Physically interpretable machine learning algorithm on multidimensional non-linear fields

Rem-Sophia Mouradi, Cédric Goeury, Olivier Thual, Fabrice Zaoui, Pablo Tassi

Physically interpretable machine learning algorithm on multidimensional non-linear fields

Rem-Sophia Mouradi[1,2], Cédric Goeury[1], Olivier Thual[2,3], Fabrice Zaoui[1], and Pablo Tassi[1,4]

[1]EDF R&D, National Laboratory for Hydraulics and Environment (LNHE), 6 Quai Watier, 78400 Chatou, France
[2]Climate, Environment, Coupling and Uncertainties research unit (CECI) at the European Center for Research and Advanced Training in Scientific Computation (CERFACS), French National Research Center (CNRS), 42 Avenue Gaspard Coriolis, 31820 Toulouse, France
[3]Institut de Mécanique des Fluides de Toulouse (IMFT), Université de Toulouse, CNRS, Toulouse, France
[4]Saint-Venant Laboratory for Hydraulics (LHSV), Chatou, France

25 mai 2020

Résumé

In an ever-increasing interest for Machine Learning (ML) and a favorable data development context, we here propose an original methodology for data-based prediction of two-dimensional physical fields. Polynomial Chaos Expansion (PCE), widely used in the Uncertainty Quantification community (UQ), has recently shown promising prediction characteristics for one-dimensional problems, with advantages that are inherent to the method such as its explicitness and adaptability to small training sets, in addition to the associated probabilistic framework. Simultaneously, Dimensionality Reduction (DR) techniques are increasingly used for pattern recognition and data compression and have gained interest due to improved data quality. In this study, the interest of Proper Orthogonal Decomposition (POD) for the construction of a statistical predictive model is demonstrated. Both POD and PCE have widely proved their worth in their respective frameworks. The goal of the present paper was to combine them for a field-measurement-based forecasting. The described steps are also useful to analyze the data. Some challenging issues encountered when using multidimensional field measurements are addressed, for example when dealing with few data. The POD-PCE coupling methodology is presented, with particular focus on input data characteristics and training-set choice. A simple methodology for evaluating the importance of each physical parameter is proposed for the PCE model and extended to the POD-PCE coupling.

1 Introduction

Deep Learning techniques (DL [59, 29]) and more generally Machine Learning (ML [95, 79]), and their applications to physical problems (fluid mechanics [51, 9, 74, 67] ; aerodynamics [115, 110] ; plasma physics [28, 83] ; astrophysics and astronomy [106, 50] ; particle physics [2] ; quantum mechanics [70], geosciences [46, 86, 88, 92, 27]) have made a promising take-off in the last few years. This has been particularly the case for fields where the measurement potential has dramatically increased, with increasing spatiotemporal resolution (e.g. Geoscience Data [46] ; the new SWOT satellite mission [78, 71]). In particular, multi-layer Neural Networks (NN) [91] are widely used for these applications. These techniques are of interest when data are available for a real-world problem that is difficult to model using process-based equations, because prescribing the underlying relationships is complex [95]. For such cases, the learning procedure is performed through a combination of steps: *Encoding*, *Hidden Layers* or *Latent Representation*, *Decoding* [59, 95]. In particular, transformation functions, also called *Activation Functions* (AFs), are used jointly with *weight matrices*, to transform the data from one layer to another. The popularity of NN comes from this complex structure, which makes them adaptable for various applications [94, 1]. Indeed, the combination of AF transformations helps capture complex interactions and non-linearities, making NN a widely used approach [95].

However, some limitations prevent the use of NN for physical applications.

(*i*) It is difficult to provide an explicit input-to-output formulation, due to the combinations and transformations involved. Physical interpretation of the constructed model is therefore tedious. This is why NN have been applied in physics often as a black-box and rather to optimize the accuracy of prediction than to extract physical information [41].

(*ii*) The calibration of the NN model involves various choices, such as the number of layers, number of neurons, the AFs, the weight matrices, etc. The number of hyper-parameters grows with the number of neurons and layers (*curse of dimensionality*), which limits the maximum complexity and therefore the interest of NN for highly non linear problems [79]. The choice of AF is application and data-set dependent [55].

(*iii*) The theoretical ability of NN to approximate functions has been proven with the fundamental *Universal Approximation Theorem* either with a limited number of layers [18, 23, 38, 5, 12] or with a particular type of AFs (ReLu functions in [21, 13, 64, 87, 34]). This limits the utility of NN for non-linear physical applications [94, 1]. Consequently, simple networks (with few layers) [41] are often used for physical problems, and the necessary assumptions on the AFs are often overlooked in favor of efficiency [55].

To overcome these limitations, we propose an alternative ML, based on a coupling between Proper Orthogonal Decomposition (POD) [14] and Polynomial Chaos Expansion (PCE) [57, 58], suitable for the prediction of spatially-distributed physical fields. Here, POD is used for both *Encoding* and *Decoding* whereas PCE is used as a *Latent Representation*, as represented in Figure 1.

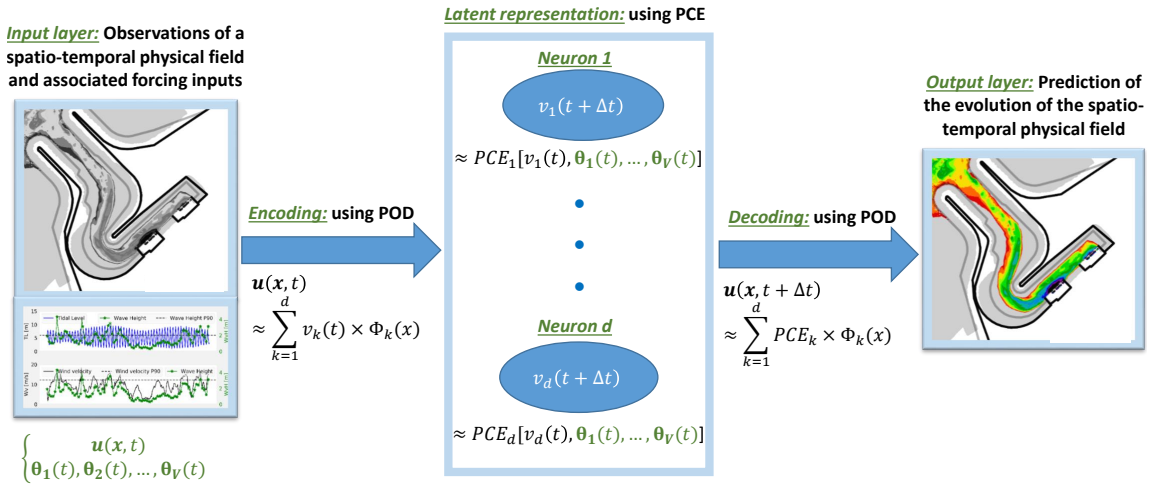


FIGURE 1: Representation of the POD-PCE ML approach.

The proposed POD-PCE addresses these drawbacks of ML.

(*i*) It is explicit and simple to implement, as it consists of the association of two linear decompositions. POD is a linear separation of the spatiotemporal patterns [66], shown to be accurate for both linear and non-linear problems [101], combining and relevance. PCE is a well-established method in Uncertainty Quantification (UQ) [114, 112, 65, 100, 107], widely used for the study of stochastic behavior in physics [8, 102, 80, 45]. It is a linear polynomial expansion that allows non-linearities to be gradually added to the model by increasing the polynomial degree. The linearity and orthonormality of the POD and PCE components and the probabilistic framework of PCE make the output's statistical moments easier to study [63, 49, 99, 40], enabling straightforward physical interpretation of the model [10, 52, 24].

(*ii*) It only has two hyper-parameters: a number of POD components, and a PCE polynomial degree. Both can be chosen according to quantitative criteria [14, 6]. All other forms of parameterization (choice of the polynomial basis) can be achieved with robust physical and/or statistical arguments [65, 107, 97], as assessed in the present paper. Furthermore, the orthonormality of the POD and PCE bases minimizes the number of components necessary to capture essential variations in data. Additionally, the POD modes capture more energy than any other decomposition [73, 16], PCE is known to exponentially converge with polynomial degree [58, 113], and the cardinality of the latter can be reduced by sparse basis selection [6].

(*iii*) It can be considered as a universal expansion for physical field approximation: a physical field has a finite variance, which implies that it belongs to the Hilbert space of random variables with finite second order moments. There therefore exists a numerable set of orthogonal random variables, that form the basis

of this Hilbert space, on which the field of interest can be expanded (strict equality, not approximation) [100]. A mathematical setting for basis construction based on input was established by Soize and Ghanem [97] for the general case of dependent variables with arbitrary density, provided that the set of inputs is finite.

The present study consisted in: i) combined use of POD and PCE in ML for point-wise prediction ; ii) application to field data with the inherent challenges not encountered with numerical data (e.g. paucity) ; iii) a focus on model explicitness as a key condition for physical understanding and iv) the influence of forcing variables study, based on a classical measure of importance (Garson weights [25]) directly computed with the POD-PCE expansion coefficients. Firstly, associating regression techniques to POD, and more generally to Reduced Order Models (ROM), is not novel [53, 76, 108, 32]. The cited studies, however, focused on dimensionality reduction, whereas the explicit formulation and applicability to complex physical processes are emphasized in the present study. Secondly, coupling PCE to ROM was recently addressed [36, 77, 54] and the use of PCE as ML is consistent with the methodology proposed by Torre et al. [103] , where they showed that PCE is as powerful as classical ML techniques, but neither spatiotemporal fields nor physical interpretability were addressed. The data in these studies were either obtained from numerical experiments, emulated from analytical benchmark functions such as Sobol or Ishigami, or based on one-dimensional data sets [103]. Our methodology is assessed on high-resolution two-dimensional field measured data.

The assessment of the proposed methodology was based on the study of sedimentation processes in a cooling water intake located in a coastal area, subject to tide and wave forcing. This application in geosciences is characterized by high non-linearities and is conditioned *a priori* with various parameters [4]. The disparity of the space and time scales and the interaction between various processes make process-based modeling difficult [4, 82]. A complete overview of the use of ML in coastal sediment transport modeling is given in [27].

The paper is organized as follows. Section 2 gives a detailed explanation of the methodology. Section 3 deals with assessment of the methodology based on a physical application. Firstly, the study case and data are described in 3.1. POD and PCE performances are then demonstrated independently in 3.2 with a deep physical analysis using adequate measures. The performance of the POD-PCE predictor is discussed in 3.3. A summary of the study and perspectives of the proposed methodology are presented in Section 4.

2 Theoretical framework

In this section, the objective is to define the framework of the proposed POD-PCE Machine Learning methodology, along with physical influence indicators for the inputs. This is the object of Subsection 2.3, but first, a reminder of the existing POD and PCE theoretical bases is presented in 2.1 and 2.2 respectively.

2.1 Proper Orthogonal Decomposition

POD is a dimensionality reduction technique [66] that is well documented in literature [14, 101, 61, 62]. Theoretical details and demonstrations can be found in descending chronological order in [3, 73, 16]. For clarity's sake, the essential elements of POD are summarized below.

The goal of POD is to extract the spatial patterns of a continuous time and space function. These patterns, when added and multiplied by appropriate temporal coefficients, explain the dynamics of the variable of interest: a real-valued physical field.

Let $\mathbf{u} : \Omega \times \mathbb{T} \to \mathbb{D}$ be a continuous function of two variables $(\mathbf{x}, t) \in \Omega \times \mathbb{T}$. The following relationships and properties hold for any $\Omega \times \mathbb{T}$ and Hilbert space $\mathbb{D}$ characterized by its scalar product $(.\ ,.)_{\mathbb{D}}$ and induced norm $||.||_{\mathbb{D}}$. However, and as is the case for a majority of physical fields, we shall consider Ω as a set of spatial coordinates (e.g. $\mathbb{R}^2$ or $\mathbb{R}^3$), $\mathbb{T}$ as a set of time coordinates (e.g. $\mathbb{R}^+$ or subset $[0, T]$), and $\mathbb{D}$ as a set of scalar real values or vector real values (e.g. $\mathbb{R}$ or $\mathbb{R}^2$). POD consists in an approximation of $\mathbf{u}(\mathbf{x}, t)$ at a given order $d \in \mathbb{N}$ (Lumley [66]) as in Equation 1,

$$\mathbf{u}(\mathbf{x}, t) \approx \sum_{k=1}^{d} v_k(t)\boldsymbol{\phi}_k(\mathbf{x}) \ , \tag{1}$$

where $\{v_k(.)\}_{k=1}^d \subset \mathcal{C}(\mathbb{T},\mathbb{R})$ and $\{\boldsymbol{\phi}_k(.)\}_{k=1}^d \subset \mathcal{C}(\Omega,\mathbb{D})$, with $\mathcal{C}(\mathbb{A},\mathbb{B})$ denoting the space of continuous functions defined over $\mathbb{A}$ and arriving at $\mathbb{B}$. The objective of POD is to identify $\{\boldsymbol{\phi}_k(.)\}_{k=1}^d$ that minimizes the distance of the approximation from the true value $\mathbf{u}(.,.)$, over the whole $\Omega \times \mathbb{T}$ domain, with an orthogonality constraint for $\{\boldsymbol{\phi}_k(.)\}_{k=1}^d$ using the scalar product $(.\ ,.)_{\mathbb{D}}$. This can be defined, in the least-squares sense, as a minimization problem.

The minimization problem is defined for all orders $d \in \mathbb{N}$, so that the members $\boldsymbol{\phi}_k$ are ordered according to their importance. In particular, for order 1, $\boldsymbol{\phi}_1$ is the linear generator of the sub-vector space most representative of $\mathbf{u}(\mathbf{x},t)$ in $\mathbb{D}$. For $\mathbb{D} = Im(\mathbf{u})$, the family $\{\boldsymbol{\phi}_k(.)\}_{k=1}^d$ is called the POD basis of $\mathbb{D}$ of rank d. The solution to this problem has already been established in literature [66, 96]. The theoretical aspects of POD and demonstrations of mathematical properties can, for example, be found in [73]: the POD basis of $\mathbb{D}$ of order d is the orthonormal set of eigenvectors of an operator $\mathcal{R} : \mathbb{D} \to \mathbb{D}$ defined as $\mathcal{R}\boldsymbol{\phi} = \langle (\mathbf{u},\boldsymbol{\phi})_{\mathbb{D}} \times \mathbf{u} \rangle_{\mathbb{T}}$, if the eigenvectors are taken in decreasing order of the corresponding eigenvalues $\{\lambda_k\}_{k=1}^d$.

For this expansion, an accuracy rate, also called the Explained Variance Rate (EVR), denoted e_d at rank d, can be calculated as in Equation 2 . EVR tends to 1 (perfect approximation) when $d \to +\infty$.

$$e_d = \frac{\sum_{k \leq d} \lambda_k}{\sum_{k=1}^{+\infty} \lambda_k} \ . \tag{2}$$

In practice, for $\mathbb{D} = \mathbb{R}$, when $\mathbf{u}(.,.)$ is a discrete sample on a set of $m \in \mathbb{N}$ space coordinates $\boldsymbol{\mathcal{X}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and for $n \in \mathbb{N}$ measurement times $\mathcal{T} = \{t_1, \ldots, t_n\}$, the available data set is arranged in a matrix $\mathbf{U}(\mathcal{X},\mathcal{T}) = [\mathbf{u}(\mathbf{x}_i,t_j)]_{i,j} \in \mathbb{R}^{m \times n}$, called the snapshot matrix, so as to be able to work in a discrete space. The POD problem formulated in Equation 1 can be written in its discrete form as $\mathbf{U}(\boldsymbol{\mathcal{X}},\mathcal{T}) = \boldsymbol{\Phi}^{(d)}(\boldsymbol{\mathcal{X}})\mathbf{V}^{(d)}(\mathcal{T})$, where $\boldsymbol{\Phi}^{(d)}(\boldsymbol{\mathcal{X}}) \coloneqq [\boldsymbol{\phi}_j(\mathbf{x}_i)]_{i,j} \in \mathbb{R}^{m \times d}$ and $\mathbf{V}^{(d)}(\mathcal{T}) \coloneqq [v_i(t_j)]_{i,j} \in \mathbb{R}^{d \times n}$. The problem can therefore be viewed as if working with a new function $\mathbf{U}(\boldsymbol{\mathcal{X}},.) = [\mathbf{u}(\mathbf{x}_i,\ .)]_{i \in \{1,\ldots,m\}} : \mathcal{T} \to \mathbb{D} = \mathbb{R}^M$. Then, the average over $\mathbb{T}$ can be defined as the statistical mean over the subset $\mathcal{T}$, and the scalar product $(.\ ,.)_{\mathbb{D}}$ as the canonical product over $\mathbb{R}^m$. The POD operator $\mathcal{R}$ can be written as in Equation 3,

$$\mathcal{R}\boldsymbol{\phi}(\boldsymbol{\mathcal{X}}) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{U}(\boldsymbol{\mathcal{X}},t_j)^T \boldsymbol{\Phi}(\boldsymbol{\mathcal{X}})\mathbf{U}(\boldsymbol{\mathcal{X}},t_j) = \frac{1}{n}\mathbf{U}(\boldsymbol{\mathcal{X}},\mathcal{T})\mathbf{U}(\boldsymbol{\mathcal{X}},\mathcal{T})^T \boldsymbol{\Phi}(\boldsymbol{\mathcal{X}}) \ , \tag{3}$$

where $\mathbf{U}(\mathcal{X},t_j) = [\mathbf{u}(\mathbf{x}_i,\ .)]_{i \in \{1,\ldots,m\}}$ is the column number j of the matrix $\mathbf{U}(\boldsymbol{\mathcal{X}},\mathcal{T})$ (i.e the measurement over $\boldsymbol{\mathcal{X}}$ at time t_j), and $\boldsymbol{\Phi}(\mathcal{X}) = [\boldsymbol{\phi}(\mathbf{x}_i)]_{i \in \{1,\ldots,m\}}$. As finding the POD basis is equivalent to identifying the orthonormal set of eigenvectors of the operator $\mathcal{R}$, then for this discrete representation the problem becomes equivalent to solving the eigen problem of the matrix $\mathbf{R} \coloneqq \frac{1}{n}\mathbf{U}(\boldsymbol{\mathcal{X}},\mathcal{T})\mathbf{U}(\boldsymbol{\mathcal{X}},\mathcal{T})^T$, called the covariance matrix. A number $d \in \mathbb{N}$ of eigen vectors $\boldsymbol{\Phi}(\boldsymbol{\mathcal{X}})$ are identified and stored in the columns of the matrix $\boldsymbol{\Phi}^{(d)}(\boldsymbol{\mathcal{X}})$. For the eigenvalues of the covariance matrix $\mathbf{R}$ denoted $\{\lambda_k\}_{k=1}^d$, the expansion in Equation 1 can also be written as in Equation 4, where $\{\boldsymbol{\phi}_k(.)\}_{k=1}^d$ together with $\{a_k(.)\}_{k=1}^d$ are bi-orthonormal, and $v_k(.) = a_k(.)\sqrt{n \times \lambda_k}$.

