surface reactions [14]. The interplay between these phenomena may affect the rate, uniformity and conformity of the deposition [5,15,16], as well as the decomposition of the deposited product, thus having an adverse effect on its properties [9]. As the contributions of these mechanisms and phenomena depend on the process conditions, a specified range of conditions is often provided for the efficient deposition of films with desired properties.

In CVD processes, the effect of process conditions is typically studied using the Arrhenius plot, showing the deposition rate as a function of temperature. Usually, two regimes are identified [1]: (i) the reaction-limited regime, where the reactions are slow and hence the rate-limiting step, and (ii) the diffusion-limited regime, where the mass transfer is slower than the reactions and therefore is the rate-limiting step. In the reaction-limited regime, the deposition rate increases with temperature, while in the diffusion-limited regime, which occurs above a certain critical temperature, the deposition rate can be unaffected or slightly decrease with a further temperature increase.

Most of the recipes used in the industrial application of such processes are based on the selection of process conditions that lead to the required regime. However, the regime definition from the Arrhenius plot is not always straightforward. Several studies have shown that other mechanisms may limit the deposition at higher temperatures besides the precursor mass transfer, such as precursor depletion [12], pre-decomposition in the gas phase [17], desorption [15,18], or catalyst deactivation [9,19], among others. Often, the number of regimes is larger, with transition regimes also identified [17,18]. Therefore, the choice of the appropriate process conditions requires the identification and mapping of these regimes on the process parameter space. As these complex regimes are an outcome of the interplay between various mechanisms, this demands a thorough study of the effect of process parameters on these mechanisms, in order to obtain optimal sets of parameters.

The rigorous investigation of such a complex process is an experimentally challenging task. Therefore, computational modeling of CVD processes has been employed to provide an insight of the different mechanisms that take place [19–21]. Using combined experimental and computational approaches, it is possible not only to identify the mechanisms and phenomena that determine the deposition process [12, 15], but also to study the effect of process parameters on such mechanisms [19], in order to provide a process parameter space for optimal process design. Nevertheless, although combined experimental and computational mechanistic approaches can provide a complete insight and understanding of CVD processes, they remain time-consuming, and demanding in terms of experimental and computational cost, as well as in terms of scientific information required as input, such as chemical reactions, their rate constants, and activation energies. Therefore, the efficient identification of the process conditions that correspond to each regime in the Arrhenius plot remains an open problem.

Data-driven approaches have been introduced, to provide predictive and robust models for CVD processes [22,23]. Such models may use either experimental or computational data [24] as input in order to predict process outputs or parameters [25]. The data-driven models have lower computational cost than the mechanistic models [22,26], while they do not require specific information (such as reaction parameters) as input to complement the data. Along these lines, uncertainty quantification and sensitivity analysis of complex systems in the context of chemical processes, as well as data-driven surrogate models have become outstanding methods for emulating such systems with high accuracy and low computational cost, and for providing physical insight [27,28]. Although these methods have been extensively studied, when faced with applications related to data-driven surrogate modeling for industrial-scale processes, some important obstacles and challenges remain to be addressed.

1. The amount and quality of data collected (either by experiments or simulations). In industrial-scale experiments, these (high-fidelity) data can be extremely costly to obtain, leading to a reduced amount of information. Moreover, as far as computational models are concerned, a factor to be considered is the ratio between the number of variables and the number of simulations necessary to train a data-driven model, which generally implies a high computational cost.
2. Concerning the Arrhenius plots, there is lack of analytical criteria to distinguish the different regimes especially in industrial-scale processes. Furthermore, the imbalance of dataset size in the various regimes introduces further obstacles.
3. Proper quantification of the influence of input variability on the output responses of models that mimic complex industrial-scale processes to determine their importance in the process and to provide them with a physical sense that can serve as criteria for system optimization.

The ambition of this work is to propose a data-driven workflow based on unsupervised methods, in order to deliver predictions of the relative effect of the physical phenomena on the final process outcome. The goal is to avoid the time and cost involved in the line of experiments typically required for determining the Arrhenius plot of a deposition process. Moreover, an accurate and efficient surrogate modeling approach is presented, which further allows to perform sensitivity analysis of the input parameters. Here the validation of our model via comparison to experimental data, confirms the fundamental theory of Fe deposition. This sets stage for the ultimate goal, which is to develop a computational tool that enables efficient design of novel processes or the optimization of existing industrial set-ups, where data production are already available.

2. Methods

2.1. Process description and experimental set up

Fe deposition was conducted in a vertical, cylindrical, stagnant flow, cold wall CVD reactor which has been previously described in detail [18,29].

Independent experiments were conducted: (i) at 13 different surface temperature values, T_s , in the range 130–250 °C for a fixed pressure of $P = 10$ Torr, and (ii) at four different P , in the range 10–40 Torr, for a fixed $T_s = 180$ °C. The deposition duration for all experiments was 1 h. $10 \times 10 \times 1$ mm³ silicon coupons are cut from 4 in Si(100) wafers (Sil'tronix). They are cleaned in an ultrasonic bath with acetone and ethanol, they are dried under argon flux and baked in a furnace at 60 °C for 30 min to cast off humidity. The deposition rate is evaluated directly by weight difference of the substrates before and after deposition, using a microbalance (Sartorius) with ± 10 µg accuracy. Three independent weight measurements are carried out for each substrate before and after the experiment and an average value is calculated. The Fe precursor, Fe(CO)₅, is supplied by Sigma-Aldrich and Fischer Scientific and is used as received. Pure nitrogen (99,998%, Air Products) is fed through computer-driven mass flow controllers. During experiments, the lines and walls of the reactor are maintained at room temperature (regulated at 25 °C), while the N₂ dilution gas flow and the N₂ carrier gas flow through the precursor equal 302 and 3 standard cubic centimeters per minute (sccm), respectively.

2.2. CFD model and data collection

A CFD model was developed based on the work of Aviziotis et al. [18,30] for the CVD of Fe from Fe(CO)₅. The original CFD model [18] is 3D, its equations are discretized using the finite volume method and solved using Fluent CFD code. In this work, we construct a 2D-axisymmetric CFD model, the equations are then discretized using 13 987 finite elements and solved using COMSOL Multiphysics®.

Table 1

Gas phase reactions along with their pre-exponential factors and activation energies. The forward reactions are denoted as $G1 - G4$ and the reverse reactions as $G1' - G3'$.

ID	Reaction	Pre-exponential factor (k_0)	Activation energy (E_a) [kJ mol ⁻¹]
G1	$\text{Fe}(\text{CO})_5 \rightarrow \text{Fe}(\text{CO})_4 + \text{CO}$	$9.65 \cdot 10^{12} \text{ s}^{-1}$	136.7
G2	$\text{Fe}(\text{CO})_4 \rightarrow \text{Fe}(\text{CO})_3 + \text{CO}$	$8.96 \cdot 10^{12} \text{ s}^{-1}$	79.9
G3	$\text{Fe}(\text{CO})_3 \rightarrow \text{Fe}(\text{CO})_2 + \text{CO}$	$1.25 \cdot 10^{11} \text{ s}^{-1}$	97.5
G4	$\text{Fe}(\text{CO})_2 \rightarrow \text{FeCO} + \text{CO}$	$3.96 \cdot 10^{11} \text{ s}^{-1}$	139.1
G1'	$\text{Fe}(\text{CO})_4 + \text{CO} \rightarrow \text{Fe}(\text{CO})_5$	$3.5 \cdot 10^7 \text{ m}^3 \text{ kmol}^{-1} \text{ s}^{-1}$	10.5
G2'	$\text{Fe}(\text{CO})_3 + \text{CO} \rightarrow \text{Fe}(\text{CO})_4$	$1.3 \cdot 10^{10} \text{ m}^3 \text{ kmol}^{-1} \text{ s}^{-1}$	9.5
G3'	$\text{Fe}(\text{CO})_2 + \text{CO} \rightarrow \text{Fe}(\text{CO})_3$	$1.8 \cdot 10^{10} \text{ m}^3 \text{ kmol}^{-1} \text{ s}^{-1}$	9.5

Table 2

Surface reactions along with their pre-exponential factors and activation energies.

ID	Reaction	Pre-exponential factor (k_0)	Activation energy (E_a) [kJ mol ⁻¹]
S1	$\text{Fe}(\text{CO})_5 \rightarrow \text{Fe}_{(s)} + 5\text{CO}_{(g)}$	$6.00 \cdot 10^2 \text{ m s}^{-1}$	27.9
S2	$\text{Fe}(\text{CO})_3 \rightarrow \text{Fe}_{(s)} + 3\text{CO}_{(g)}$	$5.30 \cdot 10^7 \text{ m s}^{-1}$	75.3
S3	$\text{Fe}(\text{CO}) \rightarrow \text{Fe}_{(s)} + \text{CO}_{(g)}$	$3.70 \cdot 10^{10} \text{ m s}^{-1}$	19.3
S4	CO adsorption	$1.90 \cdot 10^{11} \text{ Torr}^{-1}$	89.9

The governing equations include the conservation of mass, momentum, and energy, along with the conservation equations of the chemical species that partake in the chemical reactions. For more details, the interested reader is referred to [30].

A constant mass inflow rate of $5.9177 \cdot 10^{-6} \text{ kg/s}$ is imposed at the inlet of the reactor. The inlet stream consists solely of N_2 and the precursor ($\text{Fe}(\text{CO})_5$). At the outlet boundary, the pressure is set equal to the operating pressure of the reactor. A no-slip condition is set for all of the inside walls of the reactor. The flow is considered laminar.

A constant temperature is set for the substrate surface (T_s), the side walls of the reactor ($T_w = 25^\circ\text{C}$), and the gas reactant lines ($T_l = 25^\circ\text{C}$). The ideal gas assumption is made for calculating the density of the gas mixture. The species' heat capacity, molar enthalpy, and molar entropy are calculated using the relevant CHEMKIN coefficients and NASA polynomials. Furthermore, a Stefan–Maxwell diffusion model is used to compute the diffusion fluxes.

The gas phase reactions are summarized in Table 1. They consist of four decomposition (forward) reactions ($G1 - G4$) and three recombination (reverse) reactions ($G1' - G3'$). The reaction rates for the forward reactions (R_{G_i}) can be calculated as such:

$$R_{G_i} = k_{0,G_i} \exp(-E_{a,G_i}/RT) C_{\text{Fe}(\text{CO})_i}, \quad (2.1)$$

and for the reverse reactions ($R_{G'_i}$):

$$R_{G'_i} = k_{0,G'_i} \exp(-E_{a,G'_i}/RT) C_{\text{Fe}(\text{CO})_i} C_{\text{CO}}. \quad (2.2)$$

The surface reactions that ultimately lead to the deposition of Fe are summarized in Table 2. For the rates of these reactions (R_{S_i}), a Langmuir–Hinshelwood type kinetic expression is implemented, given by:

$$R_{S_i} = k_{0,S_i} \exp(-E_{a,S_i}/RT_s) C_{\text{Fe}(\text{CO})_i} / (1 + k_{\text{CO}} \exp(-E_{a,\text{CO}}/RT_s) P_{\text{CO}}). \quad (2.3)$$

The kinetic parameters of the model are fitted to the experimental data [30] for temperatures between $130 - 250^\circ\text{C}$ and an operating pressure of $P = 10 \text{ Torr}$.

For pressures higher than 10 Torr, the pre-exponential factor of S1 is adjusted:

$$k_{0,S_1, \text{adapted}} = 600 [\text{m/s}] 20 [\text{Torr}^2] / P^2. \quad (2.4)$$

Following model validation, data are collected for two parameters of interest:

1. The temperature of the deposition surface (T_s).
2. The operating pressure of the reactor (P).

Table 3

Range of parameters for data-collection purposes.

Parameter	Range	Units	# of values	Sampling
T_s	[130, 250]	$^\circ\text{C}$	601	Uniform distribution
P	[10, 40]	Torr	4	Uniform distribution

Computational experiments are conducted for T_s varying from $130 - 250^\circ\text{C}$ and P at values of 10, 20, 30 and 40 Torr. The design of experiments (DoE) is summarized in Table 3 below:

This DoE leads to a total of 2404 computational experiments.

2.3. Data-driven workflow

The data-collection process involves collecting data from the CFD model based on the parameter values reported in Table 3. At each combination of parameters,¹ the values for the solved variables at each discretization point are arranged in columns, as shown schematically in Fig. 2. The proposed data-driven workflow is detailed in Section 2.3.5. But first, we provide a brief overview of the methods applied in this work.

2.3.1. Dimensionality reduction using Principal component analysis

Principal component analysis (PCA) is a method widely used for reduction of the dimension of a scaled dataset $\tilde{\mathbf{Y}}$ (containing numerical simulations or experimental measurements) [31,32]. Other alternatives include nonlinear methods, such as Diffusion Maps [25], UMAP [33,34] and kernel PCA [35,36], among others. The analysis presented in this work is not restricted by the choice of dimensionality reduction method and it can be implemented in conjunction with whichever one is best suited for the application. We select PCA which seeks to find a suitable low-dimensional space, based on the *singular value decomposition* (SVD) of the covariance matrix of $\tilde{\mathbf{Y}}$, $\text{cov}(\tilde{\mathbf{Y}})$ (see Fig. 3), to represent the original high-dimensional data while capturing its total variability. PCA has been successfully applied for dimensionality reduction in the context of CVD processes [37–39]. The key-ingredients of PCA are presented in detail in Appendix.

2.3.2. Unsupervised learning using hierarchical (agglomerative) clustering

Clustering algorithms function by evaluating the similarity or dissimilarity among data points, thereby forming distinct clusters. The aim is to group similar observations to uncover hidden patterns within the dataset. Generally, these algorithms utilize a dissimilarity matrix

¹ The set of parameter combinations constitute the experimental design, denoted by $\mathcal{X}$.

as input, which details pairwise dissimilarities among observations. For quantitative data, Euclidean distance is the prevalent metric, representing the shortest path between two points in feature space. Nonetheless, other distance measures can be applied, potentially altering clustering outcomes [40,41].