$$\mathbf{u}(\mathbf{x},t) \approx \sum_{k=1}^{d} a_k(t)\sqrt{n \times \lambda_k}\boldsymbol{\phi}_k(\mathbf{x}) \ . \tag{4}$$

By defining the matrix $\mathbf{A}^{(d)}(\mathcal{T}) \coloneqq [a_i(t_j)]_{i,j} \in \mathbb{R}^{d \times n}$ and the operator $\mathbf{D}^{(d)}(\lambda_1,...,\lambda_d)$ corresponding to the diagonal matrix of elements λ_i, we have $\mathbf{U}(\boldsymbol{\mathcal{X}},\mathcal{T}) = \boldsymbol{\Phi}^{(d)}(\boldsymbol{\mathcal{X}})\mathbf{D}^{(d)}(\sqrt{n \times \lambda_1},...,\sqrt{n \times \lambda_d})\mathbf{A}^{(d)}(\mathcal{T})$. Therefore the transposed form is $\mathbf{U}(\boldsymbol{\mathcal{X}},\mathcal{T})^T = \mathbf{A}^{(d)}(\mathcal{T})^T\mathbf{D}^{(d)}(\sqrt{n \times \lambda_1},...,\sqrt{n \times \lambda_d})\boldsymbol{\Phi}^{(d)}(\boldsymbol{\mathcal{X}})^T$. Thanks to the orthonormality of $\{a_k(.)\}_{k=1}^d$, the covariance matrix reads $\mathbf{R} = \frac{1}{n}\boldsymbol{\Phi}^{(d)}(\boldsymbol{\mathcal{X}})\mathbf{D}^{(d)}(n\times\lambda_1,...,n\times\lambda_d)\boldsymbol{\Phi}^{(d)}(\boldsymbol{\mathcal{X}})^T = \boldsymbol{\Phi}^{(d)}(\boldsymbol{\mathcal{X}})\mathbf{D}^{(d)}(\lambda_1,...,\lambda_d)\boldsymbol{\Phi}^{(d)}(\boldsymbol{\mathcal{X}})^T$.

When $n << m$, it is more computationally efficient to solve the eigenproblem of $\mathbf{R}^T$ instead of the eigenproblem of $\mathbf{R}$ as highlighted by Sirovich [96] . This is often the case when a two-dimensional physical field is measured over a domain at specific times, and is the case encountered for our application described in Section 3.

When an order $d << min(m,n)$ corresponds to a high EVR as defined in Equation 2, we speak of dimensionality reduction, because the data are projected in a sub-space that is of much smaller dimension than $\mathbb{R}^{m \times n}$. When diverse enough records are available for the variable under study, we may consider that the resulting POD basis $\{\boldsymbol{\phi}_k(\boldsymbol{\mathcal{X}})\}_{k=1}^d = \{[\boldsymbol{\phi}_k(\mathbf{x}_i)]_{i \in \{1,\ldots,m\}}\}_{k=1}^d$ is a generator of all possible states. Predicting the associated temporal

coefficients $\{a_k(t)\}_{k=1}^d$ at a given time t would therefore be enough to predict the whole state at time t. Hence, we propose to use the POD as a spatial basis extractor. This would first enable study of the spatial dynamics of the variable of interest and eventually extraction of physical information, as shown in the application Section 3. Then, the basis can be used as a generator for the prediction of future states. This implies predicting the evolution of $\{a_k(t)\}_{k=1}^d$, for which we propose to use Polynomial Chaos Expansion (PCE), as described in the following Section 2.2.

2.2 Polynomial Chaos Expansion

A reminder of the theoretical base of PCE is presented in Subsection 2.2.1. Theoretical details, demonstrations and interesting references can be found in [6, 99, 114, 57, 58]. After the theoretical introduction, a simple indicator is proposed in Subsection 2.2.2 for the analysis of the variables influence on the output value. The latter is later generalized for POD-PCE in Section 2.3.

2.2.1 Learning

The idea behind Polynomial Chaos Expansion (PCE) is to formulate an explicit model that links a variable of interest (output) to conditioning parameters (inputs), both in a probability space. This enables the propagation path of probabilistic information (uncertainties, occurrence frequencies) to be mapped from the input space to the output space. The variable of interest, $\mathbf{Y}$, and the input parameters $\mathbf{\Theta} = (\theta_1, \theta_2, ..., \theta_V)$ are therefore considered random variables, characterized by a given Probability Density Function (PDF) denoted $f_{\mathbf{\Theta}}$. It should be kept in mind that the outputs of our problem are the temporal coefficients $\mathbf{Y} = [a_k(t)]_{k\in\{1,...,d\}}$ generated by POD, and that among the inputs there can be a set of physical forcings, as described later in Section 3. The objective is to derive the evolution of the temporal coefficients as the outcome of the forcings. Let us now recall some fundamentals of the mathematical probabilistic framework, taking the example of a one dimensional real-valued variable. The definitions can be easily extended to $\mathbb{R}^M$.

Let $(\Omega, F, \mathbb{P})$ be a probability space, where Ω is the event space (space of all the possible events ω) equipped with σ-algebra F (some events of Ω) and its probability measure $\mathbb{P}$ (likelihood of a given event occurrence). A random variable defines an application $Y(\omega) : \Omega \to D_Y \subseteq \mathbb{R}$, with realizations denoted by $y \in D_Y$. The PDF of Y is a function $f_Y : D_Y \to \mathbb{R}$ that verifies $\mathbb{P}(Y \in E \subseteq D_Y) = \int_E f_Y(y)dy$. The k^{th} *moments* of Y are defined as $\mathbb{E}[Y^k] \coloneqq \int_{D_Y} y^k f_Y(y)dy$, the first being the expectation denoted $\mathbb{E}[Y]$. In the same manner, we define the k^{th} *central moments* of Y as $\mathbb{E}[(Y - \mathbb{E}[Y])^k]$, the first being 0 and the second the variance of Y denoted by $\mathbb{V}[Y]$. The covariance of two random variables is defined as $cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$ and a resulting property is $\mathbb{V}[Y] = cov(Y, Y)$.

Returning to the PCE construction, inputs $\mathbf{\Theta} = (\theta_1, \theta_2, ..., \theta_V)$ are considered to live in the space of real random variables with finite second moments (and finite variances). This space is denoted by $\mathcal{L}^2_{\mathbb{R}}(\Omega, F, \mathbb{P}; \mathbb{R})$ and is a Hilbert space equipped with the inner product $(\theta_1, \theta_2)_{\mathcal{L}^2_{\mathbb{R}}} \coloneqq \mathbb{E}[\theta_1\theta_2] = \int_\Omega \theta_1(\omega)\theta_2(\omega)d\mathbb{P}(\omega)$ and its induced norm $||\theta_1||_{\mathcal{L}^2_{\mathbb{R}}} \coloneqq \sqrt{\mathbb{E}[\theta_1^2]}$. The PCE objective is to map the output space from the input space with a model $\mathcal{M}$ as in Equation 5:

$$\begin{aligned} Y &= \mathcal{M}(\mathbf{\Theta}) = \textstyle\sum_{\mathcal{I}\subseteq\{1,...,V\}} \mathcal{M}_{\mathcal{I}}(\theta_{\mathcal{I}}) \\ &= \mathcal{M}_0 + \textstyle\sum_{i=1}^V \mathcal{M}_i(\theta_i) + \sum_{1\le i<j\le V} \mathcal{M}_{i,j}(\theta_i, \theta_j) + ... + \mathcal{M}_{1,...,V}(\theta_1, \theta_2, ..., \theta_V) \ , \end{aligned} \tag{5}$$

where $\mathcal{M}_0$ is the expectation of Y and $\mathcal{M}_{\mathcal{I}\subseteq\{1,...,V\}}$ represents the common contribution of the variables $\mathcal{I} \subseteq \{1, ..., V\}$ to the variation in Y. For the PCE model, these contributions have a polynomial form. We shall define, for each input variable θ_i, an orthonormal univariate polynomial basis $\left\{\xi_\beta^{(i)}(.), \beta \in [|0, p|]\right\}$ where $p \in \mathbb{N}$ is a chosen polynomial degree and $\xi_\beta^{(i)}(.)$ is of degree β. The orthonormality is defined with respect to the inner product $(.\ ,\ .)_{\mathcal{L}^2_{\mathbb{R}}}$. If we introduce the multi-index notation $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_V) \in \mathbb{N}^V$ so that $|\boldsymbol{\alpha}| = \sum_{i=1}^V \alpha_i$, we can define a multivariate basis $\left\{\zeta_{\boldsymbol{\alpha}}^{\mathbf{\Theta}}(.), |\boldsymbol{\alpha}| \in [|0, p|]\right\}$ as $\zeta_{\boldsymbol{\alpha}}^{\mathbf{\Theta}}(\theta_1, \theta_2, ..., \theta_V) \coloneqq \prod_{i=1}^V \xi_{\alpha_i}^{(i)}(\theta_i)$. Therefore, the model in Equation 5 can be written as:

$$Y = \mathcal{M}(\mathbf{\Theta}) = \sum_{|\boldsymbol{\alpha}|\le P} c_{\boldsymbol{\alpha}} \zeta_{\boldsymbol{\alpha}}^{\mathbf{\Theta}}(\theta_1, \theta_2, ..., \theta_V) \ , \tag{6}$$

where $c_{\boldsymbol{\alpha}} \in \mathbb{R}$ are deterministic coefficients that can be estimated thanks to different methods. It can be formulated as a minimization problem, and regularization methods can be used when dealing with small data sets. In the present study, we used the Least Angle Regression Stagewise method (LARS) in order to construct an adaptive sparse PCE. In this approach, a collection of possible PCE, ordered by sparsity, is provided and an optimum can be chosen with an accuracy estimate. It was performed in this study using corrected leave-one-out error [7]. The reader can refer to the work of Blatman [6] for further details on LARS and more generally on sparse constructions.

The choice of the basis is crucial and is directly related to the choice of input variable marginals, via the inner product $(.,.)_{\mathcal{L}^2_{\mathbb{R}}}$. Chaos polynomials were first introduced in [109] for input variables characterized by Gaussian distributions. The orthonormal basis with respect to this marginal is the Hermite polynomials family. Later, other Askey scheme hypergeometric polynomial families were associated to some well-known parametric distributions [113]. For example, the Legendre family is orthonormal with respect to the Uniform marginals. This is called *gPC* (generalized Polynomial Chaos) when variables of different PDFs are used as inputs. In practice however, the input distributions of physical variables can be different from usual parametric marginals. In such cases, the marginals can be inferred by empirical methods such as the Kernel Smoother (see [35] for theoretical elements). In this case, an orthonormal polynomial basis with respect to arbitrary marginals can be built with a Gram-Schmidt orthonormalization process as in [111] or via the Stieltjes three-term recurrence procedure as in [107].

To highlight the importance of the marginals and choice of polynomial basis for the learning process, several configurations are attempted in Section 3. Different input sets and distributions (Gaussian, Uniform, inferred by Kernel Smoothing) were tested. The influence of the polynomial basis on the ML is investigated in Section 3.2.2.

2.2.2 Physical importance measures

Once the PCE construction is achieved, a physical interpretation can be performed. It is notable that classical NN indicators can be used [25]. The PCE can be represented in the Feedforward NN paradigm as in Figure 2. Such networks are classically composed, in addition to the input and output layers, of successive *hidden layers*. Each hidden layer is composed of *neurons* that transform the variables of the previous layer (outputs of the previous *neurons*) into a new set of variables. This is done by combining a linear transformation, giving different *weights* to the previous *neurons*, and a transformation function, called *Activation Function* (AF). This succession of layers is called the *latent representation*. For a number of hidden layers $L \geq 1$, the latent representation, in addition to the input and output layers, can be formally written as $\mathbf{Y} \approx f_{out}(\mathbf{A}_L\ f_L(\dots \mathbf{A}_2\ f_2(\mathbf{A}_1\ f_1(\mathbf{A}_{in}\ \boldsymbol{\Theta}))))$ where $\{\mathbf{A}_k\}_{k\in[|1,L|]}$ and $\{f_k\}_{k\in[|1,L|]}$ are the hidden layer weight matrices and AFs, $\mathbf{A}_{in}$ is the input-to-hidden connection matrix and f_{out} is the final hidden-output transformation [59, 95].

The PCE-based NN represented in figure 2 is a single layer feedforward, composed of $l \in \mathbb{N}$ neurons, that can be written as $\mathbf{Y} \approx f_{out}(\mathbf{A}_1\ f_1(\mathbf{A}_{in}\boldsymbol{\Theta}))$. The first matrix $\mathbf{A}_{in}$ is the input-to-hidden connection matrix of dimension $V \times V$, that links the input layer to the PCE hidden layer containing the multivariate polynomials $\left\{\zeta^{\boldsymbol{\Theta}}_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \{\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_l\}\right\}$, where V is the number of inputs and the multivariate indexes $\{\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_l\}$ are conditioned by the chosen polynomial degree p such as $\forall i \in [|1, l|]\ 0 \leq |\boldsymbol{\alpha}_i| \leq p$, and by the number of selected features if a sparse polynomial is constructed, as in the present case using LARS [6].

Matrix $\mathbf{A}_{in}$ represents the contributions of the V variables to the multivariate polynomials $\left\{\zeta^{\boldsymbol{\Theta}}_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \{\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_l\}\right\}$. It is a diagonal matrix such that $[\boldsymbol{A}_{in}]_{j,j\in[|1,V|]^2}$ is 0 if $\forall i \in [|1, l|]\ (\alpha_i)_j = 0$ and 1 if not. The first multidimensional AF f_1 is a vector of multivariate functions that transforms the set of selected inputs corresponding to $[\boldsymbol{A}_{in}]_{i,i\in[|1,V|]^2} = 1$ to the multivariate polynomials of the chosen basis (Hermite, Legendre, etc.) by tensor product over the univariate basis. The hidden layer weight matrix $\mathbf{A}_1$ gives different weights to the constructed polynomial features. It is a diagonal matrix composed of the PCE expansion coefficients such as $[\mathbf{A}_1]_{i,j\in[|1,l|]^2} = [c_{\boldsymbol{\alpha}_i}]_{i\in[|1,l|]}$.

The final hidden-output transformation f_{out} is a summation. Figure 2 can also be presented differently: another hidden layer can be added to the PCE latent representation as $\mathbf{Y} \approx f_{out}(\mathbf{A}_2\ f_2(\mathbf{A}_1\ f_1(\mathbf{A}_{in}\ \boldsymbol{\Theta}))$. The first AF f_1 would represent a transformation of each input variable to a list of monomials of degrees 1 to p (here, $\mathbf{A}_{in}$ is identity). The second AF f_2 therefore represents the tensor product that transforms the different monomials to multivariate features, with $\mathbf{A}_1$ appropriately filled with zeros and ones, and $[\mathbf{A}_2]_{i,j\in[|1,l|]^2} = [c_{\boldsymbol{\alpha}_i}]_{i\in[|1,l|]}$.

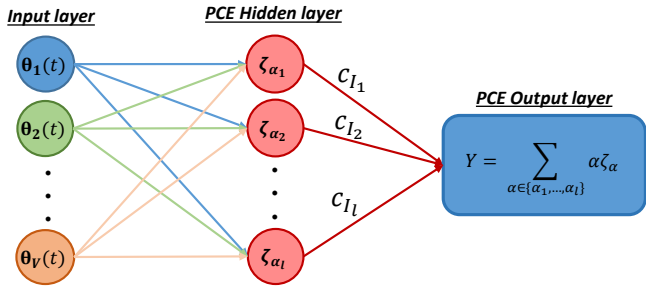


FIGURE 2: Representation of the PCE learning in the NN paradigm.

To capture the importance of each feature, the Garson relative Weights (GW) defined in Equation 7 are a classical measure to quantify the relative importance of each neuron of the last hidden layer, and therefore of each polynomial pattern, for the output value [25, 39, 104].

$$w_{\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}} = \frac{|c_{\boldsymbol{\alpha}}|}{\sum_{0\leq\boldsymbol{\beta}\leq 1}|c_{\boldsymbol{\beta}}|} \,. \tag{7}$$

This measure can be used to understand the importance given by the NN algorithm to the variables and their possible interactions, especially when using feature selection algorithms as LARS: "feature interactions [...] are created at hidden units with nonlinear activation functions, and the influences of the interactions are propagated layer-by-layer to the final output" [104]. In the particular case of a polynomial expansion, the interpretation is straightforward, the importance of each variable alone corresponds to its monomials, and the importance of its interactions with other variables corresponds to the multivariate polynomials in which it is involved.

For the particular case of the orthonormal basis provided by PCE, the GW defined in 7 can be interpreted in terms of Pearson's correlations between output Y and the polynomial basis elements $\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}$ denoted $\rho_{Y,\,\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}}$, with $\alpha \neq (0,...,0)$. Indeed, Pearson's correlations $\rho_{Y,\,\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}}$ are defined as in Equation 8,

$$\rho_{Y,\,\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}} = \frac{\mathbb{E}\left[(Y-\mathbb{E}(Y))(\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}-\mathbb{E}(\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}))\right]}{\sqrt{\mathbb{V}(Y)\mathbb{V}(\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}})}} = \frac{c_{\boldsymbol{\alpha}}}{\sqrt{\sum_{1\leq|\boldsymbol{\beta}|\leq p} c_{\boldsymbol{\beta}}^2}} \,, \tag{8}$$

thanks to the orthonormality of the basis with respect to the scalar product $(.\,,.)_{\mathcal{L}^2_{\mathbb{R}}}$ that guarantees:

- $\mathbb{E}\left[\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}\right] = \left(\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}, \zeta_{\boldsymbol{\beta}=(0,\dots,0)}^{\boldsymbol{\Theta}} = 1\right)_{\mathcal{L}^2_{\mathbb{R}}} = 0\,;$
- $\mathbb{E}\left[Y\right] = \left(\sum_{0\leq|\boldsymbol{\beta}|\leq p} c_{\boldsymbol{\beta}}\zeta_{\boldsymbol{\beta}}^{\boldsymbol{\Theta}}\,, \zeta_{\boldsymbol{\beta}=(0,\dots,0)}^{\boldsymbol{\Theta}}\right)_{\mathcal{L}^2_{\mathbb{R}}} = c_{\beta=(0,\dots,0)}\,;$
- $\mathbb{E}\left[Y, \zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}\right] = \left(\sum_{0\leq|\boldsymbol{\beta}|\leq p} c_{\boldsymbol{\beta}}\zeta_{\boldsymbol{\beta}}^{\boldsymbol{\Theta}}\,, \zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}\right)_{\mathcal{L}^2_{\mathbb{R}}} = c_{\boldsymbol{\alpha}}\,;$
- $\mathbb{V}\left[\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}\right] = \mathbb{E}\left[\left(\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}} - \mathbb{E}\left[\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}\right]\right)^2\right] = \mathbb{E}\left[\left(\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}\right)^2\right] = ||\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}||^2_{\mathcal{L}^2_{\mathbb{R}}} = 1\,;$
- $\mathbb{V}\left[Y\right] = \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y\right]\right)^2\right] = \left(\sum_{1\leq|\boldsymbol{\beta}|\leq p} c_{\boldsymbol{\beta}}\zeta_{\boldsymbol{\beta}}^{\boldsymbol{\Theta}}\,, \sum_{1\leq|\boldsymbol{\beta}|\leq p} c_{\boldsymbol{\beta}}\zeta_{\boldsymbol{\beta}}^{\boldsymbol{\Theta}}\right)_{\mathcal{L}^2_{\mathbb{R}}} = \sum_{1\leq|\boldsymbol{\beta}|\leq p} c_{\boldsymbol{\beta}}^2\,.$

Therefore, the weights $w_{\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}}$ can also be computed as $|\rho_{Y,\,\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}}| / \sum_{1\leq|\boldsymbol{\beta}|\leq p}|\rho_{Y,\,\zeta_{\boldsymbol{\beta}}^{\boldsymbol{\Theta}}}|$. This means that they measure the relative importance of the basis element in the expansion of the output, in terms of linear correlation, regardless of the sign of the latter. These "relative Pearson's correlations" can be seen as a physical contribution since the PCE model is strictly linear.

The sum of the GW $w_{\zeta_{\boldsymbol{\alpha}}^{\boldsymbol{\Theta}}}$ for all the polynomial features equals 1. This means that they allow $\{\zeta_{\boldsymbol{\alpha}}\}_{|\boldsymbol{\alpha}|\leq p}$ to be ranked in terms of relative contribution to the output Y. The contributions can be analyzed either for each polynomial pattern separately, or for a single variable θ_i by adding all the polynomial shares related to this variable alone (1st Sobol indice analogy [99]) or by adding all the polynomial shares related to this variable and its interactions (total Sobol indice analogy [99]).

2.3 POD-PCE based predictor

POD and PCE were introduced separately in Subsections 2.1 and 2.2 respectively. We are now fully equipped with the adequate theoretical basis and mathematical notations, to present the POD-PCE ML methodology for a data-based model learning of a physical spatiotemporal field. In this Subsection, we will first summarize the proposed approach, then the formal details of the coupling will be given with the definition of adequate accuracy measures. Finally the previously discussed importance measures will be generalized for the POD-PCE physical study.

The proposed POD-PCE ML consists of five steps, in a learning and a prediction phase, summed up as follows:

- Learning phase:
  - ∗ POD basis construction: given a set of measurements stored in the snapshot matrix $\mathbf{U}(\mathcal{X}, \mathcal{T}) = [\mathbf{u}(\mathbf{x}_i, t_j)]_{i,j} \in \mathbb{R}^{m \times n}$, construct a spatial POD basis accordingly ;
  - ∗ PCE learning: construct PCE models that map each POD temporal coefficient, obtained in the previous step along with the spatial basis, from time $t_j \in \mathcal{T}$ to time $t_{j+1} \in \mathcal{T}$, using the realizations of associated input variables denoted $\mathbf{\Theta}(t_j \to t_{j+1})$ on the set $\mathcal{T}$;
- Prediction phase:
  - ∗ POD projection: given a new measurement of the physical field $\mathbf{U}(\boldsymbol{\mathcal{X}}, t_k)$, obtain the values of the POD temporal coefficients using appropriate projection ;
  - ∗ PCE prediction: given the constructed PCE models and an estimate of the inputs from current time t_k to future time t_{k+1} denoted $\mathbf{\Theta}(t_k \to t_{k+1})$, calculate an estimation of the future POD temporal coefficients ;
  - ∗ POD-PCE ML prediction: reconstruct the estimate of the future state $\mathbf{U}(\boldsymbol{\mathcal{X}}, t_{k+1})$ by calculating the expansion on the POD basis, using the estimate obtained for the future temporal coefficients.

2.3.1 Machine learning methodology

Here, the formal hypothesis behind the POD-PCE ML reasoning and its mathematical formulation are discussed. Let $\mathbf{U}(\boldsymbol{\mathcal{X}}, .) = [\mathbf{u}(\mathbf{x}_i, .)]_{i \in \{1,...,m\}}$ be a field of interest defined on a set of $m \in \mathbb{N}$ space coordinates $\boldsymbol{\mathcal{X}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. Let $\mathbf{\Theta(.)} = (\theta_1(.), \theta_2(.), ..., \theta_V(.))$ be a vector of the inputs supposed to condition the evolution of $\mathbf{U}(\boldsymbol{\mathcal{X}}, .)$ over time. The dynamic model, denoted $\mathcal{H}$, that gives an estimation of a future state $\mathbf{U}(\boldsymbol{\mathcal{X}}, t_{j+1})$ from a past state $\mathbf{U}(\boldsymbol{\mathcal{X}}, t_j)$ and an estimation of $\mathbf{\Theta}(t_j \to t_{j+1})$ over the time interval $[t_j, t_{j+1}]$, where $t_j < t_{j+1} \in \mathbb{R}^+$, is formulated as in Equation 9 .

$$\mathbf{U}(\boldsymbol{\mathcal{X}}, t_{j+1}) \approx \mathcal{H}\left[\mathbf{U}(\boldsymbol{\mathcal{X}}, t_j), t_{j+1} - t_j, \mathbf{\Theta}(t_j \to t_{j+1})\right] \ . \tag{9}$$

If the field of interest has been recorded over a set of past times $\mathcal{T} = \{t_1, \ldots, t_n\} \subset \mathbb{R}^+$, where $t_n < t_j < t_{j+1}$, a POD basis can be constructed as in Section 2.1, consisting of $d \in \mathbb{N}$ vectors of dimension m stored in a matrix as $\mathbf{\Phi}^{(d)}(\boldsymbol{\mathcal{X}}) = (\mathbf{\Phi}_1^{(d)}(\boldsymbol{\mathcal{X}}), \ldots, \mathbf{\Phi}_d^{(d)}(\boldsymbol{\mathcal{X}})) \in \mathbb{R}^{m \times n}$, and can be seen as a generator of all possible states if enough records are available. If so, any future state $\mathbf{U}(\boldsymbol{\mathcal{X}}, t_j)$ can be expanded on this POD basis and the associated temporal coefficients are simply the weights of $\mathbf{U}(\boldsymbol{\mathcal{X}}, t_j)$ on the POD basis. They are therefore obtained using the canonical scalar product over $\mathbb{R}^m$, as in Equation 10.

$$\begin{aligned} \mathbf{U}(\boldsymbol{\mathcal{X}}, t_j) &\approx \sum_{k=1}^{d} a_k(t_j) \sqrt{n \times \lambda_k} \mathbf{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}}) \\ &\approx \sum_{k=1}^{d} (\mathbf{U}(\boldsymbol{\mathcal{X}}, t_j), \mathbf{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}}))_{\mathbb{R}^m} \mathbf{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}}) \\ &\approx \sum_{k=1}^{d} \mathbf{U}(\boldsymbol{\mathcal{X}}, t_j)^T \mathbf{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}}) \mathbf{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}}) \ . \end{aligned} \tag{10}$$

Hence, the variable part of $\mathbf{U}(\boldsymbol{\mathcal{X}}, t_j)$ is fully expressed in the temporal coefficients $a_k(t_j)$. The field of interest $\mathbf{U}(\boldsymbol{\mathcal{X}}, t_j)$ can be either a field measurement, a laboratory or a numerical experiment. In any-case, it can be considered as being generated by a random process "in the sense that nature happens without consideration of what could be the best realizations for the learning algorithm" [95]. Therefore, the coefficients $a_k(t_j)$ can also be seen as the j^{th} realization of a random variable A_k. We can therefore construct a PCE approximation $\mathcal{H}_k$ that maps A_k from its input space. The latter is taken as a collection of random variables, composed from the set $(A_1, ..., A_d)$ at a previous time, the duration of the dynamic, and the input variables $\mathbf{\Theta}(t_j \to t_{j+1})$. This is formulated as a classical dynamic model in Equation 11 .

$$a_k(t_{j+1}) \approx \mathcal{H}_k\left[a_1(t_j), \ldots, a_d(t_j), t_2 - t_1, \mathbf{\Theta}(t_j \to t_{j+1})\right] \ . \tag{11}$$

The model $\mathcal{H}$ in Equation 9 is approximated as in Equation 12 .

$$\mathcal{H}\left[\mathbf{U}(\boldsymbol{\mathcal{X}},t_j),t_{j+1}-t_j,\boldsymbol{\Theta}(t_j\rightarrow t_{j+1})\right]\approx\sum_{k=1}^{d}\mathcal{H}_k\left[a_1(t_j),\ldots,a_d(t_j),t_{j+1}-t_j,\boldsymbol{\Theta}(t_j\rightarrow t_{j+1})\right]\sqrt{n\times\lambda_k}\boldsymbol{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}})\ . \quad (12)$$

The choice of input variables for regression models is an ongoing research question in statistics [33, 81]. The chosen input variables can be extracted from a transformed version of a larger input set with the help of PCA [44] for DR. However, this approach was not studied here and will be the topic of a future study. Different input configurations will be evaluated, to investigate the influence of variable selection on the proposed learning. For example, the hypothesis of dependence between the random variables $(A_1,\ldots,A_d)$ could be relaxed. This would imply writing the approximation in Equation 11 in a relaxed form as $\mathcal{H}_k\left[a_k(t_j),t_2-t_1,\boldsymbol{\Theta}(t_j\rightarrow t_{j+1})\right]$. In that case a simpler model $\mathcal{H}$, under the strong independence assumption, can be formulated as in Equation 13.

$$\mathcal{H}\left[\mathbf{U}(\boldsymbol{\mathcal{X}},t_j),t_{j+1}-t_j,\boldsymbol{\Theta}(t_j\rightarrow t_{j+1})\right]\approx\sum_{k=1}^{d}\mathcal{H}_k\left[a_k(t_j),t_{j+1}-t_j,\boldsymbol{\Theta}(t_j\rightarrow t_{j+1})\right]\sqrt{n\times\lambda_k}\boldsymbol{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}})\ . \quad (13)$$

Both alternatives are tested in Section 3. To investigate the influence of input selection on learning accuracy, a quantitative evaluation of the hypothesis is needed. More generally, whether for the above-mentioned simplifications or for the approximated form of the model in general, accuracy estimators are needed. These are presented below.

2.3.2 Accuracy tests for the approximation

There are two determining parts in the POD-PCE learning process. Firstly, the PCE learning $\mathcal{H}_k(.)$ of each mode A_k should be as accurate as possible. Secondly, the reconstructed field $\sum_{k=1}^{d}\mathcal{H}_k(.)\sqrt{n\times\lambda_k}\boldsymbol{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}})$ for a given rank d should be as close to the real field $\mathbf{U}(\boldsymbol{\mathcal{X}})$ as possible.

The distance between each mode and its PCE approximate can be evaluated using the *generalization error*, denoted $\delta(A_k,\mathcal{H}_k)$ and defined as in Equation 14.

$$\delta(A_k,\mathcal{H}_k)=\mathbb{E}\left[(A_k-\mathcal{H}_k(.))^2\right]\ . \quad (14)$$

For the model defined in Equation 13 , this error can be estimated, on a set of paired realizations $(a_k(t_1),\ldots,a_j(t_n))$ and $(\boldsymbol{\Theta}(t_0\rightarrow t_1),\ldots,\boldsymbol{\Theta}(t_{n-1}\rightarrow t_n))$, as in Equation 15 as explained by Blatman [6] . This approximated version of the *generalization error* is called the *empirical error*.

$$\delta(A_k,\mathcal{H}_k)\approx\delta_{emp}(A_k,\mathcal{H}_k)\coloneqq\frac{1}{n}\sum_{j=1}^{n}\left(a_k(t_j)-\mathcal{H}_k\left[a_k(t_{j-1}),t_2-t_1,\boldsymbol{\Theta}(t_{j-1}\rightarrow t_j)\right]\right)^2\ . \quad (15)$$

Its relative estimate denoted $\epsilon_{emp}(A_k,\mathcal{H}_k)$ can be defined as in Equation 16 .

$$\epsilon_{emp}(A_k,\mathcal{H}_k)\coloneqq\frac{\delta_{emp}(A_k,\mathcal{H}_k)}{\mathbb{V}[A_k]}\ . \quad (16)$$

Once the PCE learnings can be trusted, the distance at time t_j between the true state $\mathbf{U}(\boldsymbol{\mathcal{X}},t_j)$ and the POD-PCE approximation $\mathcal{H}\left[\mathbf{U}(\boldsymbol{\mathcal{X}},t_j),t_{j+1}-t_j,\boldsymbol{\Theta}(t_j\rightarrow t_{j+1})\right]$ can be defined. It might be estimated using the relative Root Mean Squared Error (relative RMSE), denoted $r[\mathbf{U},\mathcal{H}](t_j)$ and calculated as in Equation 17 , where $\mathbf{h}(\mathbf{x}_i,t_j)$ refers to the value of the POD-PCE approximation at coordinate $\mathbf{x}_i$ and time t_j.

$$r[\mathbf{U},\mathcal{H}](t_j)\coloneqq\sum_{i=1}^{m}\frac{\left(\mathbf{u}(\mathbf{x}_i,t_j)-\mathbf{h}(\mathbf{x}_i,t_j)\right)^2}{\left[\mathbf{u}(\mathbf{x}_i,t_j)\right]^2}\ . \quad (17)$$

A mean value of the relative RMSE is calculated over a set of realizations corresponding to a set of times $\mathcal{T}=\{t_1,\ldots,t_n\}$. It is denoted $r[\mathbf{U},\mathcal{H}]^{(\mathcal{T})}$ and estimated as in Equation 18.

$$r[\mathbf{U},\mathcal{H}]^{(\mathcal{T})}\coloneqq\frac{1}{n}\sum_{j=1}^{n}r[\mathbf{U},\mathcal{H}](t_j)\ . \quad (18)$$

Once the accuracies of the PCE learnings and the POD-PCE coupling have been evaluated, a final model, which will be the most accurate one, can be chosen. This model would, for our ML set-up, be the best representation of the dependence structure between inputs and outputs. It is used to shed light on the underlying physical relationships. Therefore the inputs are ranked in terms of physical influence, using an appropriate ranking indicator, presented in the following Subsection.

2.3.3 Physical influence of inputs based on the POD-PCE model

The GW influence measures presented for the PCE models in Subsection 2.2 are here extended for the POD-PCE coupling. These indicators are adequate for the analysis of each PCE model $\mathcal{H}_k$: i.e., for interpreting the contribution of the inputs to each random variable A_k separately. However, calculating the contributions to each A_k independently precludes putting them in perspective according to the importance of A_k in the final reconstructed model $\mathcal{H}$ that approximates $\mathbf{U}(\boldsymbol{\mathcal{X}}, .)$. Hence, adapted indicators should be calculated.