Clustering algorithms categorize into various types based on methodologies. Partitional methods, like k-means clustering, allocate observations into clusters iteratively using centroid distances, which requires prior determination of cluster numbers and is sensitive to initial conditions [42].

Density-based methods, such as OPTICS and DBSCAN, define clusters by locating dense regions separated by sparse ones, with fundamental parameters being the minimum point count for a cluster and the minimum distance among core points [43–45].

Hierarchical clustering (HC) is another unsupervised machine learning method used to group data based on the similarities between elements in an agglomerative manner [46]. The algorithm creates a hierarchical nested tree, known as a dendrogram, from this grouping process (see Fig. 4). At the bottom of this tree, we have individual clusters: one cluster for each dataset element. Then, we merge pairs of clusters and order them hierarchically until we reach the top of the tree, where a single cluster contains all the elements. A cut-off of the dendrogram is then chosen at a certain distance level to form several homogeneous groups with common characteristics.

To determine the distance between groups, we can use several similarity measures depending on the elements of the dataset, such as cosine similarity [47], Manhattan and Euclidean distances [48,49]. In this work, we considered the latter distance because the linkage is performed using the Ward method [50,51], which minimizes the sum of the squared distances between cluster centers.

2.3.3. Polynomial chaos expansion

Polynomial chaos expansion (PCE) [52–55] is a surrogate² modeling technique that approximates the behavior/response of complex physics-based models [56–58]. What makes PCE unique is that it can accelerate tasks such as uncertainty quantification (UQ) and sensitivity analysis [59–61].

In the model described in Section 2.2, the conservation equations for momentum, energy, and mass are solved for two input parameters: the substrate temperature (T_s) and the chamber pressure (P). We express the response of the model as:

$$\mathbf{y} = \mathcal{M}(\mathbf{x}), \quad (2.5)$$

where $\mathcal{M}(\cdot)$ denotes the CFD model, $\mathbf{x} = (x_{T_s}, x_P) \in \mathbb{R}^2$ is the vector containing the values of the two input parameters, and $\mathbf{y} = (y_1, y_2, \dots, y_N) \in \mathbb{R}^N$ contains the model responses (see Fig. 2), with $y_i \in \mathbb{R}$ for $i = 1, \dots, N$, and N denoting the dimension of the domain discretization. In the subsequent analysis, we only consider two precursor concentrations as output variables, i.e., $\mathbf{y} \in \mathbb{R}^2$ (see Section 3 for more details).

Polynomial chaos computes an approximation of the response, $\bar{\mathbf{y}}$, using a set of multi-dimensional polynomials that are orthogonal to the probability density function (PDF) of the input random variables $\mathbf{x} = (x_{T_s}, x_P)$. PCE was initially designed for variables following a standard Gaussian distribution, employing Legendre polynomials [57]. Subsequently, it was extended to other standard statistical distributions (e.g., uniform and beta distributions utilizing Hermite and Jacobi polynomials, respectively [62,63]), or even to arbitrary distributions [64, 65].

² A surrogate or metamodel is a mathematical model that learns the mapping between the input and the output of a computational model.

Thus, for the vector-value model $\mathcal{M}$, we construct a surrogate model using PCE, denoted as $\mathcal{M}^{PCE}$, in which each component of the model response at each spatial location is represented as follows:

$$\mathcal{M}^{PCE}(\mathbf{x}) = \sum_{\alpha \in \mathbb{N}^2} c_\alpha \boldsymbol{\psi}_\alpha(\mathbf{x}), \quad (2.6)$$

where $\boldsymbol{\psi}_\alpha(\cdot)$ represents a multivariate polynomial³ that is orthonormal⁴ with respect to the PDF of the input parameters, $\alpha = (\alpha_1, \alpha_2) \in \mathbb{N}^2$ is a multi-index that indicates the polynomial degree of the components of $\boldsymbol{\psi}_\alpha$, and c_α are the associated coefficients that quantify the relationship between the polynomial expansion $\mathcal{M}^{PCE}$ and the computational model $\mathcal{M}$.

In practice, the expansion (2.6) needs to be truncated to a finite sum $\mathcal{M}_d^{PCE}$, where $d \in \mathbb{N}$ is the maximum degree of the polynomials in the expansion:

$$\mathcal{M}^{PCE}(\mathbf{x}) \approx \mathcal{M}_d^{PCE}(\mathbf{x}) + \varepsilon_d = \sum_{\alpha \in A_d \subset \mathbb{N}^2} c_\alpha \boldsymbol{\psi}_\alpha(\mathbf{x}) + \varepsilon_d, \quad (2.7)$$

where ε_d denotes the error introduced due to the truncation of the infinite series, and A_d is a set of multi-indices defined as $A_d = \{\alpha \in \mathbb{N}^2 : |\alpha| = \alpha_1 + \alpha_2 \leq d\}$.

Next, we estimate the coefficients c_α while minimizing the error ε_d [66]. This minimization problem can be solved, for instance, using nonintrusive methods such as least squares regression (LSR) [67], LASSO regression [68] and least angle regression (LAR) [69].

Furthermore, to measure the performance of the surrogate model and to avoid overfitting [70], we use the *leave-one-out* (LOO) cross-validation error ε_{LOO} [71], which can be implemented as follows:

$$\varepsilon_{LOO} = \sum_{i=1}^m \left[\mathcal{M}_d^{PCE}(\mathbf{x}^{(i)}) - \mathcal{M}_d^{PCE \setminus i}(\mathbf{x}^{(i)}) \right]^2 / \text{Var}[\mathcal{M}_d^{PCE}(\mathcal{X})], \quad (2.8)$$

where $\mathbf{x}^{(i)}$ is an input vector that belongs to the experimental design $\mathcal{X}$ (see Section 2.3), and $\mathcal{M}_d^{PCE \setminus i}$ is the PCE constructed on the reduced experimental design that excludes $\mathbf{x}^{(i)}$.

2.3.4. PCE-based sensitivity analysis (Sobol' indices)

Sobol' indices [72] are widely used for sensitivity analysis as they provide a robust measure of the input parameters' influence on the output variability. This understanding of parameter importance is particularly valuable in chemical applications [73,74], as it offers crucial physical insights into the process.

Sobol' indices decompose the variance of the model output into the sum of the contributions of individual inputs. As discussed in Section 2.3.3, the PCE surrogate model [55,75,76] facilitates sensitivity analysis by utilizing the orthogonality of the polynomial basis to estimate the statistical moments of the model response by rewriting expression (2.7), as follows:

$$\mathcal{M}_d^{PCE}(\mathbf{x}) = c_0 + \sum_{\alpha \in A_d \setminus \{\mathbf{0}\}} c_\alpha \boldsymbol{\psi}_\alpha(\mathbf{x}),$$

then, its mean and variance are given by

$$\mathbb{E}[\mathcal{M}_d^{PCE}(\mathbf{x})] = c_0 \quad \text{and} \quad \text{Var}(\mathcal{M}_d^{PCE}(\mathbf{x})) = \sum_{\alpha \in A_d \setminus \{\mathbf{0}\}} c_\alpha^2. \quad (2.9)$$

Thus, we can employ the expressions for the mean and variance above to perform sensitivity analysis using the first-order and total-order Sobol' indices [76]. However, in the following analysis, we are only interested in the total-order Sobol' indices (S^T) defined for our two input parameters as

$$S_{T_s}^T = \sum_{\alpha \in A_d, \alpha_1 \neq 0} c_\alpha^2 / \sum_{\alpha \in A_d \setminus \{\mathbf{0}\}} c_\alpha^2,$$

³ Defined as the product of the univariate polynomials: $\boldsymbol{\psi}_\alpha(\mathbf{x}) = \psi_{\alpha_1}^{(T)}(x_{T_s}) \psi_{\alpha_2}^{(P)}(x_P)$.

⁴ $\mathbb{E}_x[\boldsymbol{\psi}_\alpha(\mathbf{x}) \boldsymbol{\psi}_\beta(\mathbf{x})] = \delta_{\alpha\beta}$, where $\delta_{\alpha\beta}$ is the Kronecker delta.

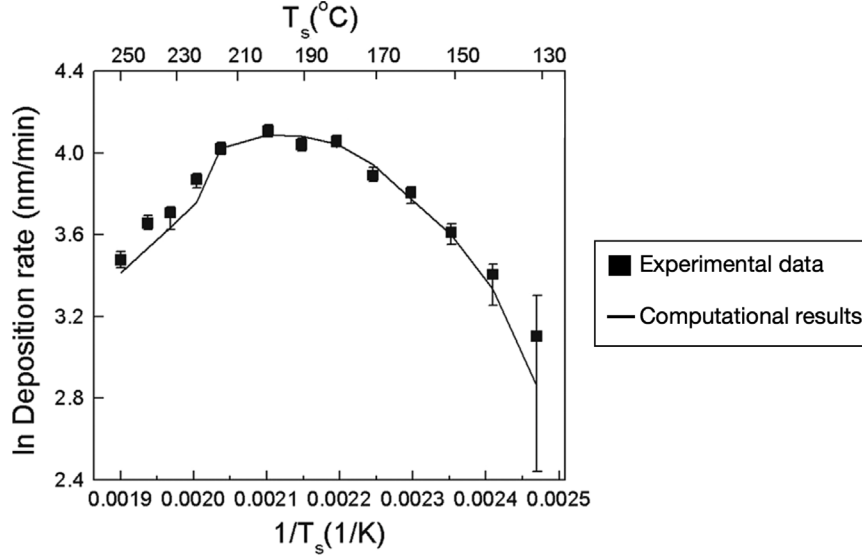


Fig. 1. The Arrhenius plot of a CVD reactor for the Fe from $\text{Fe}(\text{CO})_5$ and a fixed pressure $P = 10$ Torr. Experimental measurements (squares) and computational results (line) are shown. The drop in deposition rate is $T > 220$ °C is attributed to poisoning of the surface by CO [18].

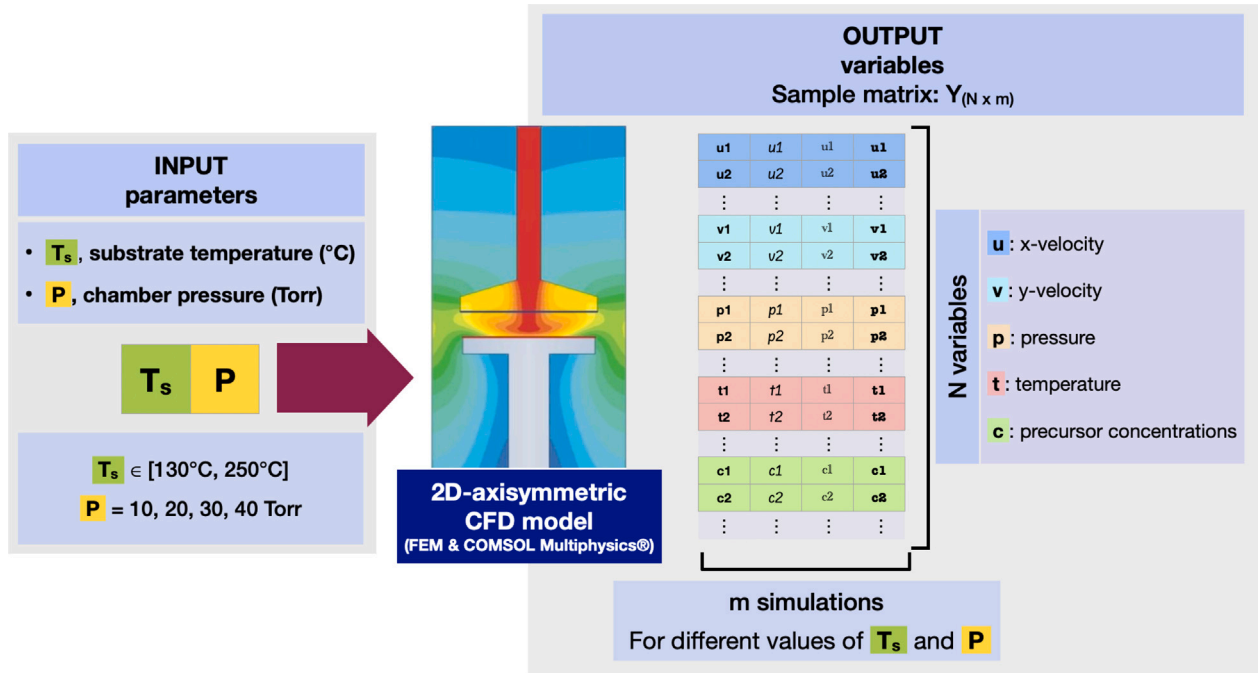


Fig. 2. The dataset comes from a CFD model for the Fe CVD reactor. The input parameters are the substrate temperature T_s and the chamber pressure P . This computational model simulates seven variables at each discretization point, including the distributions of the x-y velocities u and v , pressure p , temperature t , and three different precursor concentrations presented in Table 2.

$$S_P^T = \sum_{\alpha \in A_d, \alpha_2 \neq 0} c_\alpha^2 / \sum_{\alpha \in A_d \setminus \{0\}} c_\alpha^2.$$

These indices are written in terms of PCE coefficients which capture the total influence of the input parameters and their interactions on the variability of the output variables. When S^T is close to one, the contribution of the input parameter is highly significant, whereas values close to zero indicate a negligible influence.