Let $\mathbf{U}(\boldsymbol{\mathcal{X}}, .)$ be the random spatiotemporal field approximated by the POD-PCE ML, for prediction from time t_j to time t_{j+1} and let $\mathcal{H}_k$ be the PCE approximation at degree $p^{(k)}$ that maps the random POD temporal coefficient A_k from a set of input variables, using the expansion on the multivariate polynomial basis $\left\{\zeta_{\boldsymbol{\alpha}}^{(k)(.)}\right\}_{|\boldsymbol{\alpha}|\leq p^{(k)}}$. The POD-PCE model formulated in Equation 12 is written as in Equation 19:

$$\mathbf{U} \approx \sum_{k=1}^{d} A_k \sqrt{n \times \lambda_k} \mathbf{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}}) \approx \sum_{k=1}^{d} \left(\sum_{|\boldsymbol{\alpha}|\leq p^{(k)}} c_{\boldsymbol{\alpha}}^{(k)} \zeta_{\boldsymbol{\alpha}}^{(k)}(.) \right) \sqrt{n \times \lambda_k} \mathbf{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}}) \ . \tag{19}$$

Thanks to its linearity, the POD-PCE ML can be represented as a single-layered NN, as shown in Figure 3.

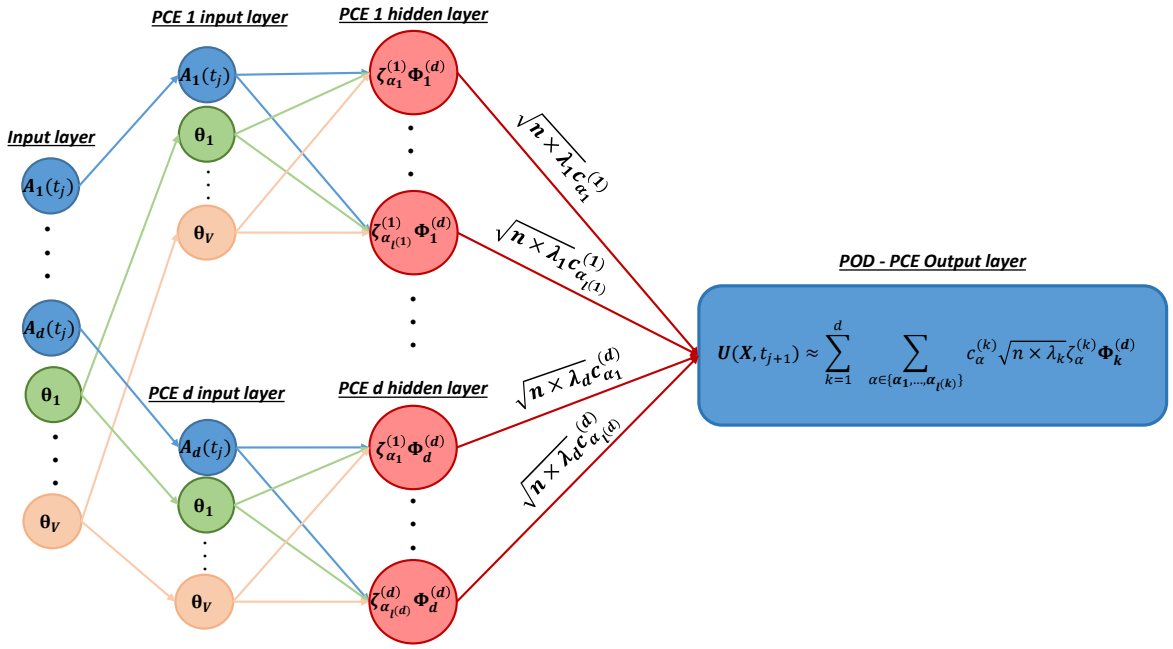


FIGURE 3: Representation of the POD-PCE ML approach in the NN paradigm.

Therefore, a new indicator, *Generalized Garson Weights* (GGW), denoted $W_{\zeta_{\boldsymbol{\alpha}}^{(k)}}$, is computed and simply re-evaluated from the PCE Garson weights (GW), here denoted $w_{\zeta_{\boldsymbol{\alpha}}^{(k)}}$, as in Equation 20.

$$\begin{aligned} W_{\zeta_{\boldsymbol{\alpha}}^{(k)}} &:= \frac{|c_{\boldsymbol{\alpha}}^{(k)}| \sqrt{n \times \lambda_k}}{\sum_{e=1}^{d} \sum_{|\boldsymbol{\beta}|\leq p^{(e)}} \left(|c_{\boldsymbol{\beta}}^{(e)}| \sqrt{n \times \lambda_e} \right)} \\ &= \frac{\left(\sum_{|\boldsymbol{\beta}|\leq p^{(k)}} |c_{\boldsymbol{\beta}}^{(k)}| \right) w_{\zeta_{\boldsymbol{\alpha}}^{(k)}} \sqrt{\lambda_k}}{\sum_{e=1}^{d} \sum_{|\boldsymbol{\beta}|\leq p^{(e)}} \left(|c_{\boldsymbol{\beta}}^{(e)}| \sqrt{\lambda_e} \right)} = \left(\frac{\sum_{|\boldsymbol{\beta}|\leq p^{(k)}} \left(|c_{\boldsymbol{\beta}}^{(k)}| \sqrt{\lambda_k} \right)}{\sum_{e=1}^{d} \sum_{|\boldsymbol{\beta}|\leq p^{(e)}} \left(|c_{\boldsymbol{\beta}}^{(e)}| \sqrt{\lambda_e} \right)} \right) w_{\zeta_{\boldsymbol{\alpha}}^{(k)}} \ . \end{aligned} \tag{20}$$

These GGW indicators show that the contribution of the polynomials $\{\zeta_{\boldsymbol{\alpha}}^{(k)}\}_{|\boldsymbol{\alpha}|\leq p^{(k)}}$ of A_k are enhanced with the eigenvalue λ_k, which is directly linked to the importance of the POD mode $\mathbf{\Phi}_k^{(d)}(\boldsymbol{\mathcal{X}})$ (EVR in Equation 2). An analogy can be drawn with the generalization of Sobol indices for a reduced order model [52, 24]. The property

$\sum_{k=1}^{d}\sum_{|\boldsymbol{\alpha}|\leq p^{(k)}} W_{\zeta_{\underline{\alpha}}^{(k)}} = 1$ holds. This means that the indices allow $\{\{\zeta_{\boldsymbol{\alpha}}^{(k)}\}_{|\boldsymbol{\alpha}|\leq p^{(k)}}\}_{k\in\{1,\dots,d\}}$ to be ranked altogether in terms of contribution to output $\mathbf{U}$. The contributions can be analyzed either for each polynomial pattern separately, or for a single variable θ_i by adding all the polynomial shares related to this variable alone or by adding all the polynomial shares related to this variable and its interactions (analogy with first and total Sobol indices respectively [99]).

3 Application to a cooling water intake in a coastal zone

This section deals with the application of the POD-PCE ML described in Section 2 to a physical problem, introduced, with an industrial study case and inherent challenges, in Subsection 3.1. The physics and data are described. Subsection 3.2 deals with application of the POD-PCE learning phase to the data and assessment of accuracy and robustness with respect to the numerical choices (data set, inputs, marginals and polynomial basis). Finally, the prediction phase using POD-PCE is dealt with in Subsection 3.3, and the ability of the proposed ML to predict mean quantities and spatial details is demonstrated.

3.1 Study case

Sedimentation processes in nearshore areas can be responsible for the excessive sediment deposition commonly observed in cooling water intakes in power plants. As a result, the carrying capacity of the water intake can be drastically reduced, by decreasing its effective area of transport [98].

Cooling water intakes usually incorporate jetties, of which the angle with the shoreline and position relative to the direction of the net longshore sediment transport influence the amount of sediments diverted into the channel inlet by waves and tidal currents. Jetties also reduces littoral drift, resulting in localized sediment accretion against the shore-normal structure due to the longshore sediment transport being trapped by the jetty [19]. In addition, a return current is prone to develop, in the form of a swirling vortex at the end of the structure, and can induce sediment deposition in the vicinity of the channel entrance, consequently affecting the amount sediment delivered into the cooling water intake [15].

Therefore, effective water intake management involves frequent dredging, with high operational costs and usually hindered by a tight schedule. It is consequently necessary to assess intake sedimentation under different natural forcing and plant operation scenarios in order to optimize dredging operations to help mitigate the potentially adverse impact of the waves and tidal currents and meteorological forcing combined with plant functioning.

Site characteristics

The study site is located on the eastern English Channel coast in northern France. Tide in the study zone is classified as mega-tidal and is dominated by semi-diurnal circulation, with low-tide water depth of $10-15$ m, and a mean tidal range of approximately 8.5 m, reaching 10 m during the spring tide [56]. Hydrodynamics are influenced by asymmetrical current velocities, with flood and ebb currents in the E-NE and W-SW directions, respectively. Current velocity at 2.2 m above seabed vary from 0.70 to 0.98 m/s, depending on flood/ebb phase, respectively [68]. Wave activity in this open exposed environment is moderate, with significant annual and decennial wave height of 3.8 m and 4.7 m, respectively, with maximum values of $4.2-5.8$ m, averaged period of $7-9$ s and a predominant W direction. Orbital velocities measured during the spring-tide period ranges between 0.5-1.3 m/s. An example of tidal levels, wind direction and velocity and wave height and direction in January 2016 is shown in Figure 4 . In the study area, bed sediment varies from medium to fine silted sands, with a morphology characterized by the presence of mega-rides parallel to the coast. In this zone, rock occupies less than 4% of bed surface [15].

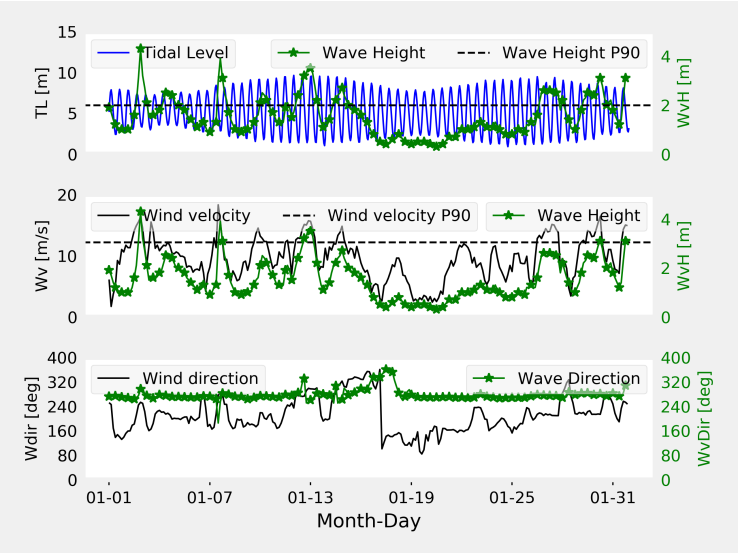


FIGURE 4: Measurements of Tidal Level (TL), Wind velocity (Wv), Wind direction ($Wdir$), Wave Height (WvH) and Wave Direction (WvD) on January 2016. (P: Percentile).

Data

Hydrodynamic and meteorological information comprise wave height, period and direction and wind velocity and direction, provided by the VAG prediction model of the sea state [31], using retrospective 3-hourly simulations between 2009 and 2018. Tidal water levels were obtained from the SHOM-REFMAR tide gauge station located in the vicinity of the study zone, with hourly survey frequency [85].

Bathymetric measurements were available from Single-Beam Echo Sounding on 39 cross-sectional profiles of intake measured at 25 m intervals, collected fortnightly between 2005 and 2018. Mean profiles were 100 m long with 0.5 m spatial resolution of bathymetric data. Additional information such as the daily coolant flow rates, and channel dredging volumes and frequency, were provided by the plant operator. A summary of the data is shown in Appendix A.

The available measurements of the forcings did not have the same frequencies. One solution to homogenize frequencies consists in reducing the measured data to representative statistics over the sedimentation interval $\Delta t \approx 15$ *days* separating two bed elevations measurements. Hence, the following statistics were used:

- *Tidal level indicators:* average low tide ($TLmean$), minimum low tide ($TLmin$), maximum tidal range ($TLrange$) and standard deviation ($TLstd$);
- *Wind indicators:* average wind velocity ($Wmean$) and average direction weighted by velocity ($Wdir$);
- *Wave indicators:* average wave height (WvH), standard deviation ($Wvstd$), average wave period and average wave direction weighted by height (*resp.* $Wvper$ and $Wvdir$), average wave height exceeding the 90^{th} percentile (arbitrary storm indicator, $Wv2m$) and percentage occurence ($Wv2m\%$);
- *Operational indicators:* average pumping flowrate (Qmean); time lapse since last dredging (Dp), and last dredged volume (Dv).

These statistical indicators were calculated for each sedimentation interval; a resulting scatter plot is shown in Figure 23. There was, for example, a positive correlation between mean low tide $TLmean$ over the studied sedimentation periods and wave parameters $Wvper$ and WvH. Mean wave periods $Wvper$ and mean wave heights WvH were also positively correlated.

For the learning part, the data overlapped only over a limited period. A maximum of 60 measurements could therefore be used, with up to 15 forcing variables. Obviously, this "small data" configuration is a considerable handicap for the dimension of the problem, especially given that the variable of interest is a two-dimensional bathymetric field. However, permanent intake monitoring ensures that the data set will always grow and can be used to update the learning. This limitation shall not prevent testing the accuracy of the methodology on small sets

such as are often encountered in physical applications, as attempted below, where learning and prediction using POD-PCE is applied to the described data. For the learning algorithm, input variables are needed, corresponding to the reduced statistical indicators described above, and denoted $(\theta_1, ..., \theta_V)$, where V is the supposed dimension of the problem.

3.2 Measurement-based learning of a physical field using POD and PCE

This section concerns learning the spatio-temporal bathymetric field using POD and PCE independently. The modes are extracted in Subsection 3.2.1 using POD and the temporal patterns are learned as a function of the forcing parameters using PCE in Subsection 3.2.2. Throughout this investigation, particular attention is given to the convergence of the learning and to its robustness with respect to the numerical choices. Trusted POD-PCE learning is immediately used for physical interpretation and the most important physical insights resulting from it are summarized in Subsection 3.2.3.

3.2.1 Physical analysis and data reduction using POD

First, POD was applied on the bathymetry measurements. The aim was to identify morphodynamic patterns so as to better understand the sediment deposition inside the channel, and to characterize variations in depositions with the external forcing variables. After setting aside poor-quality measurements (e.g. incomplete bathymetries), a total $n = 156$ realizations were used. The bathymetry points were sonar boat measurements on $m_p = 39$ cross-sections inside the intake. Linear interpolation was performed on $m_i = 100$ fixed points for each profile, in order to express all measurements on the same grid, giving a total $m = m_i \times m_p = 3,900$ spatial points. The interpolated realizations were then stored in a snapshot matrix $\boldsymbol{Z}(\boldsymbol{\mathcal{X}}, \mathcal{T}) = [z(x_i, t_j)]_{i,j} \in \mathbb{R}^{m \times n}$ and POD-processed as explained in Section 2.1. The EVR defined in Equation 2 and the mean relative RMSE between the POD approximation and the complete measurement (averaged over the realization set as in Equation 18) were calculated for each POD approximation rank and are plotted in Figure 5.

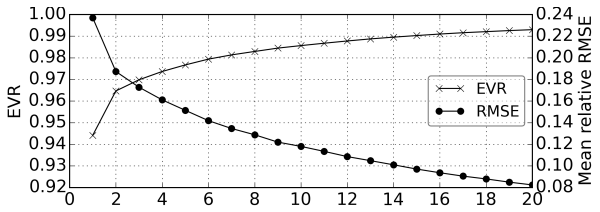
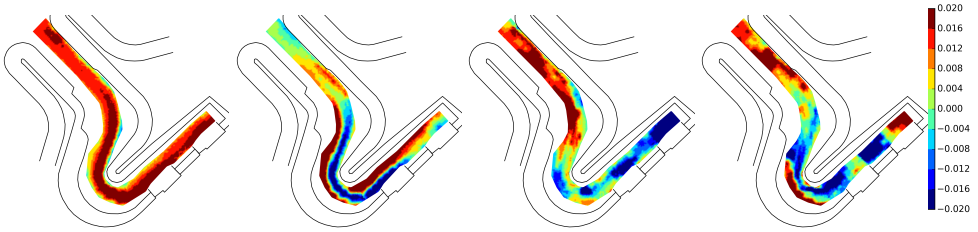


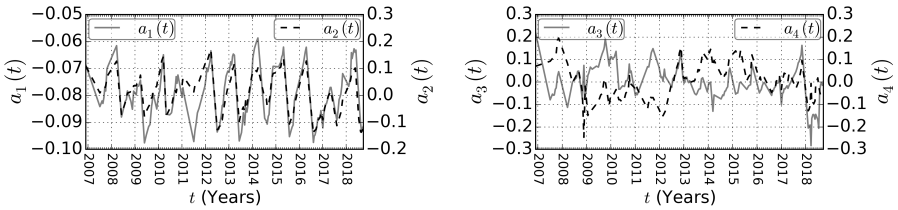
FIGURE 5: Evolution of the EVR and mean relative RMSE with mode number for the POD applied to the intake bathymetries.