2.3.5. Proposed workflow

In this section, we describe the proposed workflow illustrated in Fig. 5, where each step is enumerated according to the list below that presents a more detailed description:

1. *Set an input parameter space and generate the experimental design:* We perform deterministic sampling (DS) [77] as a computationally efficient technique for conducting UQ using data from

a CFD model [78]. This sampling technique also allows us to incorporate specific input parameter values into the analysis, for which we have experimental data obtained in [18], facilitating comparison with our results. Thus, given that the input parameters $\mathbf{x} = (x_{T_s}, x_P)$ belong to certain production ranges, we uniformly subdivide these ranges with a step size denoted as h_i (see Table 4) and consider all possible combinations of these two input values as experimental design $\mathcal{X}$.

Using the experimental design, we solve the CFD model and obtain a dataset containing the model responses at all points in the spatial discretization (see Fig. 2).

2. *Dimensionality reduction:* In this application we consider only the chemical species concentration in an area close to the deposition surface (see Fig. 6), since it is in this area where the critical

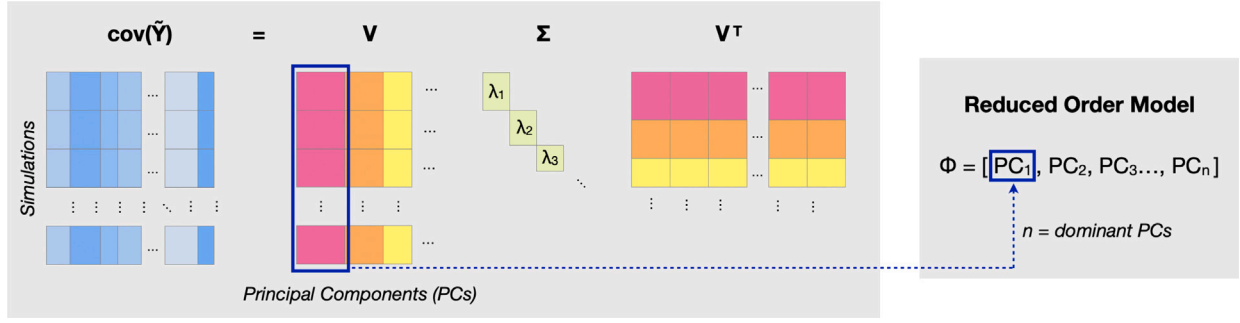


Fig. 3. PCA decomposes the covariance matrix of Y ($\text{cov}(\tilde{Y})$) into n dominant principal components (PCs): $(PC_1, PC_2, \dots, PC_n)$, from which we can obtain the reduced space ϕ .

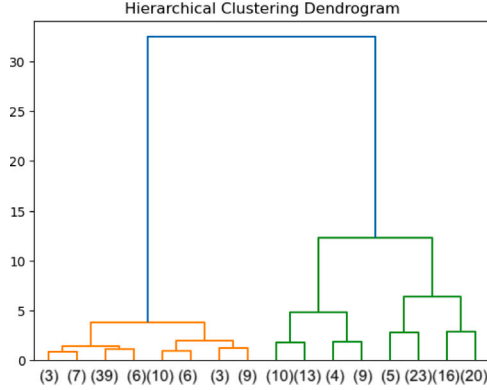


Fig. 4. HC gradually merges smaller clusters to form a nested tree structure (dendrogram) where each branch corresponds to a group (or cluster). The three largest clusters are colored, while the number of elements in each cluster is shown in parentheses at the bottom.

Table 4

Description of the input parameter space.

Parameter	Description	Range (units)	h_i
x_{T_e}	Substrate temperature	[130.0 °C, 250.0 °C]	0.2 °C
x_p	Chamber pressure	[10 Torr, 40 Torr]	10 Torr

on the polynomial coefficients. These moments are used for sensitivity analysis through the Sobol' indices, which explain the influence of input variability on the model responses. This allows us to identify important input parameters, as explained in Section 2.3.4.

Steps 2, 3, 4, and 5 were implemented in Python using the Sklearn and UQpy packages [79,80].

Finally, the parameters used in the models discussed in this section are summarized in Table 5.

3. Results and discussion

Inspired by the Arrhenius plot shown in Fig. 1, the results and analysis presented below are based on the concentration values of two different species: $\text{Fe}(\text{CO})_5$ and $\text{Fe}(\text{CO})$, as calculated by the CFD model described in Section 2.2 for the parameter values reported in Table 4. These species were chosen because their decomposition has the lowest activation energy (see Table 2). This implies that their contribution to the overall deposition rate, determined by an Arrhenius model, is considerably more significant than those obtained with other species.

Moreover, in this application it makes sense to consider data in the proximity to the deposition surface, shown in orange color in Fig. 6, since the species concentration in that area determines the deposition rate and its distribution.

3.1. Dimensionality reduction

For this initial stage, the region of interest is discretized by a computational mesh with 828 nodes, where the species concentration are calculated. The size of the data scales with the discretization of the physical domain, making dimensionality reduction necessary to keep the meaningful information in a low-dimensional representation, which will be used to build a more efficient surrogate model.

Specifically, PCA allows us to quantify the variability explained by each direction in which the data is projected. We repeat this exercise for different numbers of PCs as we want to examine the potential effects of dimensionality reduction in our subsequent analysis. Here we have considered retaining 3, 5, and 10 PCs. The evolution of the explained variability as a function of the number of retained PCs is shown in Fig. 7, where we observe that it is sufficient to retain three principal components for the established threshold to be exceeded with 99%.

On the other hand, when observing the behavior of the reconstruction error presented in Fig. 8 as a semi-log plot, it becomes clear that the error drops to $O(-3)$ with 3 PCs, and it drops below $O(-6)$ when at least 8 PCs are considered. This suggests that in this application at most 10 PCs are required to accurately reconstruct the data.

for deposition phenomena mostly take place; this still constitutes a high-dimensional problem because the model response $y \in \mathbb{R}^{2N}$, meaning its dimension is twice the number of nodes considered in the spatial discretization of that area. Then, to reduce the dimensionality of data while capturing its variability and to illustrate the clusters in step 3, we apply the PCA method discussed in Section 2.3.1. In the first case, where we reduce the dimensionality of the dataset, we retain the first n principal component modes that describe more than 99% of the data variability, where $n = 3, 5$ and 10 (See Table 6). In the second case, we retain 10 PCs, but for visualization purposes, we only show the case where we consider 3 PCs.

3. **Clustering and the Arrhenius plot:** With the low-dimensional representation of the data obtained in the previous step, we apply hierarchical clustering method to identify clusters that correspond to meaningful process regimes.
4. **Construct the PCE surrogate model:** Based on the distributions followed by the input parameters in each regime, the corresponding families of orthogonal polynomials are considered to construct the PCE model for each of the n retained principal components. The polynomial coefficients are estimated using a Linear Regression method, as discussed in Section 2.3.3. In addition, we analyze the evolution of the LOO-error, introduced in (2.8), for each polynomial degree, until reaching an optimal degree for which this error is sufficiently small (lower than order $O(-3)$). Additionally, it is important to mention that this surrogate model is twenty thousand times faster than the standalone CFD simulation for this model.
5. **PCE-based uncertainty quantification and sensitivity analysis:** At this stage, we use the PCE surrogate model to directly, and accurately, estimate the moments of the model responses based

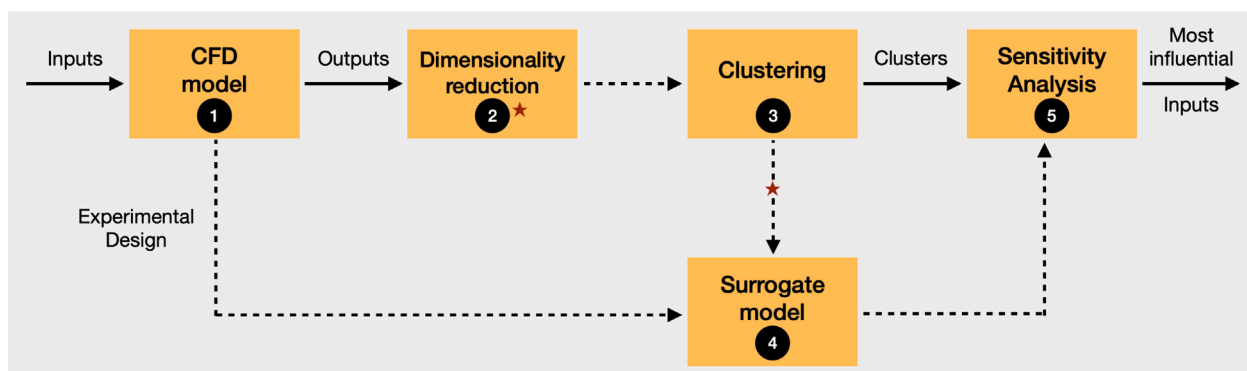


Fig. 5. Data-driven workflow using machine learning methods to distinguish regimes of the Arrhenius plot and identify the most influential input parameters. The numbers in each box correspond to the steps detailed in Section 2.3.5. The red star on the diagram indicates that it is possible to perform dimensionality reduction at either of the two stages, depending on the size of the dataset.

Table 5
Summary of parameter settings.

Model	Parameter	Setting	Selection criteria
Dimensionality reduction (PCA)	Scaling	Standard scaler	
	Number of components	Set to 3	Reconstruction error
Hierarchical clustering	Number of clusters	Set to 3	Experimental observations
	Linkage type	Ward method	Experimental observations
	Distance	Euclidean	Experimental observations
Surrogate model (PCE)	Polynomial basis	Legendre polynomials	Associated with the distributions followed by the input parameters
	Maximum degree	Set to 3	Cross-validation error
	Regression method	LSR	Cross-validation error

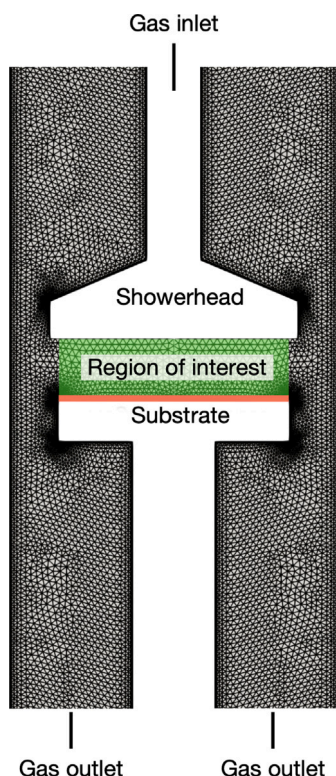


Fig. 6. Computational domain corresponding to the actual experimental reactor geometry. The discretization mesh is shown; in the data-driven analysis, only data from the nodes close to the deposition surface (green area) are used. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Clustering: identification of deposition regimes

In the analysis that follows, each chamber pressure P is treated separately, by implementing Hierarchical Clustering (refer to Section 2.3.2) to data collected in each value of chamber pressure. In each case, three entirely disjoint clusters are identified.

To visualize the clusters identified the data is projected on a lower-dimensional space, identified by the PCA method described in Section 2.3.1, as illustrated in Fig. 9. Within each group corresponding to different values of chamber pressure P , we can identify three separate clusters, each represented by a distinct color, determined by the average temperature of the substrate from which the values of each cluster are derived. It is worth noting that the clustering process focused solely on species concentrations, making it entirely oblivious of the input parameters of the computational model, i.e. the operational parameters T_s and P . This practical approach highlights the capability of HC to efficiently identify clusters of interest using only computational experiments.

As discussed in the previous section, we explore how the clustering results change as we increase the number of principal components, as depicted in Table 6. We examine cases with 3, 5, and 10 PCs, revealing the convergence of temperature ranges associated with regimes for each pressure value, which means that the clusters we obtained are insensitive to the number of principal components considered. Specifically, the results presented in Table 6, corresponding to $P = 30$ Torr and $P = 40$ Torr, indicate that the range associated with the intermediate regime (in the green rows) shows no discernible changes beyond the results obtained when considering 3 PCs or more. For $P = 20$ Torr, this range narrows for 5 PCs, but no further changes occur beyond that point. For the last case ($P = 10$ Torr), we observe that this intermediate range remains unchanged for a small number of PCs, but when considering 10 PCs, it tightens up slightly. Therefore, the subsequent analyses focuses on the scenario where 10 PCs are taken into account, ensuring the elimination of potential effects resulting from dimensionality reduction. Moreover, intervals corresponding to 10 PCs are colored in the fourth column of Table 6 to identify them with each regime.

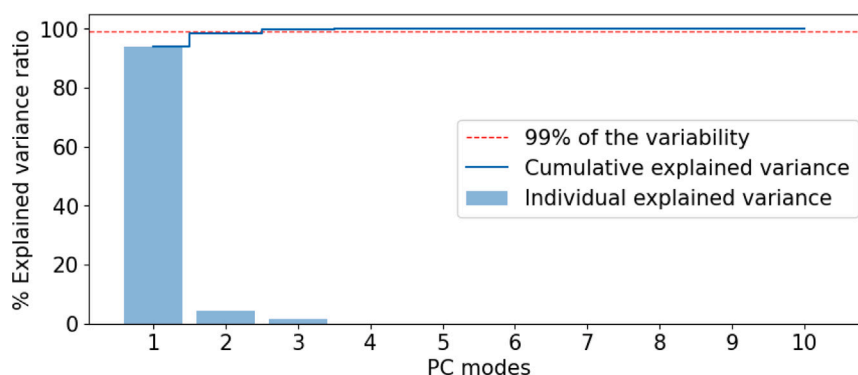


Fig. 7. Evolution of the individual (blue bars) and cumulative (blue line) explained variance as a function of the number of retained PC modes. In addition, a minimum threshold (red dashed line) is set at 99% of the total variability. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

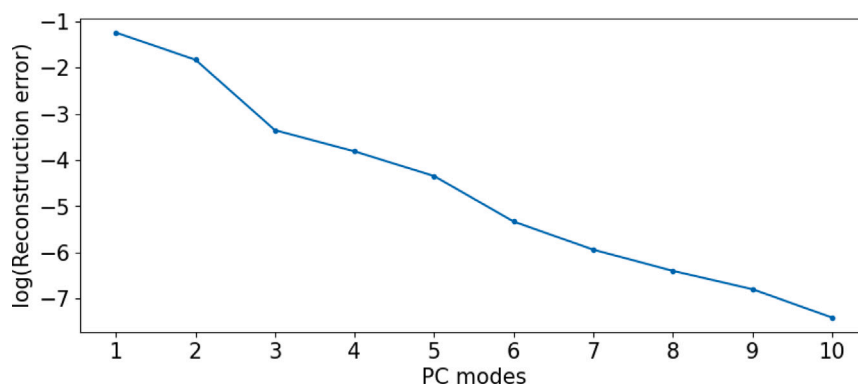


Fig. 8. Evolution of the reconstruction error (blue line) in a logarithmic scale as a function of the number of retained PC modes.