The first pattern represents over 94% of the variance, and explains most of the variation in dynamics. The variance percentage reached 99% at rank 14, where the mean error was slightly over 10%, decreasing to 8% at rank 20. Dimensionality reduction is therefore a realistic option for this specific dynamic problem. This encouraged the learning and prediction attempts undertaken in Subsections 3.2.2 and 3.3 respectively. The spatial and temporal components of the first four POD modes corresponding to an EVR higher than 97% are respectively plotted in Figures 6 and 7.



(a) Mode $\Phi_1(x)$ (b) Mode $\Phi_2(x)$ (c) Mode $\Phi_3(x)$ (d) Mode $\Phi_4(x)$ (e)

FIGURE 6: The first four spatial patterns of the POD applied to intake bathymetries.



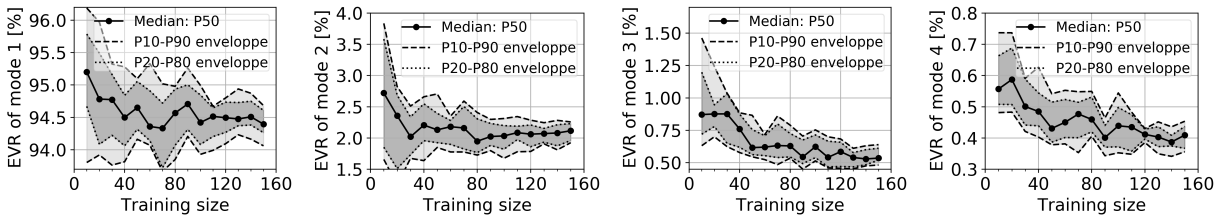
(a) Coefficients $a_1(t)$ and $a_2(t)$ (b) Coefficients $a_3(t)$ and $a_3(t)$

FIGURE 7: The first four temporal coefficients of the POD applied to intake bathymetries.

The first spatial pattern (Figure 6-a) represents the channel's slope. Its temporal coefficient (Figure 7-a) shows regularity in time that is almost periodicity. When it increased, overall sediment deposition in the channel increased, because the difference between the upstream and the downstream bed elevations, and therefore the slope, diminished. The sediment deposition in the channel might be related to the increasing sediment supply caused by the external forcing influence. Decrease always corresponded to a dredging episode. The apparent periodicity is therefore not natural or seasonal but due to periodicity of operational intervention: sediment deposition in the channel is tolerated up to a certain level and then dredging is always undertaken at a certain point, which corresponds to the maximum of the temporal coefficient. The second pattern (Figure 6-b) acts as a geometric distribution function of the sediment deposition. In general, when the first temporal coefficient was maximal, the second coefficient was positive, meaning that the sedimentation mainly occured in the middle of the first portion of the channel (upstream), on the right bank of the bend and on the left bank in front of the pumps. This spatial distribution can be associated to the internal flow characterized by a velocity distribution inside the channel. In fact, the sediments settle where velocity is the lowest, which is probably the case where the banks appear. The computed sediment deposition and erosional patterns are analogous to those commonly observed in meandering rivers [42].

The third pattern (Figure 6-c) shows sediment deposition concentrated in the first portion of the intake, and the fourth pattern (Figure 6-c) emphasizes sediment dynamics, particularly in the downstream part of the channel. This behavior is statistical proof and quantification of finer sediment supply. The finer sediment fraction was transported in suspension and deposited at the end of the intake channel. The temporal coefficients associated with the third and fourth mode (Figure 7-b) were less regular than those of the first and second mode, and seemed to follow a more stochastic dynamic. The peaks may represent unusual sediment supply, probably linked to extreme events (extreme tides, storms, etc.).

To check the robustness of the statistical conclusions deduced from POD, convergence analysis is necessary. This was performed on the EVR values associated with the first four patterns, using bootstrap analysis [20]. The results are shown in Figure 8. The convergence of the mean values and the tightening of the confidence intervals around the mean whith increasing matrix size are clear for these first four modes. However, whereas the confidence intervals represent at most an error of $\pm 0.6\%$ around the mean for the first mode, they reached respectively $\pm 12.5\%$, $\pm 25\%$ and $\pm 12.5\%$ for the second, third and fourth modes.



(a) Mode 1 (b) Mode 2 (c) Mode 3 (d) Mode 4

FIGURE 8: EVR convergence of the first four bathymetry POD modes, using a bootstrap of size 20. Plots show median values and confidence intervals (P: Percentile).

The analysis proved that the POD results could be used to pursue the learning. Firstly, a high EVR and low

RMSE were associated with a small number of modes, guaranteeing optimal data reduction ($d << min(m,n)$ as explained in Section 2.1). The number of POD modes to accurately represent the bathymetry can be chosen accordingly. In the present study, the configuration was $d = 11$ modes (discussed in *Step 3*), guaranteeing EVR $\geq$ 98% and information loss $\leq$ 12% (mean relative RMSE). Secondly, the EVRs were guaranteed to converge statistically at least for the first four modes, with error of $\pm$0.6% around the mean for the most important mode, representing over 94% of the variance. Thirdly, the deduced patterns were physically coherent. Lastly, more than a decade of evolution was used to extract the POD basis, under variable operational and environmental conditions. As long as the operating conditions of the intake remained unchanged, it can be assumed that a wide range of evolutions has been covered, except for extreme events that rarely occur and that are not specifically treated in this study [26]. Hence, the POD basis can be considered as a physically trustworthy and mathematically complete basis to understand past evolutions and to predict future ones. The learning of temporal coefficients is therefore attempted in Subsection 3.2.2.

3.2.2 Learning of the POD patterns using PCE

The temporal coefficients calculated with the POD in Section 3.2.1 (Figure 7) were learned using PCE (theory in Section 2.2). The aim was to learn the way these coefficients evolve over time, as a function of the forcing parameters presented in Table 2 , with the ultimate objective of field prediction as explained in Section 2.3 and applied in Section 3.3. The present section focuses strictly on the learning phase and the physical analysis of the learned model, highlighting quality of learning (robustness, convergence, etc.).

The investigation of learning is organized in four steps.

- *Step 1 - Sensitivity of learning to inputs and marginals:* different configurations were tested to practically demonstrate the implications of these choices on the accuracy of fit.
- *Step 2 - Convergence and Robustness of fit:* The best model resulting from Step 1 was studied more deeply. Its convergence and robustness with respect to the choice of training members are were analyzed.
- *Step 3 - Physical interpretation of the best learned model:* the best model was chosen, and the most influential forcings were ranked using the *Evolution Weights* (EW) and *Generalized Evolution Weights* (GEW) presented in Sections 2.2 and 2.3 respectively.
- *Step 4 - Robustness of the physical interpretation with respect to the learning-set members:* the physical conclusions of the model were shown to be statistically meaningful.

These steps, in the aboved order, follow the logic of statistical model construction to build a trustworthy prediction algorithm, used in Section 3.3.

Step 1 - Sensitivity of learning to inputs and marginals

Input variable selection is capital, and marginals must be chosen wisely. Below, we demonstrate the influence of these choices on the performance of the learning. Different configurations were tested.

Before introducing the tested configurations, the training steps that are common to all configuration need to be defined. A learning data-set is classically separated into different sub-sets, corresponding to different steps of the learning algorithm. This is commonly referred to as the "Train-Validation-Test Split" [95]: a training set is used for the learning, a validation set is used to check the learning and for further calibration, and a test set is used to assess the prediction capability of the statistical model. However, the data-set used in this study was small: the bathymetry and forcings measurements shown in Subsection 3.1 overlap for the 2012-2017 period only, leaving 64 sedimentation periods to study. Therefore, only a "Train-Predict" split was performed, where the prediction set played the role of both the test set and validation set. Hence, the numerical choices were calibrated on the training set, and validated on the prediction set for both statistical accuracy and physical prediction. The learning was then performed with an arbitrary choice of training-set size at 50, which left a prediction set of 14. The training data were chosen in chronological order (first 50 records), to mimic the learning process in an industrial context. This arbitrary training-set choice had consequences for learning ; the sensitivity of learning to training set choice was investigated (*Step 2*). All the model choices presented below (choice of inputs and marginals) were assessed on this training configuration. To assure that comparison is made between models at their best performances, the PCE polynomial degree was optimized for each separately. Degrees from 1 to 7 were tested, and the associated relative empirical errors on the training and prediction sets, respectively ϵ_T and ϵ_P, were calculated as in Equation 16 . The PCE degree that minimized the training and prediction errors for each model was chosen ; the corresponding

result is referred to as "optimal" learning.

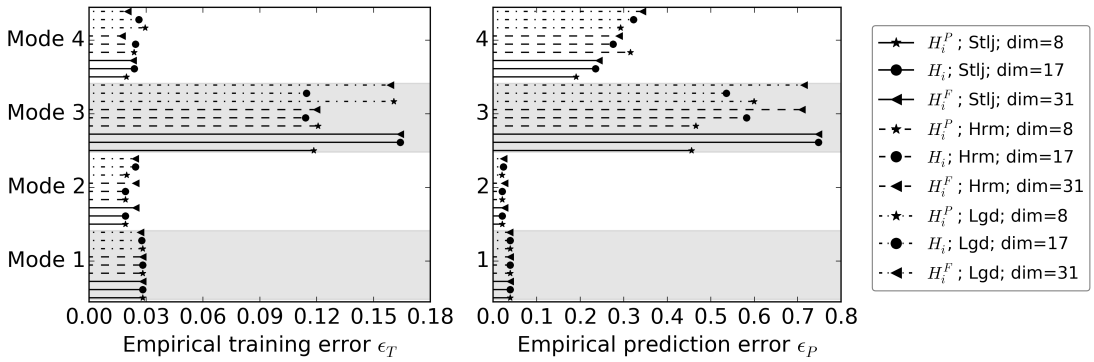
Three different input configurations were used for the learning of each temporal coefficient a_i as generally formulated in Equation 11 .

- $\mathcal{H}_i$-model: a first simple configuration where all the statistical indicators described in Subsection 3.1 were used and an independence hypothesis between the POD temporal coefficients is considered. The model is written as in Equation 13 : $a_i(t_{j+1}) \approx \mathcal{H}_i\left[a_i(t_j), t_{j+1} - t_j, \mathbf{\Theta}(t_j \to t_{j+1})\right]$. This is a model of dimension 17.
- $\mathcal{H}_i^F$-model: a more complex configuration where a "Full" 15-mode POD approximation is considered with possible dependencies between the temporal coefficients. Of course, the choice of the basis size and the dependency structure can be optimized, but the objective here was to make a first step toward a more optimal configuration. The model can be written as: $a_i(t_{j+1}) \approx \mathcal{H}_i^F\left[a_1(t_j), \ldots, a_{15}(t_j), t_{j+1} - t_j, \mathbf{\Theta}(t_j \to t_{j+1})\right]$. This is a model of dimension 31.
- $\mathcal{H}_i^P$-model: a smaller set of inputs, used by the operators to qualitatively evaluate sediment deposition risk, was used. It corresponds to the six variables $TLmean$, WvH, $Wvper$, $Wvdir$, $Wv2m$ and $Wv2m\%$. This mimics the physical expertise that may be engaged when building a statistical model. It is written as $a_i(t_{j+1}) \approx \mathcal{H}_i^P\left[a_i(t_j), t_{j+1} - t_j, \mathbf{\Theta^P}(t_j \to t_{j+1})\right]$, where $\mathbf{\Theta^P}$ stands for the "physical". This is a model of dimension 8.

To these variable choices were associated three choices of marginals, conditioning the choice of the PCE orthonormal polynomial (Section 2.2).

- *Lgd:* all the variables follow a Uniform PDF. The bounds of the marginal were set to the minimum and maximum chronological values ±1% as in [103]. The associated orthonormal polynomial basis is the Legendre family.
- *Hrm:* all the variables have Gaussian marginals characterized by the empirical mean and variance deduced from the data. The associated orthonormal polynomial basis is the Hermite family.
- *Stlj:* the marginals were inferred from the data using Gaussian Kernel density estimates. The orthonormal polynomial basis was constructed from the knowledge of the marginal using a Stieltjes orthogonalization.

The three marginal choices *(Lgd, Hrm, Stlj)* were trained with the three dimension choices ($\mathcal{H}_i^P$ dim=8, $\mathcal{H}_i$ dim=17, $\mathcal{H}_i^F$ dim=31). The empirical errors of the "optimal" learnings are compared in Figure 9 .



(a) Training error (b) Prediction error

FIGURE 9: The empirical training error ϵ_T and prediction error ϵ_P corresponding to the optimal fitting of models with different dimensions and marginals. The figure is organized as follows: errors are plotted for each mode vertically, separated by a gray band. Each marginal type corresponds to the same line style, and each dimension to the same marker style. The legend is shown in the order of the plots, down to top for each mode.

For Modes 1 and 2, training and prediction errors were almost identical for all configurations, although with a slight advantage with the smallest dimensions for all the marginal types in the learning of Mode 2. Starting from Mode 3, bigger differences emerged. At the learning step of Mode 3, models of dimension 17 and 31 were poorly fitted for the *Stlj* and *Lgd* configurations compared to others. At the prediction step of Mode 3, the errors of models with dimension 31 were much greater than smaller dimensions for all marginal choices. There seemed to be an overfitting of the model by selecting a larger number of inputs. The best models for Mode 3 were those of the

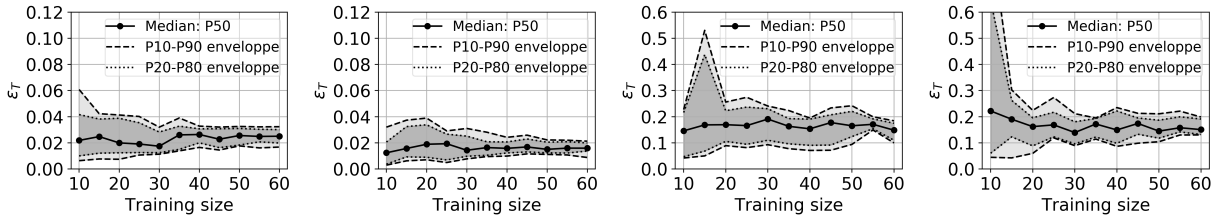
smallest dimension, 8, with either the *Stlj* or the *Hrm* model. Lastly, for Mode 4, two orderings were observed for the prediction error. Firstly, for each marginal choice, prediction error increased with dimension, which confirmed the overfitting hypothesis. Secondly, error was the smallest with the *Stlj* model (Kernel density), followed by the *Hrm* model (Gaussian) and lastly by the *Lgd* model (Uniform). Here, Uniform marginals performed worst ; they were probably too different from the real data marginals and did not account for particularities in the inputs. In the parametric family, Gaussian marginals probably fitted real density better.

To conclude this comparison, the best marginal choice was the Kernel density estimate. The smallest dimensions performed the best, with the *Stlj* choice for the polynomial basis. The $\mathcal{H}_i^P; Stlj$ model of dimension 8 was therefore selected. However, the training was performed with an arbitrary split of the available statistical set. The sensitivity of the model to the learning set size and members is performed in the following step.

Step 2 - Convergence and robustness of the fit

Up to this point, an arbitrary number of 50 measurements was used for the training phase, leaving 14 prediction points for testing purposes. In the following, the influence of training set size on the learning and prediction error is assessed. The objective is to check the robustness of the previous best model $\mathcal{H}_i^P; Stlj$ with respect to the data-set size and members. The evolution of the training and prediction empirical errors according to training set size is shown in Figure 10 and 11 respectively. For each training set size, members were chosen randomly among the full data-set, and the remaining members were used for the prediction phase. For the estimation of the confidence intervals, bootstrap analysis was performed [20]. For comparison, the convergence of the $\mathcal{H}_i^P; Hrm$ model is plotted for Mode 3 and can be found in Appendix B for Modes 1 and 2.

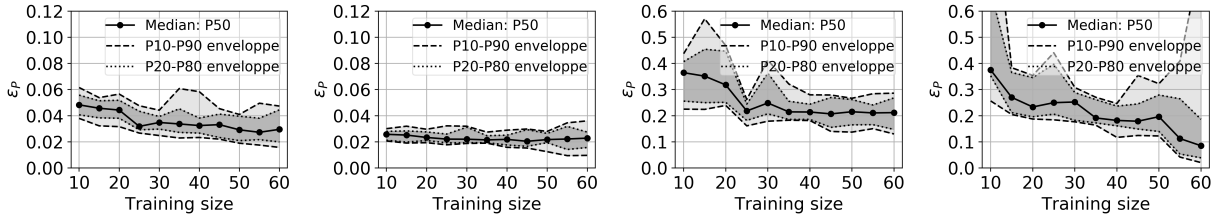
For the first two modes, the training errors in Figure 10 show a convergence of the median value and a tightening of the confidence intervals. The trainings can be considered as converging from around training size 40.



(a) Mode 1 - Stlj (b) Mode 2 - Stlj (c) Mode 3 - Stlj (d) Mode 3 - Hrm

FIGURE 10: Training empirical errors ϵ_T calculated for diverse training sizes with a Bootstrap of size 20. Plots show median value and confidence intervals (P: Percentile).

The associated median prediction errors in Figure 11 globally decreased with increasing training set size. However, although the final median values were lower for the *Hrm* model, the confidence intervals were much larger than for the *Stlj* model. The latter seems much more robust with respect to changes in training scenario.