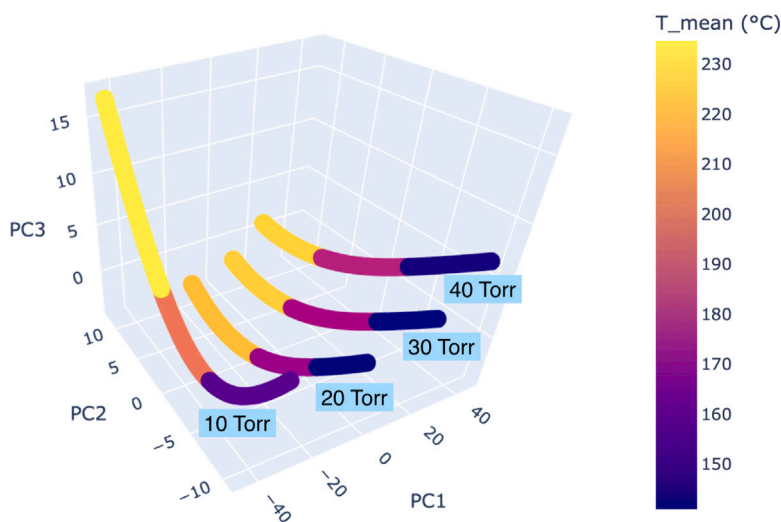


Fig. 9. The clusters are projected in a three-dimensional space using PCA, for each value of $P = 10, 20, 30$ and 40 Torr. They are colored according to the average substrate temperature T_s in $^{\circ}\text{C}$, following the temperature scale shown on the right, with blue representing the lowest temperature and yellow representing the highest ones. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We then proceed to compare the temperature range of the clusters for $P = 10$ Torr (refer to the cluster block at the bottom of Table 6) to the regimes of the experimentally determined Arrhenius plot (cf. Fig. 1). Indeed, upon analyzing these findings presented at the bottom of Table 6, it becomes apparent that Cluster 1 coincides with the temperature range for the reaction-limited regime, i.e. $130.0\text{--}183.2$ $^{\circ}\text{C}$. Ad-

ditionally, Cluster 2 coincides with the transition regime, which occurs for $183.4\text{--}219.0$ $^{\circ}\text{C}$, while Cluster 3 corresponds to the diffusion-limited regime, encompassing the span of $219.2\text{--}250.0$ $^{\circ}\text{C}$.

Notably, these values indicate a transition regime extending to slightly higher temperatures compared to the results presented in [18]. Specifically, if we rely visually on the Arrhenius plots shown in Fig. 10,

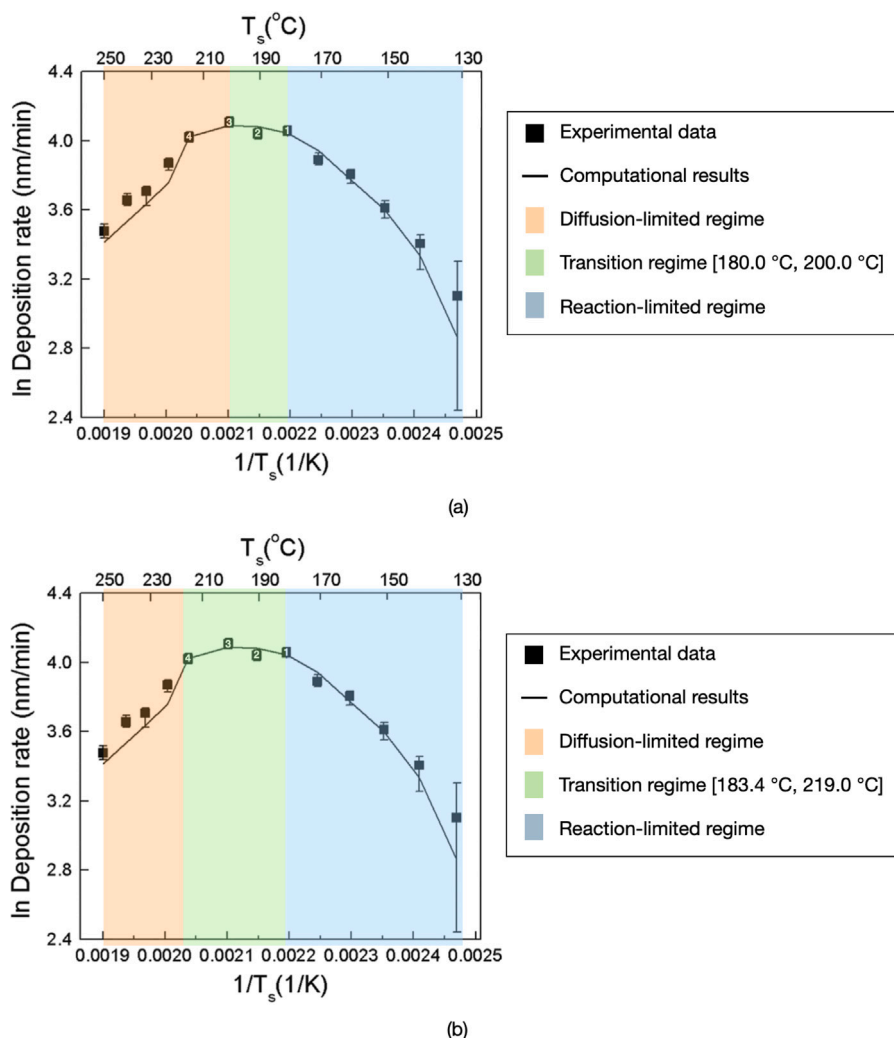


Fig. 10. The Arrhenius plot of the Fe-CVD reactor for $P = 10$ Torr. Three regions have been colored: reaction-limited regime (low temperatures) in blue, transition regime in green, and diffusion-limited regime (high temperatures) in orange. These regions were colored according to (a) the results presented in [18] and (b) temperature ranges associated with the clusters obtained in this study. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6

Temperature ranges associated with the clusters identified in Fig. 9 for each value of P arranged in ascending order.

	$T_{\min}$	$T_{\max}$	$T_{\min}$	$T_{\max}$	$T_{\min}$	$T_{\max}$	P
Cluster 3	203.4 °C	250.0 °C	203.4 °C	250.0 °C	203.4 °C	250.0 °C	40
Cluster 2	157.4 °C	203.2 °C	157.4 °C	203.2 °C	157.4 °C	203.2 °C	
Cluster 1	130.0 °C	157.2 °C	130.0 °C	157.2 °C	130.0 °C	157.2 °C	
Cluster 3	192.2 °C	250.0 °C	192.2 °C	250.0 °C	192.2 °C	250.0 °C	30
Cluster 2	153.8 °C	192.0 °C	153.8 °C	192.0 °C	153.8 °C	192.0 °C	
Cluster 1	130.0 °C	153.6 °C	130.0 °C	153.6 °C	130.0 °C	153.6 °C	
Cluster 3	221.2 °C	250.0 °C	191.0 °C	250.0 °C	191.0 °C	250.0 °C	20
Cluster 2	170.2 °C	221.0 °C	152.6 °C	190.8 °C	152.6 °C	190.8 °C	
Cluster 1	130.0 °C	170.0 °C	130.0 °C	152.4 °C	130.0 °C	152.4 °C	
Cluster 3	219.2 °C	250.0 °C	219.2 °C	250.0 °C	219.2 °C	250.0 °C	10
Cluster 2	184.2 °C	219.0 °C	184.2 °C	219.0 °C	183.4 °C	219.0 °C	
Cluster 1	130.0 °C	184.0 °C	130.0 °C	184.0 °C	130.0 °C	183.2 °C	
	3 PCs		5 PCs		10 PCs		

we observe four numbered points aligned at the top of the curve, which could be interpreted as region where the reaction and the species diffusion rates become similar. Additionally, three disjoint and colored regions corresponding to the three previously defined regimes appear in each case. Fig. 10(a) displays the Arrhenius plot ($P = 10$ Torr) where these regions are shown according to findings in [18], with the intermediate regime (in green) encompassing points 1, 2, and 3,

extending up to 200 °C. This is because, in [18], some changes in the crystalline structure were observed at that temperature, urging the authors to classify the fourth point (at $T = 220$ °C) in the diffusion-limited regime rather than in the intermediate regime. In contrast, the results obtained with hierarchical clustering, shown Fig. 10(b), indicate that the intermediate regime (in green) extends a bit further to include point 4. As mentioned, point 4 aligns with the other three

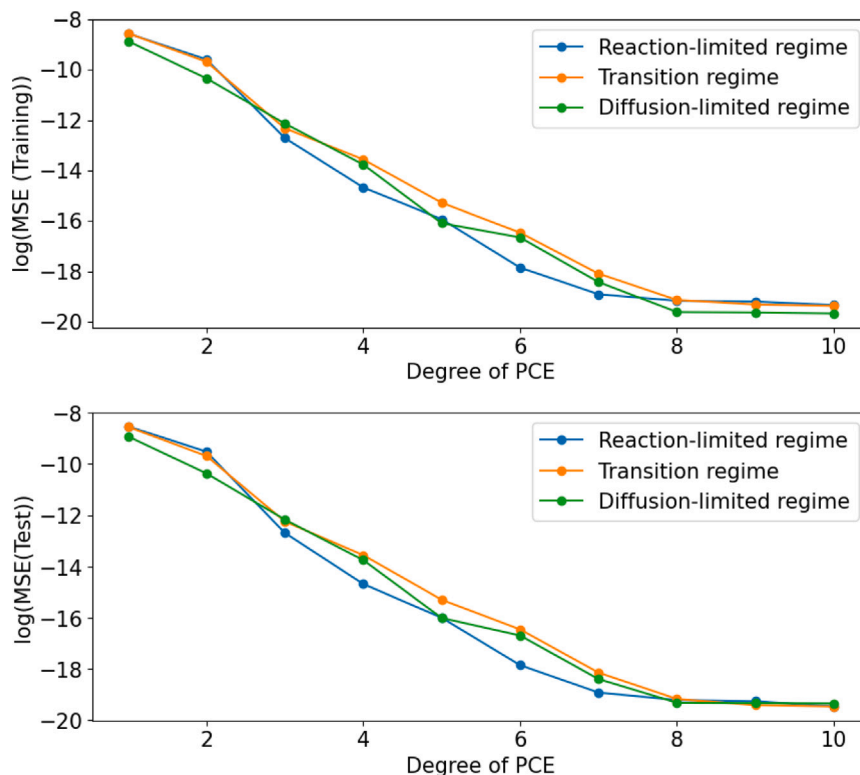


Fig. 11. Evolution of the training (at the top) and test (at the bottom) as a function of the degree of PCE for the reaction-limited regime, transition regime, and diffusion-limited regime.

points, suggesting its independence from temperature. Therefore, it seems reasonable to include it in this region, as our results indicate.

For higher pressures, the clustering results indicate a marked shift of the behavior toward lower temperatures. Namely, for the case of $P = 20$ Torr, we observe that the temperature ranges associated with the clusters are as follows: Cluster 1, linked to the reaction-limited regime, spans from 130.0–152.4 °C. Cluster 2, associated with the transition regime, falls between 152.6–190.8 °C, while Cluster 3, representing the diffusion-limited regime, ranges from 191.0–250.0 °C. Similarly, at $P = 30$ Torr, we obtain the following results: the reaction-limited regime covers temperatures between 130.0–153.6 °C, and the transition regime covers temperatures between 153.8–192.0 °C. The diffusion-limited regime remains between 192.2–250.0 °C. For $P = 40$ Torr, we observe temperature ranges analogous to the previous cases: the reaction-limited regime is between 130.0–157.2 °C, the transition regime spans in the range 157.4–203.2 °C, and the diffusion-limited regime ranges between 203.4–250.0 °C.

For higher chamber pressures $P \geq 20$ Torr, we have experimental data from only a single temperature, $T_s = 180$ °C. According to [18], at this temperature and $P = 40$ Torr, the deposition rate value is 1/8 of the deposition rate at the same temperature but at $P = 10$ Torr. This implies that the deposition rate significantly decreases at higher pressures, as will be discussed later. Our computational model, also follows this trend, finding that the deposition rate at $P = 40$ Torr is 1/7 of the value at $P = 10$ Torr. This is further confirmation of the accuracy of the CFD model and the clustering algorithm, at parameter values that are outside the training set.

Table 7

Distributions of the parameter T_s for building PCE models in each regime.

Regime	Distribution
Diffusion-limited	$\mathcal{U}(T_{s,\min,\text{diffusion}}, T_{s,\max,\text{diffusion}})$
Transition	$\mathcal{U}(T_{s,\min,\text{transition}}, T_{s,\max,\text{transition}})$
Reaction-limited	$\mathcal{U}(T_{s,\min,\text{reaction}}, T_{s,\max,\text{reaction}})$

3.3. PCE surrogate model

Proceeding to surrogate modeling, we employ Polynomial Chaos Expansion (PCE) as discussed in Section 2.3.3. We formulate surrogates model for each position in the spatial distribution in the region of interest and for each regime of the Arrhenius plot.

Furthermore, since the process relies on x_{T_s} and x_P , and both parameters are treated as random variables for constructing the PCE model, it is essential to consider their probabilistic information. However, as the distributions of these random variables are unknown, we assume they follow uniform distributions. More precisely, x_P follows the distribution $\mathcal{U}([10.0, 40.0])$ across all regimes. Meanwhile, x_{T_s} follows a uniform distribution defined in a specific interval for each regime and each pressure (according to the results presented in Table 6), as detailed in Table 7.

Under these assumptions, we construct surrogate models using Legendre polynomials associated with the uniform distribution. Their coefficients are computed employing the Least Squares Regression (LSR) method (see Section 2.3.3).

We assess the accuracy of this approach by studying the training and test errors shown in Fig. 11. These results indicate accurate predictions

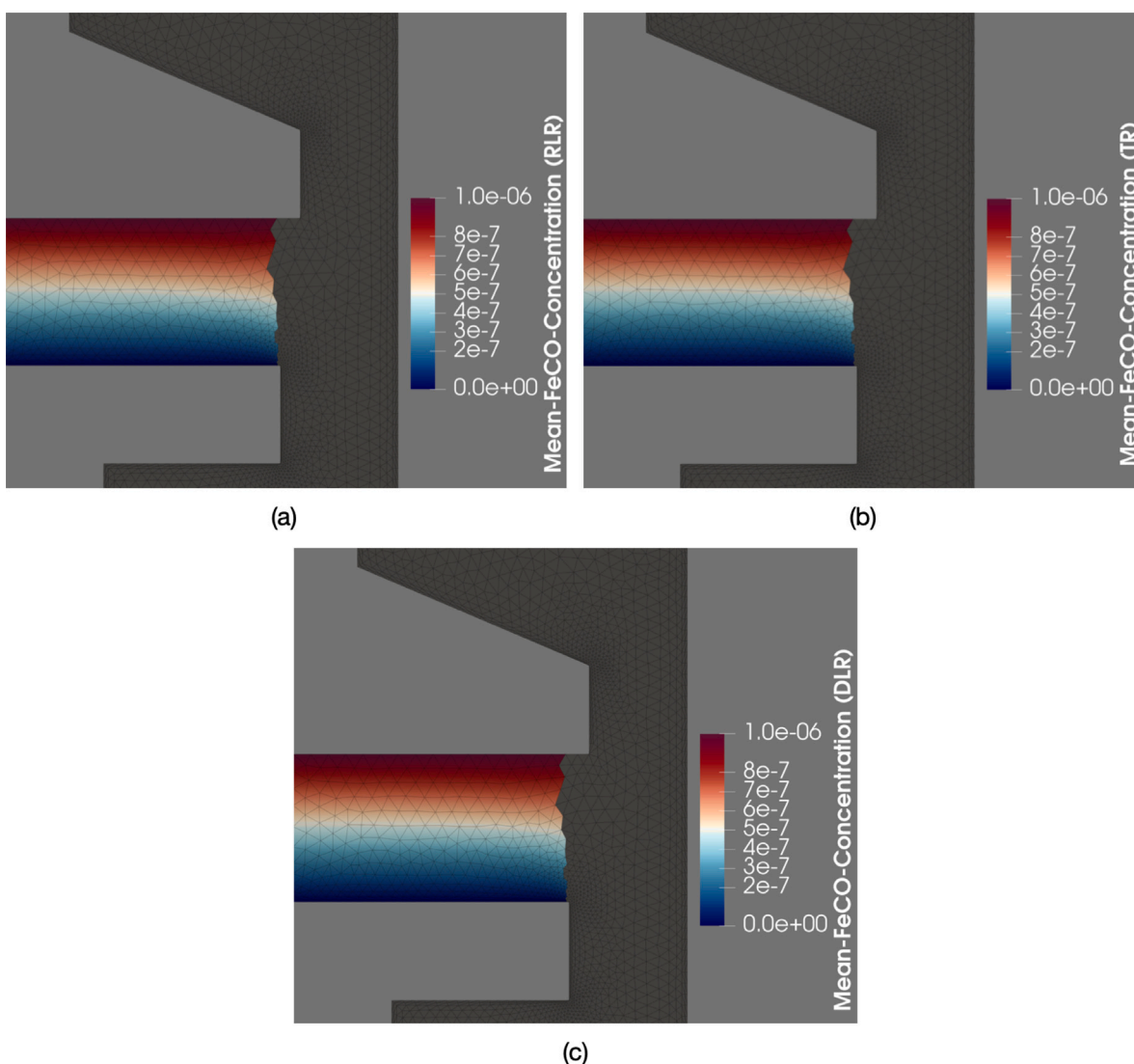


Fig. 12. Spatial distribution over the region near the deposition surface of the mean for $\text{Fe}(\text{CO})$ concentrations ($P = 10$ Torr) in (a) the reaction-limited regime, (b) the transition regime, and (c) the diffusion-limited regime. The color scale on the right indicates darker colors for smaller quantities and lighter colors for larger quantities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of spatial distribution concentrations, in every regime of the Arrhenius plot. In each case, we achieve a low-order for both errors, confirming the high accuracy of the PCE models. Based on their evolution, the maximum degree for the surrogate approximation is set at 3.

Having established that PCE can produce accurate surrogate models, we can exploit this opportunity further and extract crucial information to quantify the inputs' effects propagated through the surrogate model in terms of Sobol' indices, individually for each one of the input parameters (T , P). This is discussed in detail in the following sections.

3.4. PCE-based uncertainty quantification

The first two moments of the model responses, namely the mean and standard deviation, are calculated using the coefficients of the polynomial expansion according to (2.9). Their spatial distributions in the study area defined near the deposition surface (cf. Fig. 6) is plotted for $\text{Fe}(\text{CO})$ and $\text{Fe}(\text{CO})_5$ (cf. Figs. 12, 13, 14 and 15, respectively), and for each of the regimes identified in Section 3.2.

The mean of the $\text{Fe}(\text{CO})$ concentrations presented in Fig. 12 is lower near the deposition surface in all regimes. Furthermore, the concentration gradient is larger toward the edges of the deposition surface. The values of the standard deviation, shown in Fig. 13, are

very small everywhere in the computational domain, signifying a very small deviation of each point in the dataset from the population mean, suggesting lower variability and greater consistency.

Further examination of the results for $\text{Fe}(\text{CO})_5$ (cf. Figs. 14 and 15), reveals even smaller standard deviation overall in the domain, particularly in the vicinity of the deposition surface, suggesting smaller variability.

The prediction error for the mean and standard deviation for both species is computed in the computational domain. For $\text{Fe}(\text{CO})_5$, the relative error for the mean does not exceed $O(-3)$, while the error for the standard deviation is $O(-6)$ throughout the computational domain. Similarly, for $\text{Fe}(\text{CO})$ the magnitudes for both the mean and the standard deviation are $O(-5)$ on average.

These results elucidate the effect of uncertainty of the inputs on the distribution of precursor concentrations, in the considered region of interest. The influence of the input parameters (associated with each of the regimes) on the concentrations is revealed by sensitivity analysis in the following section.

3.5. PCE-based sensitivity analysis

The objective of this section is to estimate the effects of the two input parameters on the low-dimensional model responses across the

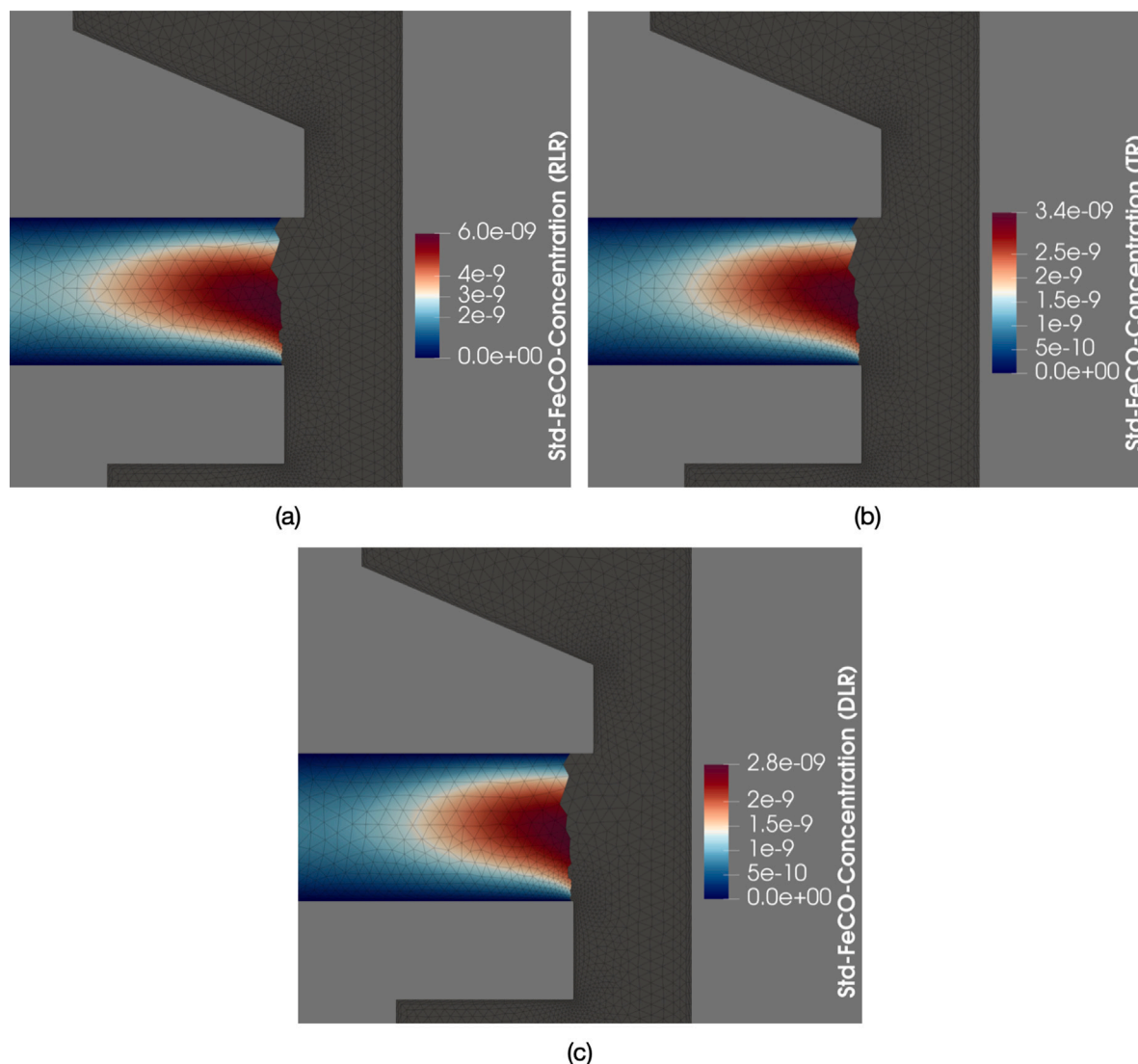


Fig. 13. Spatial distribution over the region near the deposition surface of the standard deviation for Fe(CO) concentrations ($P = 10$ Torr) in (a) the reaction-limited regime, (b) the transition regime, and (c) the diffusion-limited regime. The color scale on the right indicates darker colors for smaller quantities and lighter colors for larger quantities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

different regimes of the Arrhenius plot, using the Sobol' indices (see Section 2.3.4).

Thus, the following results correspond to the Sobol' indices calculated for the two input parameters, x_{T_s} and x_P , with respect to each of the three PC coefficients considered for each regime. In addition, since the sum of the variabilities explained by the PCs for each regime is approximately 100%, the Sobol' indices are presented using pie charts and should be read as the percentage of variability contributed by an input parameter (and its interactions) to the variability explained by each PC coefficient (PC_i).

3.5.1. Regimes of the Arrhenius plot

Reaction-limited regime

In the reaction-limited regime, chemical reaction rates increase with increasing temperature, which explains the relative significance of the substrate temperature T_s , as confirmed by the Sobol' indices in Fig. 16, from which we can infer that the substrate temperature T_s is the most influential parameter in the reaction-limited regime.

When we sum the individual contributions, we find that T_s is responsible for 68.74% of the total variability, while the remaining 31.25% is attributed to the influence of P . More precisely, at low(er) temperatures, reactions in the gas phase and on the deposition surface occur at a slower rate than the rate of chemical species diffusion. Therefore, the chemical reaction rate determines the overall deposition rate.

Transition regime

At higher temperature, the reaction rate also increases until it becomes comparable to the species diffusion rate (transition regime). The results of the sensitivity analysis show that the temperature is no longer as important as it was in the reaction-limited regime because pressure gains significance, as it dictates the transport, diffusion, and concentration of the reactive species on and to the substrate.