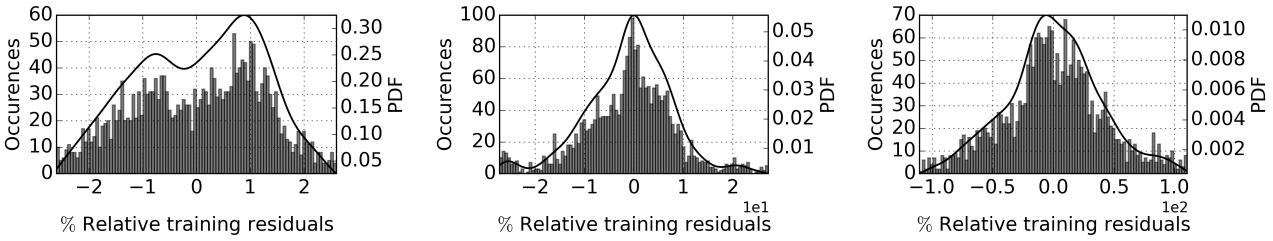


(a) Mode 1 - Stlj (b) Mode 2 - Stlj (c) Mode 3 - Stlj (d) Mode 3 - Hrm

FIGURE 11: The prediction empirical errors ϵ_P calculated for diverse training sizes with a Bootstrap of size 20. Plots show median value and confidence intervals (P: Percentile).

The residuals distributions of the $\mathcal{H}_i^P; Stlj$ model, calculated as $a_i(.) - \mathcal{H}_i^P[.]$ on all the training sizes and Bootstraps, are shown in Figure 12 and Figure 13 for training and prediction, respectively. Only the middle 80% portion

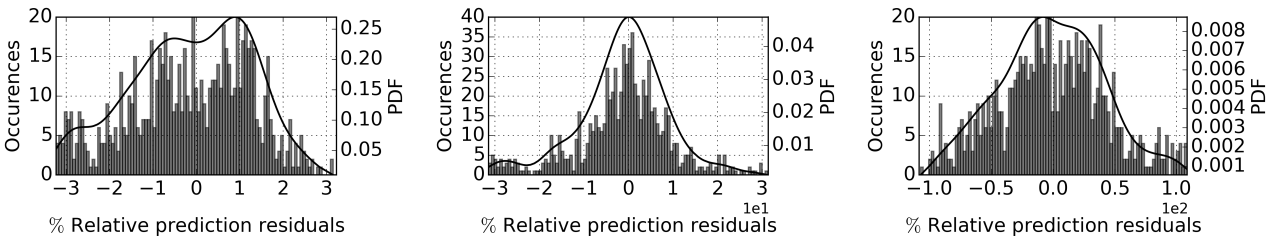
of the residuals range is plotted, in order to analyze the center of the distribution ; the full residuals distribution was long tailed, because the confidence intervals associated with small training set sizes were too large and produced extreme behaviors of the model.



(a) First mode $a_1(t)$ (b) Second mode $a_2(t)$ (c) Third mode $a_3(t)$

FIGURE 12: The training residuals distributions of the $\mathcal{H}_i^P; Stlj$ model calculated for diverse training sizes with a Bootstrap of size 20.

The training residuals were generally centered around zero: i.e., the models are unbiased. A slight asymmetry was, however, observed for Mode 1, which means that $a_1(.)$ was more often overestimated by $\mathcal{H}_i^P; Stlj$. Consequently, the mean elevation in the channel and the mean global sedimentation may be slightly exaggerated. These exaggerations, however, remained within a reasonable range, as most of the residuals fell within the $\pm 2\%$ interval. The training residuals of Modes 2 and 3 were perfectly centered, but percentage error dramatically increased. Most of the residuals fell within the $\pm 10\%$ interval for Mode 2, whereas they reached $\pm 50\%$ for Mode 3. However, this error concerns modes that represent at most 4% of the total bathymetry variance, as more than 96% of the total variance was already captured by the addition of the first two modes.



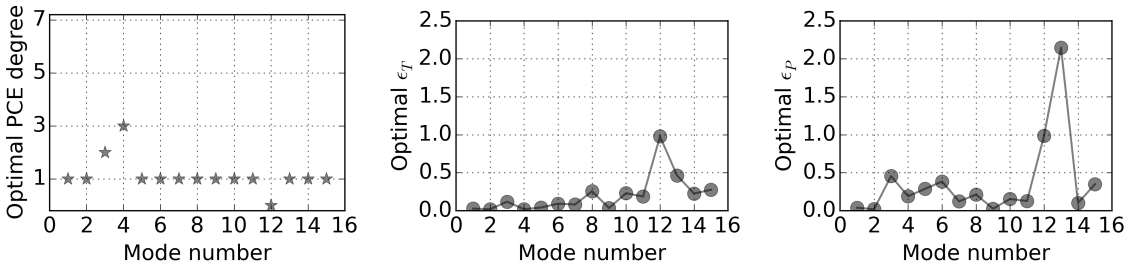
(a) First mode $a_1(t)$ (b) Second mode $a_2(t)$ (c) Third mode $a_3(t)$

FIGURE 13: The prediction residuals distributions using *Stlj* model of dimension 8 calculated for diverse training sizes with a Bootstrap of size 20.

The residuals shapes (i.e. slight overestimation for Mode 1 and perfect centering for Modes 2 and 3) were maintained through the prediction phase. Furthermore, the residuals mostly fell within the ranges identified in the training phase. $\mathcal{H}_i^P; Stlj$ model behavior was stable. The prediction uncertainty could therefore be measured and trusted and the physical interpretation was consequently robust, as discussed below in *Step 3*.

Step 3 - Physical interpretation of the best learned model

The calibrated $\mathcal{H}_i^P; Stlj$ PCE models were considered optimal, as they showed good fit, convergence and robustness with respect to the training choices. Here, they are analyzed to deduce physical information. Firstly, the optimal polynomial degrees selected for each mode and the associated training and prediction empirical errors are shown in Figure 14. Linear models were optimal for Modes 1 and 2 (degree 1), and the associated errors were low for both the training and the prediction sets.



(a) PCE degrees (b) Training errors (c) Prediction errors

FIGURE 14: Optimal PCE degrees for the $\mathcal{H}_i^P; Stlj$ model and associated empirical errors of the training (ϵ_T) and the prediction (ϵ_P) sets.

For Modes 3 and 4, the optimal polynomial degrees increased, which implies higher-order contributions and/or higher-order interactions for the input variables. For modes of higher ranks, the models were either linear (degree 1) or approximated by a simple average value (degree 0). This means that LARS rejects polynomial terms of higher degrees because they do not significantly improve the learning [6]. Prediction relative empirical errors increased from Mode 3, but remained under 50% up to mode 11. This percentage error seems surprisingly high, but must be interpreted according to the meaning of the relative empirical error, which, as calculated in Equation 16, is a measure of the missing variations (distance between the model and reality) relative to the variance of the data. It is also often called "the fraction of unexplained variance" [17], and represents the amount of variance that was not captured by the PCE model. For example, for mode 12, estimated with a degree 0 PCE model, 100% of the variance is not captured, which is natural because only the average value is accounted for with degree 0. For Mode 13, LARS did not include polynomial terms of degree higher than 1 (no improvement of the learning), but this appears to be a bad choice since the prediction error was much higher than the training error, meaning that the training set or inputs were not enough to guarantee accurate learning in Mode 13. For Modes 3 to 11 with error up 50%, this means that either the training set or the used inputs made it impossible to predict more than 50% of the variance. However, as presented in *Step 2*, this 50% error concerned at most 4% of the total bathymetry variance. Hence, the errors starting from Mode 3 represented at most 2% of missing variance. Beyond Mode 11, prediction with PE would not be optimal, as the prediction error dramatically increases.

Secondly, the explicit formulations of the PCE models (reported in Appendix B) were used to analyze the contribution of each forcing variable to the dynamics. for this, the *Garson Weights* (GW) defined in Equation 7 were used to estimate the influence of the forcings on each temporal coefficient. The global influence on the whole bathymetry field was quantified using the *Generalized Garson Weights* (GGW), as in Equation 20.

The ranking of the modes and the impact of the inputs is represented in Figure 15. The inner circle represents the contribution of each mode to the whole field. Mode 1 corresponds to a major contribution and the following modes are ranked in terms of contribution, which is the essence of POD. The outer circle represents the contribution of each polynomial term in the PCE representation. The share of each polynomial term corresponds to the GGW in relation to the global contribution (full circle). When this share is compared to the importance of the corresponding mode, it corresponds to the GW. Lastly, the polynomial terms corresponding to more than 0.5% GGW are indicated. $\zeta_{\boldsymbol{\alpha}=(.)}(.)$ corresponds to the notation introduced in Subsection 2.1, with the multi-index notation for α that represents the polynomial degree of each monomial. For example, $\zeta_{\boldsymbol{\alpha}=(\alpha_1,\alpha_2)}(\theta_1,\theta_2)$ corresponds to a polynomial of degree $\alpha_1+\alpha_2$, where θ_1 contributes as a monomial of degree α_1 and θ_2 as a monomial of degree α_2. The meaning of the variables that appear in Figure 15 can be found in Subsection 3.1

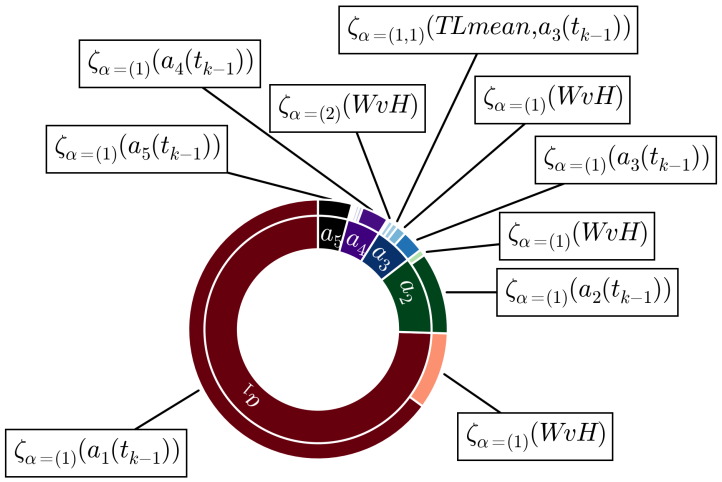


FIGURE 15: Piechart of the most influential parameters, using GW and GGW on the POD-PCE. The inner circle represents the share of each mode. The outer circle represents the share of each polynomial term. The polynomial terms corresponding to GGW higher than 0.5% are shown.

For all the temporal coefficients $a_i(.)$, the most influential contributor by far was the value of the previous state $a_i(t_{j-1})$, in the form of a monomial of degree 1. It is followed by contributions involving the mean wave height during the sedimentation period period WvH for all the modes, which makes WvH the most important external forcing, figuring in the third position among all the forcings, with a contribution of 9.6% through the first mode, a total of 12.6% if only WvH monomials are considered, and 13.3% if interactions with other variables are taken into account.

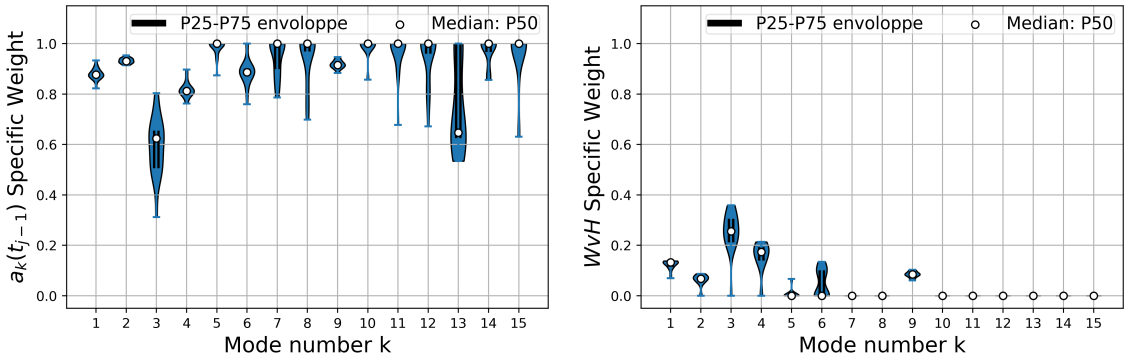
The other forcing contributions also appeared, but with much less importance: e.g., the influence of mean low tide level $TLmean$, which took an interaction form with the previous bathymetry shape for Mode 3. Firstly, this interaction makes sense in terms of physics, as sediment deposition is conditioned by the value of bed shear stress [105], which depends on velocity and water depth. The water depth value is exactly the tidal level minus the bed elevation value, which here appears as a multiplicative interaction between $TLmean$ and Mode 3. Second, the value of this contribution was only 0.7% GGW, which is negligible when compared to the first contribution of WvH. The learned model gave much more importance to waves than to tides. This does not necessarily mean that tides have no influence on sediment deposition, but may simply suggest that, in the present configuration, sediment mobilization by the tide is always more or less the same, and that the forcing that makes a considerable difference is the variation in wave heights. Waves are a determining factor for sediment mobilization in coastal configurations, through the influence they have on bed shear stress [105]. Further more, a noticeable correlation between WvH and $TLmean$ can be seen in Figure 23 in Appendix A, which means that the information of low-tide levels is to a certain extent contained in the mean wave height. There is therefore a probable dependency between these variables. In case of dependencies, the iterative process used by LARS may drop a variable that is physically important because the important information is already contained in another variable, due to their dependency.

Lastly, it is important to note that the contribution of less frequent wave events was also present but to a much smaller extent. It is represented by a polynomial term in the form $\zeta_{\alpha=(1,1)(Wv2m,Wv2m\%)}$, where $Wv2m$ and $Wv2m\%$ are respectively mean wave height exceeding $2m$ and the associated frequency of occurrence (arbitrary storm indicator chosen in Subsection 3.1). This term appears in Modes 3 and 4 for a maximum total influence of 0.3%. Higher-order interactions and less frequent events are therefore represented by modes of higher rank, associated with smaller variance percentages.

Step 4 - Robustness of the physical interpretation with respect to the learning-set members
As a last proof of the robustness of the proposed learning algorithm, specifically concerning physical interpretation, a sensitivity analysis with respect to the training set members was performed. The robustness of the calculated *Garson Weights* (GW) with respect to the choice the training members was studied. This is equivalent to studying the robustness of the polynomial basis term selection as produced by LARS, and their associated multiplicative

coefficients.

For this, a Bootstrap analysis was again used to construct different learning sets of size 50, instead of choosing the first 50 measurements. This produces a distribution of the GWs rather than a single value, for each polynomial term. The result is shown in Figure 16 for the weights of the $a_k(t_{j-1})$ and WvH monomials.



(a) Polynomial term $a_k(t_{j-1})$

(b) Polynomial term WvH

FIGURE 16: Probability density functions of the GWs associated with the degree 1 monomials of variables $a_k(t_{j-1})$ and WvH. The training size is 50 using 20 different random picks for each size.

For modes 1 to 4, the median weights P50 (Percentile 50) of the $a_k(t_{j-1})$ monomials, represented in Figure 16-a, were always over 0.6, but the variation range was strictly less than 1, with density functions centered around the median and a small standard deviation for modes 1, 2 and 4. This means that whatever the training set, the previous state $a_k(t_{j-1})$ value was always predominant but never enough to estimate the evolution of the first four modes. A tendency (in particular linear) using the last state was not sufficient, and additional information was always needed (forcing). In parallel, Figure 16-b shows that this information is certainly the waves, as the median values of the GW for the first four modes were between 10 and 25%, corresponding to the information gap left by the previous value variable $a_k(t_{j-1})$ in the fitted PCE model.

Starting from mode 5, the median values of $a_k(t_{j-1})$'s GW had greater chance of falling around 1, which means that the associated polynomial models only rely on the last recorded value of the mode for the future guess. In other terms, the constructed model consists of a linearization around the previous value (tendency capturing) and does not incorporate the correlations between the future-state and the forcing variables (causality model). This can be explained by the small variances of the higher-rank modes and the difficulty of learning the PCE models from statistics averaged over the sedimentation periods. Additionally, the P25-P75 confidence interval moved to the upper bound of the density functions.

3.2.3 Summary of the physical insights from the learning

The spatial patterns as deduced by POD express the spatial correlation in the sediment deposition from the upstream to the downstream part of the channel. The EVR reached 99% with $d = 20$ modes only, where the mean relative RMSE between the approximation and reality was slightly over 10%. This is a statistical proof that the spatial correlations expressed in the POD patterns are explanatory of the physical dynamics over their whole range of variation (at least that observed from 2010 to 2018), with a low approximation rank. In conclusion, the dynamic problem exhibits fairly strong spatial correlations, and the solution to the problem can be expressed on a finite orthonormal basis.

The temporal patterns express the evolution of the sedimentation, as they multiply the spatial patterns. They were learned using PCE as a function of the previously cited inputs (previous states and forcings). The statistical model configuration (dimension and marginals) was chosen after an investigation of different options. The associated training and prediction error converged for the first three modes, and are characterized by tight confidence intervals. The residuals of the selected model were either negligible or centered around zero, demonstrating the

unbiased character of the learning and prediction. The fitted models are of lower degree for the low-rank modes 1 and 2 and of higher degrees for modes 3 and 4, which are higher-rank, due to the emergence of interactions between the forcings, namely variables related to extreme behavior (storm events). The model mainly relies on the last state information, showing a strong correlation/continuity in time of the studied physics. Using GW, which measures the forcing influence for the first five modes, the action of waves was highlighted by the PCE model as a determining phenomenon. The first mode influenced the dynamic with a rate of 64.9%, the previous value of the second mode with a rate of 10.2% and, in third position, the mean wave height with a rate of 9.6%. The remaining 15.3% is essentially associated with previous values of higher order modes (10.3%), interactions with tides and contributions of other wave indicators. The GWs show robustness with respect to the choice of the training set members, which makes them trustworthy, at least for temporal correlation and analysis of wave influence. The main physical conclusions are that the dynamic problem is characterized by strong temporal correlations, representing more than 85% of the evolution, with an external sediment source, mainly represented by the waves, representing not more than 15%.