In Fig. 17, the Sobol' indices for this regime are presented. Here, we observe an expected balance of influences produced by the input parameters. More precisely, when we sum up the individual contributions, we obtain the following results: T_s is responsible for 53.80% of

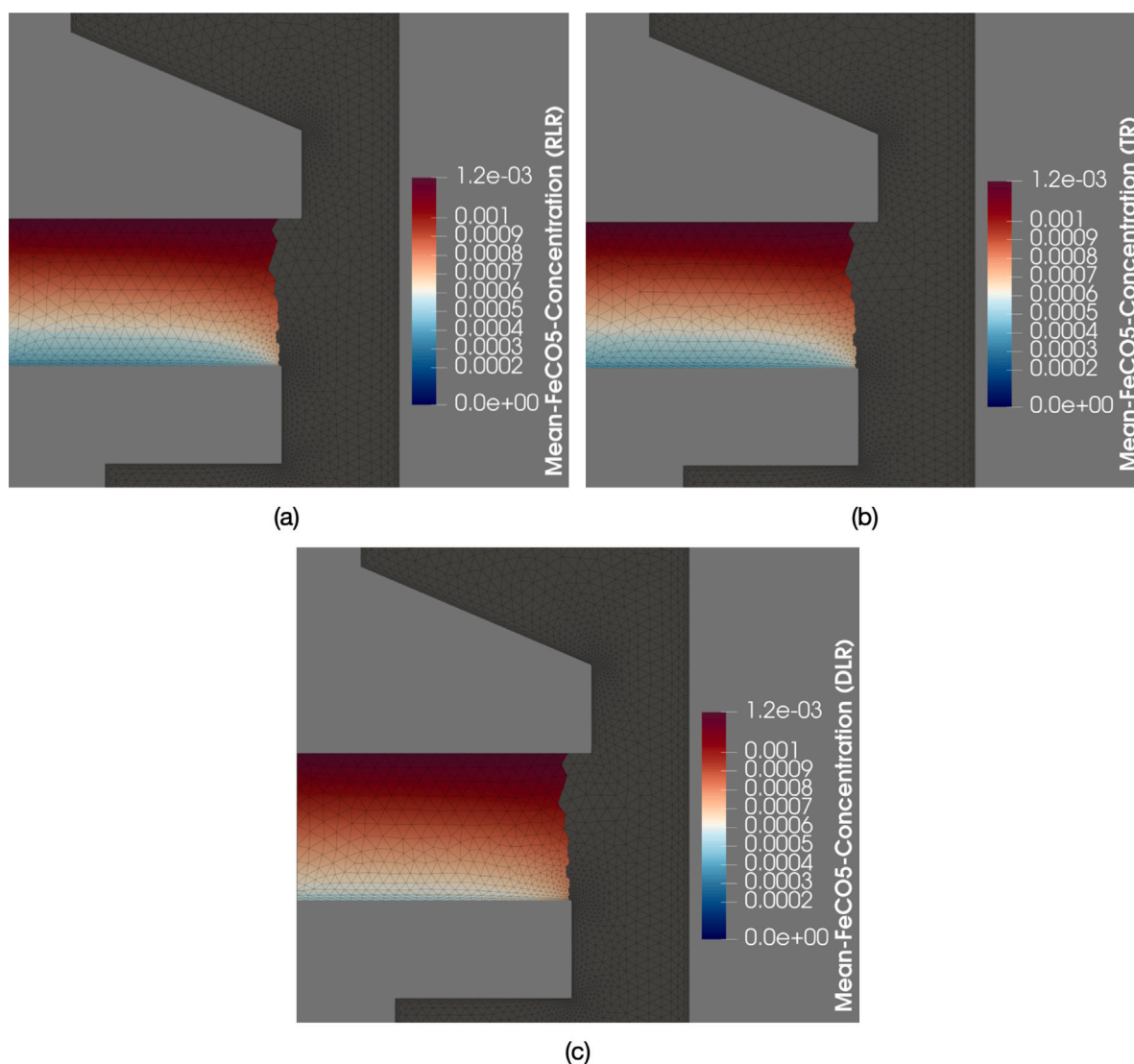


Fig. 14. Spatial distribution over the region near the deposition surface of the mean for FeCO_5 concentrations ($P = 10$ Torr) in (a) the reaction-limited regime, (b) the transition regime, and (c) the diffusion-limited regime. The color scale on the right indicates darker colors for smaller quantities and lighter colors for larger quantities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the total variability, while the remaining 46.19% corresponds to the influence of the chamber pressure P .

Diffusion-limited regime

Further increasing temperatures, typically suggests that the reaction rate is even faster and ultimately the overall reaction rate is determined by the rate of species diffusion on the deposition surface. Nevertheless, in this case, this regime is regulated by gas-phase decomposition reactions that are activated at higher temperatures causing depletion and a decrease in the deposition rate.

Indeed, Sobol' indices (refer to Fig. 18) reveal that temperature has a small effect, but they also indicate that there exists a greater influence of the chamber pressure. T_s contributes to 7.01% of the total variability, while the influence of P corresponds to 92.97%.

To interpret these results, we should revisit the Arrhenius plot Fig. 10 ($P = 10$ Torr), where the deposition rate decreases in this region due to two phenomena: (i) The high temperatures cause the precursor to decompose or be depleted in the gas-phase. (ii) There is a species diffusion from the surface back to their gas-phase. Consequently, instead of a typical straight line in the Arrhenius plot, in this case, we have a drop in the deposition rate.

4. Conclusions

This work presents a collection of data-driven methods able to (i) achieve dimensionality reduction of large data sets resulting from a CFD model of a CVD reactor; (ii) identify clusters of "outcomes" that correspond to different process regimes, where the relative influence of the various inputs changes, something which is experimentally confirmed by the Arrhenius plot; (iii) create an efficient surrogate model, without sacrificing accuracy; (iv) perform sensitivity analysis to quantify the relative effect of the process inputs on the outcome, in each one of the identified regimes. The outcomes of this analysis aids in the formulation of a hypothesis as to the phenomena taking place past the transition regime: the fact that the temperature is important even in the diffusion-limited regime of the Arrhenius plot, suggests that gas phase reactions are activated at higher temperatures.

The proposed methods offer reasonable physical insights that align with intuition that results from experimental observations and theoretical knowledge. These insights can be instrumental in decision-making for the design and optimization of the process, bypassing the need for costly and time-consuming experiments. Furthermore, this work demonstrates the potential of computational methods based on data for the design of innovative reactors.

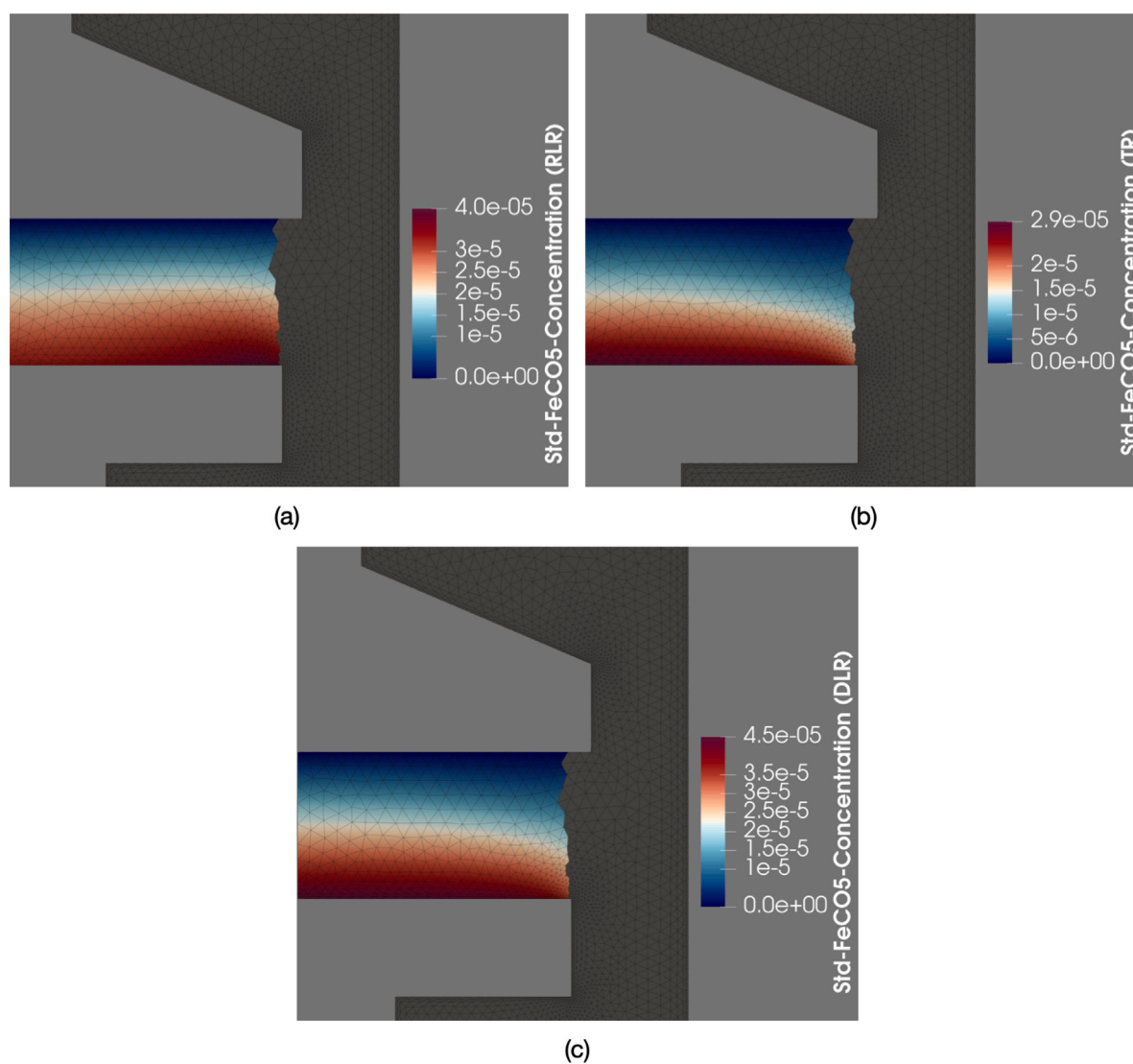


Fig. 15. Spatial distribution over the region near the deposition surface of the standard deviation for Fe(CO)_5 concentrations ($P = 10$ Torr) in (a) the reaction-limited regime, (b) the transition regime, and (c) the diffusion-limited regime. The color scale on the right indicates darker colors for smaller quantities and lighter colors for larger quantities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

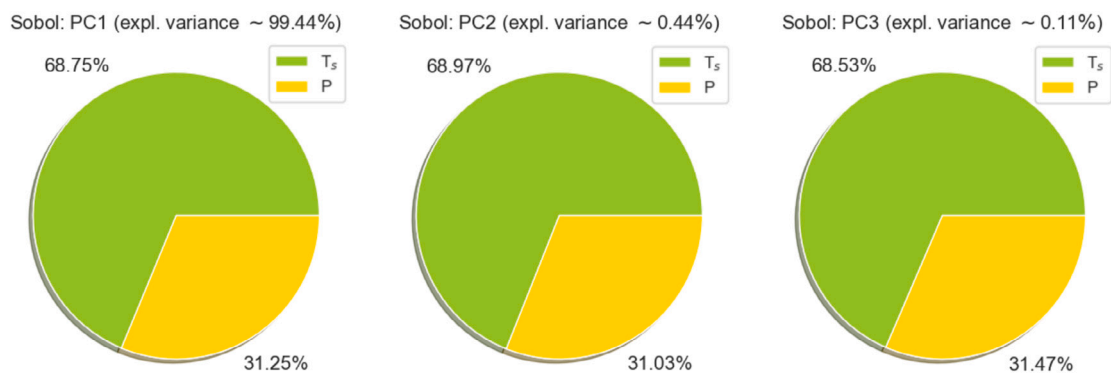


Fig. 16. Total Sobol' indices for T_s and P concerning each PC corresponding to the reaction-limited regime (cumulative variance explained > 99.99%).

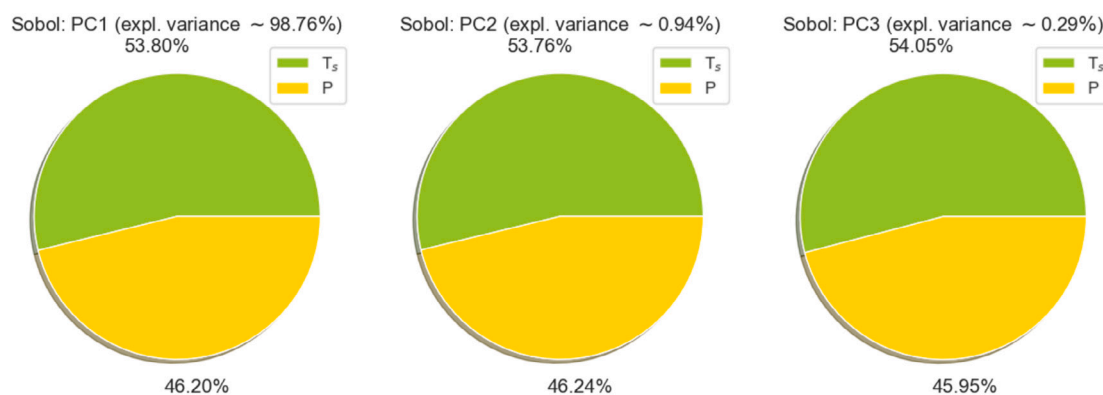


Fig. 17. Total Sobol' indices for both T_s and P concerning each PC-mode corresponding to the transition regime (cumulative variance explained > 99.99%).