3.3 Prediction of a physical field using POD-PCE coupling

After performing both POD and PCE independently, the accuracy of a Machine Learning process using a POD-PCE coupling was assessed as in Section 2.3. In the continuity with Section 3.2, the first 50 historical bathymetries were used for training and the other 14 for forecasting. First, the impact of the size of the POD basis on the prediction process is assessed in Subsection 3.3.1. Then, the best size was determined and the average prediction behavior is analyzed in Subsection 3.3.2. The accuracy of the POD-PCE ML in predicting spatial details is assessed on cross-section examples in Subsection 3.3.3 and a summary is given in Subsection 3.3.4.

3.3.1 Influence of POD basis size

In order to track the errors generated by the various steps of the algorithm (POD, PCE and coupling), the mean relative RMSE (averaged over the prediction set, as in Equation 18) was calculated for each step and for each approximation rank d, as follows:

- *Reduction error:* distance between the POD approximation $\sum_{k=1}^{d} a_k(t)\mathbf{\Phi}_k(x)$ and the corresponding measured bathymetry field $\mathbf{z}(\mathbf{x}, t)$;
- *Learning error*: distance between the POD approximation $\sum_{k=1}^{d} a_k(t)\mathbf{\Phi}_k(x)$ and the prediction using the POD-PCE coupling formulated as $\sum_{k=1}^{d} \mathcal{H}_i^P \left[a_i(t_j), t_{j+1} - t_j, \mathbf{\Theta}^{\boldsymbol{P}}(t_j \to t_{j+1})\right] \mathbf{\Phi}_k(x)$;
- *Prediction error:* the resulting final error between the prediction using POD-PCE coupling and the corresponding measured bathymetry $\mathbf{z}(\mathbf{x}, t)$.

The results are shown in Figure 17 . Reduction error decreased from 16% to 3%, with increasing approximation rank. The error followed a logarithmic trend, with a significant slowdown from rank 9. These errors are coherent with the errors averaged over the full set (rather than the prediction set only) in the POD results Section 3.2.1 (around 8%).

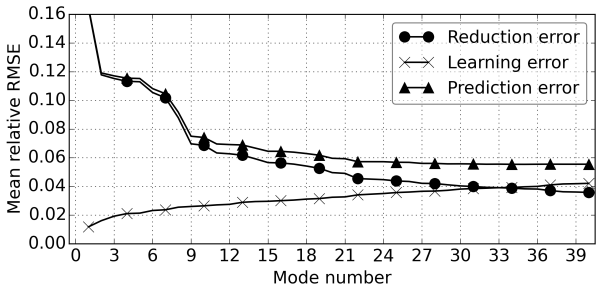


FIGURE 17: Mean relative RMSE generated by the reduction and the learning, and the resulting prediction errors for different approximation ranks.

The learning error increased from 1% to 5% with increasing approximation rank, which is natural because the complexity of the model is increased. In fact, a prediction of rank $d + 1$ has an additional temporal coefficient

that is predicted as compared to rank d. It is therefore natural that the distance between the approximation $Z(x,t) \approx \sum_{i=1}^{d} a_i(t) * \phi(x)$ and its prediction $Z(x,t) \approx \sum_{i=1}^{d} \mathcal{H}_i^P * \phi(x)$ increased with increasing rank. The learning error order of magnitude was consistent with the empirical prediction error of 4% for mode 1 (as calculated in Section 3.2.2), associated with an EVR of over 94%. Lastly, the prediction error decrease is the balance of, on the one hand, the increase in accuracy by adding POD modes and, on the other hand, the increase in forecasting error with increasing number of temporal coefficients to be predicted. Consequently, the prediction error decreased from 16 to 6% up to rank 11, following almost the same decreasing trend as the reduction error. However, the decrease rate became slower and increasingly subdued, being overtaken by the learning errors, which dramatically increased starting from mode 12, as seen in Figure 14. Hence, a POD-PCE model of size 11 was selected for prediction.

3.3.2 Average performance of the chosen model

Average sediment deposition was predicted using the POD-PCE model of rank 11, for each of the 14 prediction dates. The average sedimentation rate, denoted S_r, was calculated for time t_j representing the sedimentation over $[t_{j-1}, t_j]$, as in Equation 21. For operational estimation of sediment deposition, only the positive evolutions are of interest ; therefore, the erosion points were discarded in calculating rate S_r by cancelling negative evolutions. Indeed, $z(\mathbf{x}_i, t_j) < z(\mathbf{x}_i, t_{j-1})$ implies $(z(\mathbf{x}_i, t_j) - z(\mathbf{x}_i, t_{j-1})) = -|z(\mathbf{x}_i, t_j) - z(\mathbf{x}_i, t_{j-1})|$, and therefore a null contribution to the sedimentation rate S_r. Furthermore, only regions of considerable depth are of interest. Therefore only n_p bathymetry points under -1 m $(\mathbf{x}_i, i \in \mathcal{N}_p)$ were taken into account. The results are shown in Figure 18

$$S_r = \frac{1}{2n_P} \sum_{i \in \mathcal{N}_p} \frac{(z(\mathbf{x}_i, t_j) - z(\mathbf{x}_i, t_{j-1})) + |z(\mathbf{x}_i, t_j) - z(\mathbf{x}_i, t_{j-1})|}{|z(\mathbf{x}_i, t_j)|} \quad . \tag{21}$$

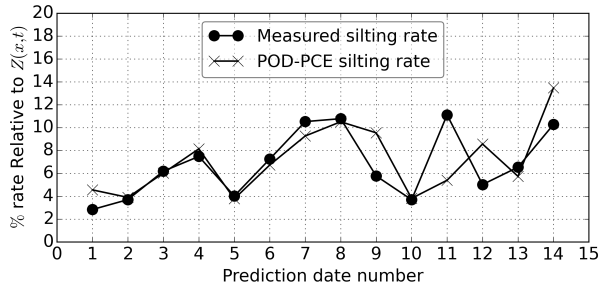


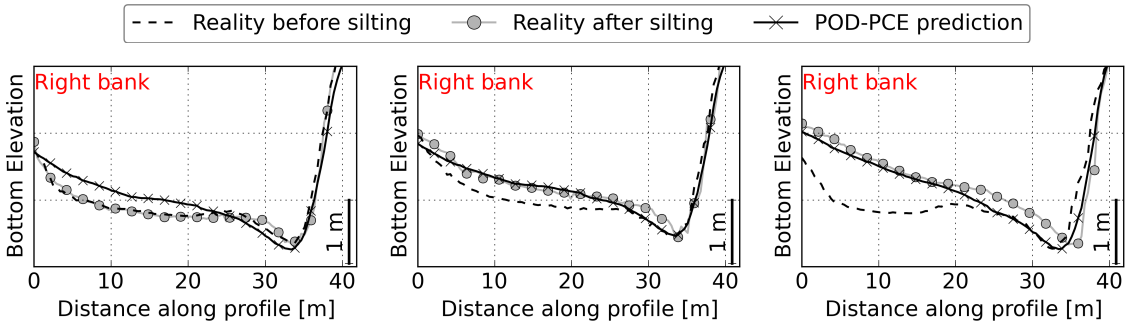
FIGURE 18: A comparison between the real sedimentation rates and the POD-PCE prediction of the sedimentation.

The POD-PCE prediction globally followed the real sedimentation trend, for example from Dates 1 to 8. When it was not equal to the real sedimentation rate, it was generally an overestimation, which is coherent with the asymmetry observed in the distribution of Mode 1 training and prediction residuals (Figure 12 in Subsection 3.2.2).

For the particular Date 11 however, half of the sedimentation was missing. Investigation of the data for this particular measurement showed that the previous record, taken as input, had been made 29 days previously, which is far from the average $\Delta t \approx 15$ days ; it is twice the mean interval, thus underestimating sedimentation by half. As measurement intervals were in general around the average, sedimentation time interval Δt was not selected as a key parameter by LARS for the dynamic model, although it was given as an input and is physically significant. For the particular case of the time variable, multiplicative enhancement can be intended as a correction. However, this shows one of the limitations of statistical modeling: statistical significance can be confused with physical importance. Indeed, for the statistical conclusions to be physically significant, the measurements should be diverse enough to account for the variations in the inputs and the impact of these variations on the output. This was unfortunately not guaranteed for sedimentation measurement intervals, as they were often equal to 2 weeks. Additionally, for Dates 12 and 13, a large part of the wave measurements were missing in the sedimentation time interval. Consequently, mean wave height WvH was estimated over only a small portion of the time interval. This may lead to a good prediction (Date 13) if the interval used is representative enough of the full interval, and to bad prediction (Date 12) when not, and highlights the limitations of statistical averaging.

3.3.3 Spatial details of prediction by the chosen model

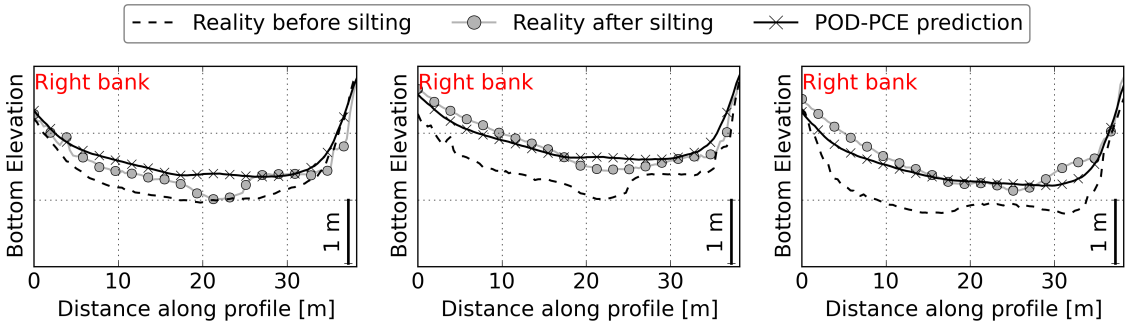
The spatial details of the prediction were analyzed on cross-sections for specific prediction dates. First, sediment deposition was observed on a cross-section at the entrance of the intake (Figure 19). In accordance with the previous conclusions from Figure 18, the POD-PCE prediction captured various sedimentation ranges, as shown with Dates 2 and 7. However, a slight artificial sedimentation was predicted whereas there was no dynamics in reality, for example for Date 1 (Figure 19-a), due to the fact that the model is continuous, whereas threshold phenomena can occur in reality.



(b) Date 1 (c) Date 3 (d) Date 7

FIGURE 19: Assessment of the POD-PCE prediction algorithm on a cross-section at the entrance of the intake.

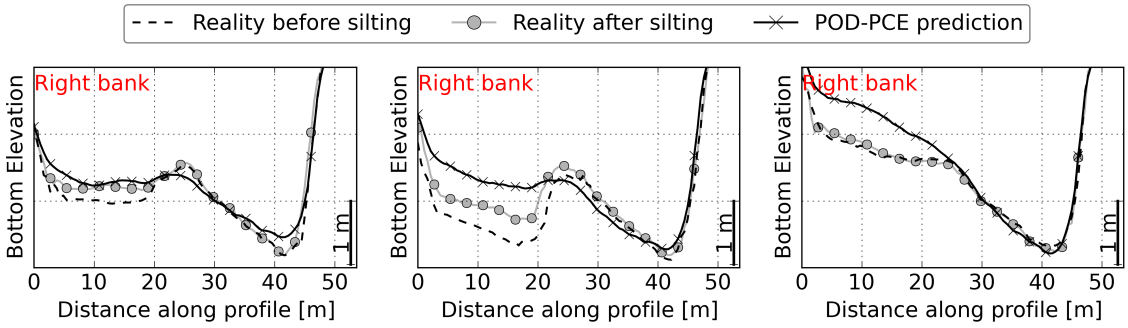
Next, sediment deposition was observed on a cross-section at the middle of the first portion of the intake (Figure 20). Mean sedimentation was well captured, but some details of the bathymetry were missing, such as formation of a new feature for Date 5 (distance 20 to 25 m) and Date 7 (distance 30 to 35 m). For Date 6, sediment deposition was slightly underestimated in the right bank and overestimated in the left bank. However, although the details of sediment deposition were not perfectly captured, the value of the sedimentation area corresponds well enough to reality. It can also be concluded that the way the RMSE and relative errors are averaged in space, for example in Figure 17, actually penalizes the accuracy of the algorithm because it does not take account of the oscillation of the prediction around an accurate mean.



(b) Date 5 (c) Date 6 (d) Date 7

FIGURE 20: Assessment of the POD-PCE prediction algorithm on a cross-section at middle of the first portion of the intake.

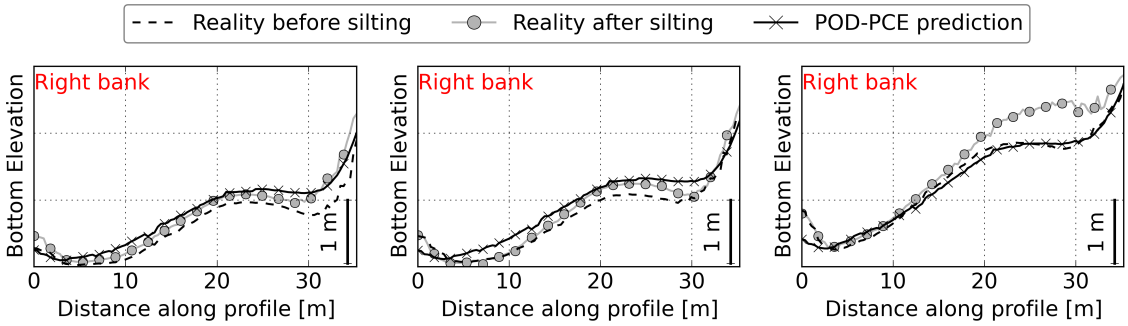
Then sediment deposition was observed on a cross-section at the bending part of the intake (Figure 21). It shows that the prediction algorithm understood that the sediment deposition mainly occured in the right bank of the channel, for example for Date 7, even though it was overestimated. Furthermore, considering modes of higher rank from the previous measurement as an input, the algorithm captured the swing in the profile throughout its history, which is here observed from Date 4 to 14.



(b) Date 4 (c) Date 7 (d) Date 14

FIGURE 21: Assessment of the POD-PCE prediction algorithm on a cross-section at the bending portion of the intake.

Lastly, a cross-section of the last portion of the channel, in front of the downstream pumping station, is shown in Figure 22. Once again, the prediction algorithm understood where the sediment deposition occurs, this time in the left bank of the channel, coherent with the pattern represented by Mode 2 (Figure 6-b). However, it can be seen that unusual sediment deposition occurred for Date 11, which was not captured by the model, and may correspond to the arrival of less frequent fine sediment that appears in Mode 4 (Figure 6-d). This also explains the sediment deposition error observed for Date 11 in Figure 18.



(b) Date 4 (c) Date 5 (d) Date 11

FIGURE 22: Assessment of the POD-PCE prediction algorithm on a cross-section at the last portion of the intake.

3.3.4 Summary of POD-PCE algorithm performance

Overall, the proposed learning algorithm showed interesting prediction characteristics. Firstly, model complexity can be increased gradually, by increasing the number of POD modes when accurate. Plotting error against the number of modes shows a convergence that helps in selecting the optimal number of modes. The average RMSE of the predicted field remains reasonably low. It was around 6% with the 11 modes selected in the present case.

Secondly, trends are well captured for spatially averaged quantities (here sedimentation rate) and for detailed spatial representations of the field. Good spatial distribution of evolution is guaranteed by the POD modes, even when evolution amplitude is over- or under-estimated.

Lastly, some disadvantages should be noted. For example, less frequent events that are represented by modes of higher ranks can be overlooked. Furthermore, sudden changes in features were not sufficiently captured, due to the high temporal correlation between last state and future state that was incorporated in the learning.

4 Summary and discussion

In this study, POD-PCE coupling for field-measurement based Machine Learning was proposed and assessed, in an industrial context in the field of geosciences, for complex physical phenomena involving non-linear dynamics

and various forcings.

POD showed excellent performance both for reducing problem dimensionality and for physical interpretation of spatial and temporal modes. This is an important property of POD [47, 48]. Even though the spatial patterns are mathematical modes that do not have explicit physical meaning, an interpretation could be given, at least for the patterns that were statistically stable, because they are statistical representations of the dynamic behavior. Our investigation showed that temporal signals are also relevant: they can show more or less regularity, which can be related either to the representation of different space or time scale physics, or to less frequent events. In our particular problem, the mean of the bathymetry field was not deduced from the field before applying POD, as can be the case for classical applications [14]. It was used as valuable information for decomposition, which allowed correlations to be studied between the first and second patterns, as well as the construction of the learning. The potential of POD for detecting biased and missing data was also assessed. POD was first applied to the whole set of measurements, but discontinuities emerged in the temporal signals of the decomposition. Such a procedure is important because, in most of cases, the data need to be filtered, which is a time-consuming task. The POD enabled fast recognition of elements that react differently from the average. However, many points of improvement are worth mentioning. Firstly, the choice of POD as a decomposition technique was here motivated by its simplicity and the possibility of interpretation when coupled to a linear learning formulation such as PCE. Other decomposition techniques exist, and many authors attempted comparisons, for example with Fourier [84], extensions of POD [37, 30] or other classes of decomposition [90]. For the present application, other decomposition techniques such as Kernel Principal Component Analysis (KPCA) [69] and Sparse PCA [43] were analyzed, without significant improvements. Secondly, data filtering using POD consisted only in deleting the poor-quality measurements and extracting the spatial zones where data were always measured. POD can however be used to reconstruct missing data, by inverse projection on POD basis elements deduced from qualitative data [89]. This could help to extend the statistical set for the learning. Lastly, a linear interpolation of the bathymetry was used to project all the measurements onto the same grid for POD application. The uncertainty that emerged from this interpolation process was not treated. This uncertainty, added to the measurement errors, can shed light on model behavior. For example, comparison of mono-beam cross-sections with multi-beam measurements and uncertainty propagation of bathymetry errors through the learning could be attempted, especially because uncertainties in the bathymetric information may impact the flow field computation [60, 11].