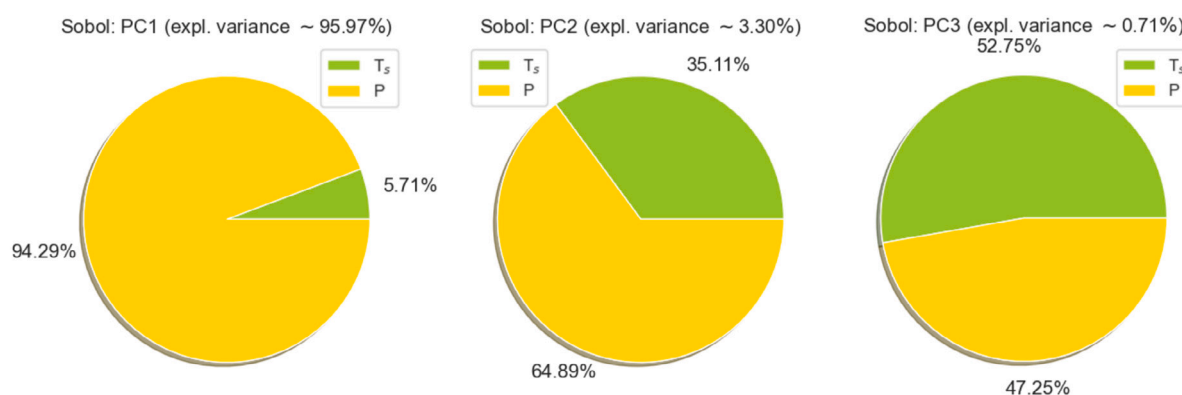


Fig. 18. Total Sobol' indices for both T_s and P concerning each PC-mode corresponding to the diffusion-limited regime (cumulative variance explained > 99.98%).

CRediT authorship contribution statement

Geremy Loachamín-Suntaxi: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation. **Paris Papavasileiou:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Eleni D. Koronaki:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Dimi- trios G. Giovanis:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis. **Georgios Gakis:** Writing – review & editing, Validation, Resources, Methodology, Investigation, Conceptualization. **Ioannis G. Aviziotis:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Martin Kathrein:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Gabriele Pozzetti:** Writing – review & editing, Validation, Methodology, Investigation. **Christoph Czettel:** Writing – review & editing, Validation, Supervision, Methodology, Investigation. **Stéphane P.A. Bordas:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis. **Andreas G. Boudouvis:** Writing – review & editing, Supervision, Formal analysis.

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.

Acknowledgments

For the purpose of open access, the authors have applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission.

Appendix. Principal component analysis

Given a dataset $\mathbf{Y}(m \times N)$, where we have collected m simulations of the N variables, we start by centering or standardizing the data (to give each variable the same importance throughout the analysis especially when data has different magnitudes) [60,81]. This transformation is expressed as follows:

$$\tilde{\mathbf{Y}} = (\mathbf{Y} - \bar{\mathbf{Y}}) \mathbf{S}^{-1},$$

where $\bar{\mathbf{Y}}$ is an $m \times N$ matrix containing the mean of each column vector of the simulated output variables and $\mathbf{S}$ is an $m \times N$ matrix containing the corresponding standard deviations ($\mathbf{S} = \mathbf{I}_{m \times N}$, when we only center the data).

To perform PCA [31,32], the covariance matrix of $\tilde{\mathbf{Y}}$ and its SVD are considered:

$$\text{cov}(\tilde{\mathbf{Y}}) = \tilde{\mathbf{Y}}^T \tilde{\mathbf{Y}} / m - 1 = \mathbf{V} \mathbf{\Sigma} \mathbf{V}^T,$$

where $\mathbf{V}$ is an $N \times N$ orthogonal matrix with eigenvectors of $\text{cov}(\tilde{\mathbf{Y}})$, whereas the $N \times N$ diagonal matrix $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_N)$ contains the associated eigenvalues arranged in decreasing order, that is, $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_N$ (see Fig. 3). More precisely, the columns of matrix $\mathbf{V}$ are uncorrelated vectors and correspond to the directions in which the data varies the most, defining a new orthogonal basis. These basis vectors are called *principal components* (PCs). Similarly, we can interpret the corresponding eigenvalues as variances of the data in such directions.

Subsequently, we projected the standardized dataset $\tilde{\mathbf{Y}}$ on the new basis. The obtained values define the *principal component scores* arranged in an $m \times N$ matrix $\mathbf{Z} = \tilde{\mathbf{Y}}\mathbf{V}$. In addition, since the variance captured by each PC decreases, it is sufficient to retain only the first n dominant PCs that describe the highest cumulative variability. Thus, from $\tilde{\mathbf{Y}}$ and $\mathbf{V}$, we can compute their truncated versions $\tilde{\mathbf{Y}}_n (m \times n)$ and $\mathbf{V}_n (N \times n)$ such that

$$\mathbf{Z} \approx \tilde{\mathbf{Z}}_n = \tilde{\mathbf{Y}}_n \mathbf{V}_n. \quad (\text{A.1})$$

A criterion for finding the smallest $n \leq N$, that suffices for the accurate data representation, is to study the total variance ratio such that it is larger than a fixed threshold $\delta > 0$, i.e., $\sum_{i=1}^n \sigma_i / \sum_{i=1}^N \sigma_i > \delta$. We retain the PCs required to explain at least 99% of the total variance of the data. Furthermore, using Eq. (A.1), the original dataset can be reconstructed as follows:

$$\tilde{\mathbf{Y}} \approx \tilde{\mathbf{Z}}_n \mathbf{V}_n^T, \quad (\text{A.2})$$

where $\mathbf{V}_n^T (m \times n)$ denotes the transpose of $\mathbf{V}_n$.

Additionally, to assess the accuracy of PCA, we use Eq. (A.2) and compute the *reconstruction error* by means of the *mean squared error*, $\text{MSE} = \mathbb{E} \left[(\tilde{\mathbf{Y}} - \tilde{\mathbf{Z}}_n \mathbf{V}_n^T) (\tilde{\mathbf{Y}} - \tilde{\mathbf{Z}}_n \mathbf{V}_n^T)^T \right]$.

Data availability

Data will be made available on request.

References

- [1] K.L. Choy, Chemical vapour deposition of coatings, Prog. Mater. Sci. 48 (2003) 57–170, [http://dx.doi.org/10.1016/S0079-6425\(01\)00009-3](http://dx.doi.org/10.1016/S0079-6425(01)00009-3).
- [2] S.M. George, Atomic layer deposition: an overview, Chem. Rev. 110 (2010) 111–131, <http://dx.doi.org/10.1021/cr900056b>.
- [3] U. Helmersson, M. Lattemann, J. Bohlmark, A.P. Ehasarian, J.T. Gudmundsson, Ionized physical vapor deposition (IPVD): A review of technology and applications, Thin Solid Films 513 (2006) 1–24, <http://dx.doi.org/10.1016/j.tsf.2006.03.033>.
- [4] M. Melloch, M. Reed, Compound Semiconductors, Compound semiconductors 1997, in: Proceedings of the IEEE Twenty-Fourth International Symposium on Compound Semiconductors, San Diego, CA, USA, 1997.
- [5] P.A. Gkinis, I.G. Aviziotis, E.D. Koronaki, G.P. Gakis, A.G. Boudouvis, The effects of flow multiplicity on GaN deposition in a rotating disk CVD reactor, J. Cryst. Growth 458 (2017) 140–148, <http://dx.doi.org/10.1016/j.jcrysgro.2016.10.065>.
- [6] G.P. Gakis, C. Vahlas, H. Vergnes, S. Dourdain, Y. Tison, H. Martínez, J. Bour, D. Ruch, A.G. Boudouvis, B. Caussat, E. Scheid, Investigation of the initial deposition steps and the interfacial layer of atomic layer deposited (ALD) Al₂O₃ on Si, Appl. Surf. Sci. 492 (2019) 245–254, <http://dx.doi.org/10.1016/j.apsusc.2019.06.215>.
- [7] C. Vahlas, B. Caussat, P. Serp, G.N. Angelopoulos, Principles and applications of CVD powder technology, Mater. Sci. Eng. R Rep. 53 (2006) 1–72, <http://dx.doi.org/10.1016/j.mser.2006.05.001>.
- [8] J.W. Suk, A. Kitt, C.W. Magnuson, Y. Hao, S. Ahmed, J. An, A.K. Swan, B.B. Goldberg, R.S. Ruoff, Transfer of CVD-grown monolayer graphene onto arbitrary substrates, ACS Nano 5 (2011) 6916–6924, <http://dx.doi.org/10.1021/nn201207c>.
- [9] G.P. Gakis, S. Termine, A.F.A. Trompeta, I.G. Aviziotis, C.A. Charitidis, Unraveling the mechanisms of carbon nanotube growth by chemical vapor deposition, J. Chem. Eng. 445 (2022) 136807, <http://dx.doi.org/10.1016/j.ccej.2022.136807>.
- [10] S.M. Rossnagel, Thin film deposition with physical vapor deposition and related technologies, J. Vac. Sci. Technol. 21 (2003) S74–S87, <http://dx.doi.org/10.1116/1.1600450>.
- [11] R.W. Johnson, A. Hultqvist, S.F. Bent, A brief review of atomic layer deposition: from fundamentals to applications, Mater. Today 17 (2014) 236–246, <http://dx.doi.org/10.1016/j.mattod.2014.04.026>.
- [12] G.M. Psarellis, I.G. Aviziotis, T. Duguet, C. Vahlas, E.D. Koronaki, A.G. Boudouvis, Investigation of reaction mechanisms in the chemical vapor deposition of Al from DMEAA, Chem. Eng. Sci. 177 (2018) 464–470, <http://dx.doi.org/10.1016/j.ces.2017.12.006>.
- [13] E.D. Koronaki, G.P. Gakis, N. Cheimarios, A.G. Boudouvis, Efficient tracing and stability analysis of multiple stationary and periodic states with exploitation of commercial CFD software, Chem. Eng. Sci. 150 (2016) 26–34, <http://dx.doi.org/10.1016/j.ces.2016.04.043>.
- [14] C.R. Kleijn, R. Dorsman, K.J. Kuijlaars, M. Okkerse, H. van Santen, Multi-scale modeling of chemical vapor deposition processes for thin film technology, J. Cryst. Growth 303 (2007) 362–380, <http://dx.doi.org/10.1016/j.jcrysgro.2006.12.062>.
- [15] G.P. Gakis, H. Vergnes, E. Scheid, C. Vahlas, A.G. Boudouvis, B. Caussat, Detailed investigation of the surface mechanisms and their interplay with transport phenomena in alumina atomic layer deposition from TMA and water, Chem. Eng. Sci. 195 (2019) 399–412, <http://dx.doi.org/10.1016/j.ces.2018.09.037>.
- [16] P.A. Gkinis, E.D. Koronaki, A. Skouteris, I.G. Aviziotis, A.G. Boudouvis, Building a data-driven reduced order model of a chemical vapor deposition process from low-fidelity CFD simulations, Chem. Eng. Sci. 199 (2019) 371–380, <http://dx.doi.org/10.1016/j.ces.2019.01.009>.
- [17] R. Fritzsche, D.R. Zahn, M. Mehning, Atmospheric pressure metal organic chemical vapor deposition of thin germanium films, J. Mater. Sci. 56 (2021) 9274–9286, <http://dx.doi.org/10.1007/s10853-021-05871-9>.
- [18] I.G. Aviziotis, T. Duguet, C. Vahlas, A.G. Boudouvis, Combined macro/nanoscale investigation of the chemical vapor deposition of Fe from Fe(CO)₅, Adv. Mater. Interfaces 4 (2017) 1601185, <http://dx.doi.org/10.1002/admi.201601185>.
- [19] G.P. Gakis, E.N. Skountzos, I.G. Aviziotis, C.A. Charitidis, Multi-parametric analysis of the CVD of CNTs: Effect of reaction temperature, pressure and acetylene flow rate, Chem. Eng. Sci. 267 (2023) 118374, <http://dx.doi.org/10.1016/j.ces.2022.118374>.
- [20] G.P. Gakis, E.D. Koronaki, A.G. Boudouvis, Numerical investigation of multiple stationary and time-periodic flow regimes in vertical rotating disc CVD reactors, J. Cryst. Growth 432 (2015) 152–159, <http://dx.doi.org/10.1016/j.jcrysgro.2015.09.026>.
- [21] P. Papavasileiou, E.D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A.G. Boudouvis, T.J. Mountziaris, S.P.A. Bordas, 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 (2022) 314–325, <http://dx.doi.org/10.1016/j.cherd.2022.08.005>.
- [22] P. Papavasileiou, E.D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A.G. Boudouvis, S.P.A. Bordas, Equation-based and data-driven modeling strategies for industrial coating processes, Comput. Ind. 149 (2023) 103938, <http://dx.doi.org/10.1016/j.compind.2023.103938>.
- [23] E.D. Koronaki, A.M. Nikas, A.G. Boudouvis, A data-driven reduced-order model of nonlinear processes based on diffusion maps and artificial neural networks, J. Chem. Eng. 397 (2020) 125475, <http://dx.doi.org/10.1016/j.ccej.2020.125475>.
- [24] R. Spencer, P. Gkinis, E.D. Koronaki, D.I. Gerogiorgis, S.P.A. Bordas, A.G. Boudouvis, Investigation of the chemical vapor deposition of Cu from copper amidinate through data driven efficient CFD modelling, Comput. Chem. Eng. 149 (2021) 107289, <http://dx.doi.org/10.1016/j.compchemeng.2021.107289>.
- [25] E.D. Koronaki, N. Evangelou, Y.M. Psarellis, A.G. Boudouvis, I.G. Kevrekidis, From partial data to out-of-sample parameter and observation estimation with diffusion maps and geometric harmonics, Comput. Chem. Eng. 178 (2023) 108357, <http://dx.doi.org/10.1016/j.compchemeng.2023.108357>.
- [26] C.P. Martín-Linares, Y.M. Psarellis, G. Karapetsas, E.D. Koronaki, I.G. Kevrekidis, Physics-agnostic and physics-infused machine learning for thin films flows: modelling, and predictions from small data, J. Fluid Mech. 975 (2023) A41, <http://dx.doi.org/10.1017/jfm.2023.868>.
- [27] A. Bhoekar, M. Ierapetritou, Advances in surrogate based modeling, feasibility analysis, and optimization: A review, Comput. Chem. Eng. 108 (2018) 250–267, <http://dx.doi.org/10.1016/j.compchemeng.2017.09.017>.
- [28] K. McBride, K. Sundmacher, Overview of surrogate modeling in chemical process engineering, Chem. Ing. Tech. 91 (2019) 228–239, <http://dx.doi.org/10.1002/cite.201800091>.
- [29] T.C. Xenidou, N. Prud'homme, C. Vahlas, N.C. Markatos, A.G. Boudouvis, Reaction and transport interplay in Al MOCVD investigated through experiments and computational fluid dynamic analysis, J. Electrochem. Soc. 157 (2010) D633, <http://dx.doi.org/10.1149/1.3493617>.
- [30] I.G. Aviziotis, Chemical Vapor Deposition of Al, Fe and of the Al₁₃Fe₄ Approximant Intermetallic Phase: Experiments and Multiscale Simulations (Ph.D. thesis), Toulouse INP, 2016.
- [31] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441, <http://dx.doi.org/10.1037/h0071325>.
- [32] A. Mackiewicz, W. Ratajczak, Principal components analysis (PCA), Comput. Geosci. 19 (1993) 303–342, [http://dx.doi.org/10.1016/0098-3004\(93\)90090-R](http://dx.doi.org/10.1016/0098-3004(93)90090-R).
- [33] L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, 2020, <http://dx.doi.org/10.48550/arXiv.1802.03426>.
- [34] B. Hoojogh, A. Ghodsi, F. Karay, M. Crowley, Uniform manifold approximation and projection (UMAP) and its variants: Tutorial and survey, 2021, <http://dx.doi.org/10.48550/arXiv.2109.02508>.
- [35] B. Schölkopf, A. Smola, K.R. Müller, Kernel principal component analysis, in: W. Gerstner, A. Germond, M. Hasler, J.D. Nicoud (Eds.), Artificial Neural Networks ICANN 1997, in: Lecture Notes in Computer Science 1327, Springer, Berlin, Heidelberg, 1997, <http://dx.doi.org/10.1007/BFb0020217>.
- [36] B. Schölkopf, A. Smola, K.R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319, <http://dx.doi.org/10.1162/089976698300017467>.

- [37] B.J. Isaac, A. Coussement, O. Gicquel, P.J. Smith, A. Parente, Reduced-order PCA models for chemical reacting flows, *Combust. Flame* 161 (2014) 2785–2800, <http://dx.doi.org/10.1016/j.combustflame.2014.05.011>.
- [38] H.J. Huang, L.H. Kau, H.S. Wang, Y.L. Hsieh, C.C. Lee, Y.K. Fuh, T.T. Li, Large-scale data analysis of PECVD amorphous silicon interface passivation layer via the optical emission spectra for parameterized PCA, *Int. J. Adv. Manuf. Technol.* 101 (2019) 329–337, <http://dx.doi.org/10.1007/s00170-018-2938-1>.
- [39] E.D. Koronaki, P.A. Gkinis, L. Beex, S.P.A. Bordas, C. Theodoropoulos, A.G. Boudouvis, Classification of states and model order reduction of large scale Chemical Vapor Deposition processes with solution multiplicity, *Comput. Chem. Eng.* 121 (2019) 148–157, <http://dx.doi.org/10.1016/j.compchemeng.2018.08.023>.
- [40] T. Hastie, R. Tibshirani, J. Friedman, in: T. Hastie, R. Tibshirani, J. Friedman (Eds.), *Unsupervised Learning, in the Elements of Statistical Learning: Data Mining, Inference, and Prediction*, in: Springer Series in Statistics, Springer, New York, NY, 2009, pp. 485–585, http://dx.doi.org/10.1007/978-0-387-84858-7_14.
- [41] G. James, D. Witten, T. Hastie, R. Tibshirani, in: G. James, D. Witten, T. Hastie, R. Tibshirani (Eds.), *Unsupervised Learning, in an Introduction to Statistical Learning: With Applications in R*, in: Springer Texts in Statistics, Springer US, New York, NY, 2021, pp. 497–552, http://dx.doi.org/10.1007/978-1-0716-1418-1_12.
- [42] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Oakland, CA, USA, 1967, pp. 281–297.
- [43] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96*, Portland, Oregon, AAAI Press, 1996, pp. 226–231.
- [44] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, OPTICS: ordering points to identify the clustering structure, *SIGMOD Rec.* 28 (1999) 49–60, <http://dx.doi.org/10.1145/304181.304187>.
- [45] E. Schubert, J. Sander, M. Ester, H.P. Kriegel, X. Xu, Dbscan revisited, revisited: Why and how you should (Still) use DBSCAN, *ACM Trans. Database Syst.* 42 (2018) 1–21, <http://dx.doi.org/10.1145/3068335>.
- [46] A.D. Gordon, A review of hierarchical classification, *J. R. Stat. Soc.* 150 (1987) 119–137.
- [47] D. Kavyasrujana, B.C. Rao, Hierarchical clustering for sentence extraction using cosine similarity measure, in: S. Satapathy, A. Govardhan, K. Raju, J. Mandal (Eds.), *Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India, CSI*, in: *Advances in Intelligent Systems and Computing*, Springer, Cham, 2015, http://dx.doi.org/10.1007/978-3-319-13728-5_21.
- [48] M.A. Mercioni, S. Holban, A study on hierarchical clustering and the distance metrics for identifying architectural styles, in: *International Conference on Energy and Environment, CIEM, Timisoara*, 2019, pp. 49–53, <http://dx.doi.org/10.1109/CIEM46456.2019.8937662>.
- [49] M. Vagni, N. Giordano, G. Balestra, S. Rosati, Comparison of different similarity measures in hierarchical clustering, in: *IEEE International Symposium on Medical Measurements and Applications, MeMeA*, 2021, <http://dx.doi.org/10.1109/memea52024.2021.94>.
- [50] J.H. Ward Jr., Hierarchical grouping to optimize an objective function, *J. Amer. Statist. Assoc.* 58 (1963) 236–244, <http://dx.doi.org/10.2307/2282967>.
- [51] G. Szekely, M. Rizzo, Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method, *J. Classification* 22 (2005) 151–183, <http://dx.doi.org/10.1007/s00357-005-0012-9>.
- [52] N. Wiener, The homogeneous chaos, *Am. J. Math.* 60 (1938) 897–936, <http://dx.doi.org/10.2307/2371268>.
- [53] D. Lucor, C.-H. Su, G.E. Karniadakis, Generalized polynomial chaos and random oscillators, *Internat. J. Numer. Methods Engrg.* 60 (2004) 0029–5981, <http://dx.doi.org/10.1002/nme.976>.
- [54] L. Le Gratiet, S. Marelli, B. Sudret, Metamodel-based sensitivity analysis: Polynomial chaos expansions and Gaussian processes, in: R. Ghanem, D. Higdon, H. Owhadi (Eds.), *Handbook of Uncertainty Quantification*, Springer, Cham, 2017, http://dx.doi.org/10.1007/978-3-319-12385-1_38.
- [55] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, *Reliab. Eng. Syst. Saf.* 93 (2008) 964–979, <http://dx.doi.org/10.1016/j.res.2007.04.002>.
- [56] T. Berkemeier, M. Krüger, A. Feinberg, M. Müller, U. Pöschl, U.K. Krieger, Accelerating models for multiphase chemical kinetics through machine learning with polynomial chaos expansion and neural networks, *Geosci. Model Dev.* 16 (2023) 2037–2054, <http://dx.doi.org/10.5194/gmd-16-2037-2023>.
- [57] D. Xiu, G.E. Karniadakis, The Wiener-Askey polynomial chaos for stochastic differential equations, *SIAM J. Sci. Comput.* 24 (2002) 619–644, <http://dx.doi.org/10.1137/S1064827501387826>.
- [58] M. Villegas, F. Augustin, A. Gilg, A. Hmaid, U. Wever, Application of the polynomial chaos expansion to the simulation of chemical reactors with uncertainties, *Math. Comput. Simulation* 82 (2012) 805–817, <http://dx.doi.org/10.1016/j.matcom.2011.12.001>.
- [59] M. Kim, S. Cho, K. Jang, S. Hong, J. Na, I. Moon, Data-driven robust optimization for minimum nitrogen oxide emission under process uncertainty, *J. Chem. Eng.* 428 (2022) 130971, <http://dx.doi.org/10.1016/j.cej.2021.130971>.
- [60] J.B. Nagel, J. Rieckermann, B. Sudret, Principal component analysis and sparse polynomial chaos expansions for global sensitivity analysis and model calibration: Application to urban drainage simulation, *Reliab. Eng. Syst. Saf.* 195 (2020) 106737, <http://dx.doi.org/10.1016/j.res.2019.106737>.
- [61] P.L. Trung Duong, W. Ali, E. Kwok, M. Lee, Uncertainty quantification and global sensitivity analysis of complex chemical process using a generalized polynomial chaos approach, *Comput. Chem. Eng.* 90 (2016) 23–30, <http://dx.doi.org/10.1016/j.compchemeng.2016.03.020>.
- [62] B. Sudret, *Uncertainty Propagation and Sensitivity Analysis in Mechanical Models - Contributions To Structural Reliability and Stochastic Spectral Methods* Habilitation thesis, Université Blaise Pascal, Clermont-Ferrand, 2007.
- [63] O.P. Le Maître, O.M. Knio, *Spectral Methods for Uncertainty Quantification with Applications to Computational Fluid Dynamics*, Springer Dordrecht, 2012, <http://dx.doi.org/10.1007/978-90-481-3520-2>.
- [64] C. Soize, R. Ghanem, Physical systems with random uncertainties: Chaos representations with arbitrary probability measure, *SIAM J. Sci. Comput.* 26 (2004) 395–410, <http://dx.doi.org/10.1137/S1064827503424505>.
- [65] X. Wan, G.E. Karniadakis, Multi-element generalized polynomial chaos for arbitrary probability measures, *SIAM J. Sci. Comput.* 28 (2006) 901–928, <http://dx.doi.org/10.1137/050627630>.
- [66] M. Berveiller, B. Sudret, M. Lemaire, Stochastic finite element: a non intrusive approach by regression, *Eur. J. Comput. Mech.* 15 (2006) 81–92, <http://dx.doi.org/10.3166/remn.15.81-92>.
- [67] M. Hadigol, A. Doostan, Least squares polynomial chaos expansion: A review of sampling strategies, *Comput. Methods Appl. Mech. Engrg.* 332 (2018) 382–407, <http://dx.doi.org/10.1016/j.cma.2017.12.019>.
- [68] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Stat. Soc. Ser. B Methodol.* 58 (1996) 267–288, <http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x>.
- [69] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, *J. Comput. Phys.* 230 (2011) 2345–2367, <http://dx.doi.org/10.1016/j.jcp.2010.12.021>.
- [70] G. Blatman, B. Sudret, An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis, *Probab. Eng. Mech.* 25 (2010) 183–197, <http://dx.doi.org/10.1016/j.probengmech.2009.10.003>.
- [71] G. Saporta, *Probabilités, analyse des données et statistique*, third ed., TECHNIP, Paris, 2011.
- [72] I.M. Sobol, Sensitivity estimates for nonlinear mathematical models, *Math. Model. Comput. Exp.* 1 (1993) 407–414.
- [73] X. Xie, R. Schenkendorf, U. Krewer, Efficient sensitivity analysis and interpretation of parameter correlations in chemical engineering, *Reliab. Eng. Syst. Saf.* 187 (2019) 159–173, <http://dx.doi.org/10.1016/j.res.2018.06.010>.
- [74] M. Santanoceto, M. Tibergh, Z. Perkó, S. Dulla, D. Lathouwers, Preliminary uncertainty and sensitivity analysis of the molten salt fast reactor steady-state using a polynomial chaos expansion method, *Ann. Nucl. Energy* 159 (2021) 108311, <http://dx.doi.org/10.1016/j.anucene.2021.108311>.
- [75] T. Crestaux, O. Le Maître, J.M. Martinez, Polynomial chaos expansion for sensitivity analysis, *Reliab. Eng. Syst. Saf.* 94 (2009) 1161–1172, <http://dx.doi.org/10.1016/j.res.2008.10.008>.
- [76] X. Sun, Y.Y. Choi, J. Choi, Global sensitivity analysis for multivariate outputs using polynomial chaos-based surrogate models, *Appl. Math. Model.* 82 (2020) 867–887, <http://dx.doi.org/10.1016/j.apm.2020.02.005>.
- [77] J.P. Hesslering, Deterministic sampling for quantification of modeling uncertainty of signals, in: F.P. García Márquez, N. Zaman (Eds.), *Digital Filters and Signal Processing*, InTech, Rijeka, Croatia, 2013, pp. 53–79, <http://dx.doi.org/10.5772/52193>.
- [78] A. Cutrono Rakhimov, D.C. Visser, E.M.J. Komen, Uncertainty quantification method for CFD applied to the turbulent mixing of two water layers – II: Deterministic sampling for input uncertainty, *Nucl. Eng. Des.* 348 (2019) 146–158, <http://dx.doi.org/10.1016/j.nucengdes.2019.04.016>.
- [79] F. Pedregosa, A. Varoquaux, A. Gramfort, V. Michel, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, *J. Mach. Learn. Res.* 12 (2011) 2825–2830, <https://dl.acm.org/doi/10.5555/1953048.2078195>.
- [80] A. Olivier, D.G. Giovanis, B.S. Aakash, M. Chauhan, L. Vandanapu, M.D. Shields, UQpy: A general purpose python package and development environment for uncertainty quantification, *J. Comput. Sci.* 47 (2020) 101204, <http://dx.doi.org/10.1016/j.jocs.2020.101204>.
- [81] A. Parente, J.C. Sutherland, Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity, *Combust. Flame* 160 (2013) 340–350, <http://dx.doi.org/10.1016/j.combustflame.2012.09.016>.