PCE was used to learn the temporal POD modes as 1D data. We showed the importance of the polynomial basis and therefore of the choice of marginals for the learning phase. In this comparison, we showed that choosing uniform distributions associated with Legendre family might not be appropriate, even though it is widely used when no input information is available [103]. Moreover, the number of inputs involved can alter the learning. When using LARS, the presence of numerous variables can mislead the algorithm in selecting useless variables that seem to decrease training error but increase prediction error. This is an overfitting phenomenon. A good combination between polynomial basis and dimension choice could significantly improve convergence speed, centering of residuals and mean training and prediction errors. The proposed contribution analysis using the PCE coefficients showed that the last-state information is often the most influential input. A robustness test was conducted by varying the training set, and the observation was stable. For the modes of small ranks associated with the largest variances (modes 1 to 4 in our example), wave height was the most influential forcing, whatever the chosen learning set. This is consistent with physical knowledge of sediment mobilization in coastal configurations, where waves are known to be determining through the influence they have on bed shear stress [105]. For modes of higher rank, however, the only selected variable by LARS is the last-state information. Firstly, the forcings that were used as PCE inputs were simple statistical estimators deduced from the data (means, percentiles, etc.). This reduction was used instead of giving all the time series as an input, because the problem would become ultrahigh-dimensional. This wastes the richness of the available data for inputs such as tidal information that are measured on an hourly basis. A more accurate statistical reduction of the inputs could be used. For example, PCA and KPCA were used for surrogate modeling with PCE and Gaussian Processes on ultrahigh-dimensional problems in [54] . Secondly, dependencies were not specifically modeled. These can be incorporated using the mathematical setting for the construction of the polynomial basis established in [97]. The dependencies, however, indirectly influenced the construction of the model via selection of basis elements by LARS, which avoids redundancy. Thirdly, the choice of tested input configurations for PCE was arbitrary. A more objective variable selection technique is necessary [81, 22]. Lastly, PCE was chosen for the interpretation possibilities that it allows when combined to POD, thanks to computation of importance measures from the expansion coefficients. Other methods can be used: i.e., for high rank POD modes that were poorly learned using PCE. Indeed, PCE seems to work better for modes associated with high than low

variances. Although it may be tempting to conclude that modeling of high rank modes is not necessary since they are associated with low variances, it should be noted that accurate prediction of modes of higher rank can make the difference between average forecasting and forecasting that captures less frequent events and smaller scale features of the 2D field. Therefore, the present ML could be enhanced by improving the learning of high-rank modes. For example, the construction of marginals and the use of random draw with confidence intervals, or extreme statistics models [26], instead of causal models like PCE, could be attempted.

Finally, the robustness and convergence properties added to the physical interpretability supported the choice of POD-PCE coupling as a ML prediction algorithm. It respects the PDR (Predictive, Descriptive, Relevant) framework defined in [75]. It is characterized by both predictive and descriptive accuracy (simplicity) and is stable with respect to data disturbance. It is interpretable, as the sparsity, simulatability and modularity defined in [75] are respected by construction. Finally, it is both interpretable at features level (POD components and their PCE) and at multidimensional output level (GW compared to the proposed GGW indicators). The POD-PCE ML was therefore implemented using the first 11 modes, after sensitivity test to number of modes. Mean information (e.g. sediment deposition rate) was in general well reproduced. Profile-by-profile investigation also showed that POD-PCE coupling was promising, as the spatial distribution of the sediment deposition patch locations and amplitudes were well represented. Some general limitations should be highlighted and could be good perspectives for improving the process. The small data-set was a clear handicap in our problem. Some events, such as sediment downstream the intake or variation in measurement intervals, were poorly represented. It would be interesting to test the methodology on an enriched data set in order to assess the real potential of POD-PCE Machine Learning. Due to the lack of such data, input distributions were certainly not well approximated. One way of improving POD-PCE coupling would be the development of hybrid measurement-based/process-based data learning [93, 72]. This could be used to enrich the data set, not only by increasing its size (emulated realistic scenarios) but also by adding new input parameters that are not measured but obtained from process-based modeling.

Acknowledgements

This work is funded by the French National Association of Research and Technology (ANRT) through the Industrial Conventions for Training through REsearch (CIFRE) in agreement with EDF R&D. The authors acknowledge their support, and are grateful for data collection and feedback from EDF operators. In particular, we would like to thank D. Rougé for providing the data-set used in this study and for his continuous availability. We also would like to thank Pr. L. Terray (CERFACS) and Dr. M. Rochoux (CERFACS) for constructive discussions on POD and PCE respectively, and Pr. B. Sudret (ETH Zurich) for providing key literature elements on the treatment of ultrahigh dimensional problems and functional inputs using PCE. Finally, the authors gratefully acknowledge the OpenTURNS open source community (An Open source initiative for the Treatment of Uncertainties, Risks'N Statistics).

Appendix A. Additional information on the data set

The available measurements are summarized in Table 1.

Variable	Frequency of data	Covered period	Source of data	Spatial coverage	Descriptive statistics
Bed elevations	Each 2 weeks	2005 - 2018	Plant operator	Multiple profiles of the channel	-
Pumping flowrate	Daily	2007-2018	Plant operator	One global information	-
Dredging date	Each 6 months	2007-2018	Plant operator	One global information	-
Dredging volume	Each 6 months	2007-2018	Plant operator	One global information	-
Tidal level	Hourly	2009-2018	REFMAR [85]	Nearest tidal gauge	min=0.17 ; P10=1.82 ; P50=5.04 ; P90=8.29 ; max=10.21 m
Wind direction	Three-hourly	2009-2018	VAG Model [31]	Nearshore grid point	-
Wind velocity	Three-hourly	2009-2018	VAG Model [31]	Nearshore grid point	min=0.20 ; P10=2.70, P50=7.00 ; P90=12.20 ; max=22.60 m/s
Wave period	Three-hourly	2009-2018	VAG Model [31]	Nearshore grid point	-
Wave height	Three-hourly	2009-2018	VAG Model [31]	Nearshore grid point	min=0.10 ; P10=0.30 ; P50=0.80 ; P90=2.00 ; max=5.70 m
Wave direction	Three-hourly	2009-2018	VAG Model [31]	Nearshore grid point	-

TABLE 1: Summary of measurements, frequencies, periods, sources and spatial coverage. Descriptive statistics are given (P: Percentile).

The forcings did not have the same frequencies. A solution for the homogenization of frequencies is to reduce the measured data to representative statistics over the sedimentation interval $\Delta t \approx 15$ days that separates two bed elevation measurements. Hence, the statistics described in Table 2 are used.

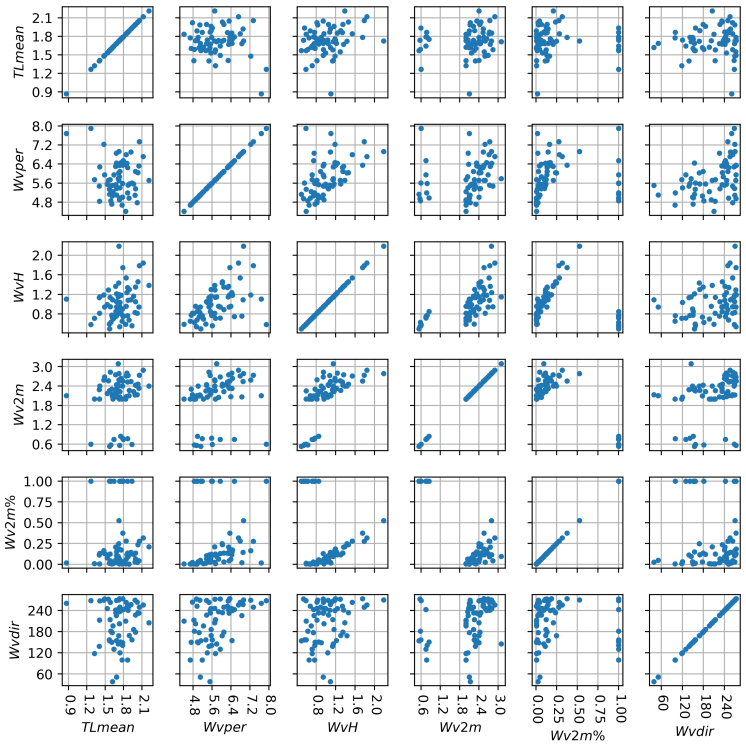


FIGURE 23: Scatter plots of variables $TLmean$, $Wvper$, WvH, $Wv2m$, $Wv2m\%$ and $Wvdir$ from Table 2.

Variable	**Statistic**	**Calculation over Δt**
Pumping flowrate	$Qmean$	Average
Dredging period	Dp	Time lapse since last dredging
Dreding volume	Dv	Last dredged volume
Tidal level	$TLmean$ $TLmin$ $TLrange$ $TLstd$	Average Low Tide Minimum Low Tide Maximum tidal range Standard deviation
Wind direction	$Wdir$	Average direction weighted by wind velocity
Wind velocity	$Wmean$	Average
Wave period	$Wvper$	Average period weighted by the wave heights
Wave height	WvH $Wvstd$ $Wv2m$ $Wv2m\%$	Average Standard deviation Average wave height exceeding $P90 = 2m$ threshold weighted by the associated periods Percentage wave height exceeding the $P90 = 2m$ threshold
Wave direction	$Wvdir$	Average direction weighted by wave height and associated periods

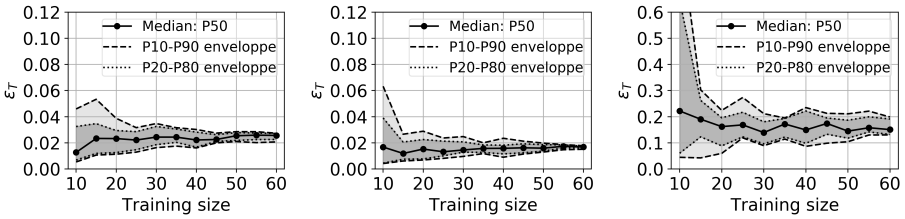
TABLE 2: Reduced variables over the sedimentation periods (P90: the 90^{th} percentile of wave height corresponding to storm events).

Scatter plots of some input variables are presented in Figure 23.

Appendix B. Prediction of a physical field using POD-PCE coupling

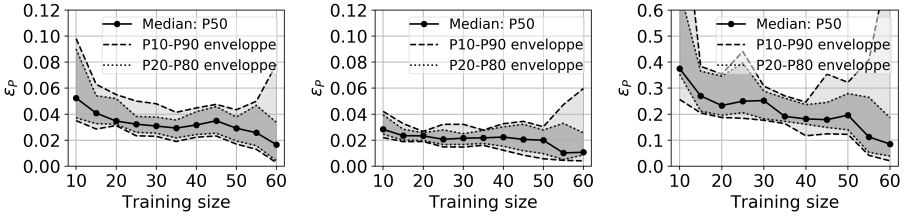
Convergence of the $\mathcal{H}_i^P; Hrm$ model

The impact of training-set size on the learning and prediction errors of the $\mathcal{H}_i^P; Hrm$ model was assessed and can be compared to that of the $\mathcal{H}_i^P; Stlj$ model from Subsection 3.2.2. Evolution of training and prediction empirical errors with training-set size is shown in Figure 10 and 11, respectively.



(a) Mode 1 - Hrm (b) Mode 2 - Hrm (c) Mode 3 - Hrm

FIGURE 24: Training empirical errors ϵ_T, calculated for diverse training-set sizes with 20 random picks among the available data. Plots show median value and confidence intervals (P: Percentile).

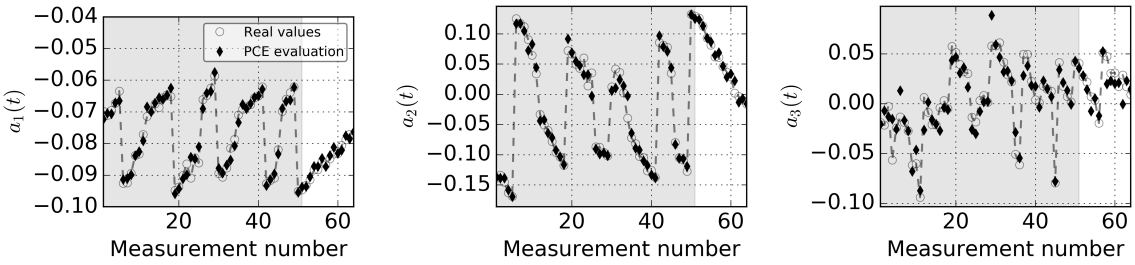


(a) Mode 1 - Hrm (b) Mode 2 - Hrm (c) Mode 3 - Hrm

FIGURE 25: Prediction empirical errors ϵ_P calculated for diverse training-set sizes with 20 random picks among the available data. Plots show median value and confidence intervals (P: Percentile).

Explicit fitted PCE model

For the first three modes, fitting with the model $\mathcal{H}_i^P; Stlj$ is illustrated in Figure 26, and shows good pointwise performance on the fitting and learning intervals.



(a) First mode $a_1(t)$ (b) Second mode $a_2(t)$ (c) Third mode $a_3(t)$

FIGURE 26: Fitting of the first four temporal coefficients using PCE-LARS. Background colored in gray is the training step.

The corresponding explicit PCE models for each POD mode are shown in Equation 22.

$$\begin{cases} a_1(t_k) = -0.0764596 + 0.00169697 * (0.921002 * WvH) + 0.0114921 * (0.976091 * a_1(t_{k-1})) \\ a_2(t_k) = -0.0297343 - 0.00724308 * (0.921002 * WvH) + 0.092532 * (0.958222 * a_2(t_{k-1})) \\ a_3(t_k) = -0.0142516 * (0.921002 * WvH) + 0.031753 * (0.900271 * a_3(t_{k-1})) \\ \qquad -0.00225851 * ((0.948137 * Dt) * (0.921002 * WvH)) \\ \qquad -0.00666008 * (-0.6476 - 0.521557 * WvH + 0.549323 * WvH^2) \\ \qquad -0.000693673 * ((0.97073 * Wv2m) * (0.989947 * Wv2m\%)) \\ \qquad +0.00796314 * ((0.946933 * TLmean) * (0.900271 * a_3(t_{k-1}))) \end{cases} \tag{22}$$

The *Garson Weights* (GW) and *Generalized Garson Weights* (GGW), respectively presented in Sections 2.2 and 2.3, were calculated for each polynomial term, for the first five modes, and shown in Table 3. A visual plot can be found in Figure 15 in Subsection 3.2.2.

Polynomial term	GGW	Total	Mode	GW
$\zeta_{\alpha=(1)}(a_1(t_{k-1}))$	0.64934	0.64934	1	0.87134
$\zeta_{\alpha=(1)}(a_2(t_{k-1}))$	0.10189	0.75123	2	0.92741
$\zeta_{\alpha=(1)}(WvH)$	0.09588	0.84711	1	0.12866
$\zeta_{\alpha=(1)}(a_5(t_{k-1}))$	0.04231	0.88942	5	1.0
$\zeta_{\alpha=(1)}(a_4(t_{k-1}))$	0.03361	0.92303	4	0.70799
$\zeta_{\alpha=(1)}(a_3(t_{k-1}))$	0.02753	0.95056	3	0.49942
$\zeta_{\alpha=(1)}(WvH)$	0.01236	0.96292	3	0.22415
$\zeta_{\alpha=(1)}(WvH)$	0.00798	0.9709	2	0.07259
$\zeta_{\alpha=(1,1)}(TLmean, a_3(t_{k-1}))$	0.0069	0.9778	3	0.12525
$\zeta_{\alpha=(2)}(WvH)$	0.00577	0.98357	3	0.10475
$\zeta_{\alpha=(1,1)}(Dt, TLmean)$	0.00426	0.98783	4	0.08976
$\zeta_{\alpha=(1)}(WvH)$	0.00375	0.99158	4	0.07904
$\zeta_{\alpha=(1,1,1)}(Dt, Wvper, TLmean)$	0.00252	0.9941	4	0.05308
$\zeta_{\alpha=(1,2)}(WvH, Wv2m\%)$	0.00242	0.99652	4	0.05099
$\zeta_{\alpha=(1,1)}(Dt, WvH)$	0.00196	0.99848	3	0.03552
$\zeta_{\alpha=(1,1,1)}(Dt, TLmean, a_4(t_{k-1}))$	0.00063	0.99911	4	0.01328
$\zeta_{\alpha=(1,1)}(Wv2m, Wv2m\%)$	0.0006	0.99971	3	0.01091
$\zeta_{\alpha=(1,1,1)}(WvH, Wv2m, Wv2m\%)$	0.0002	0.99991	4	0.00423
$\zeta_{\alpha=(1,1,1)}(Wv2m, Wv2m\%, a_4(t_{k-1}))$	8e-05	0.99999	4	0.00162

TABLE 3: Full table of the polynomial terms of calibrated PCE models for the 5 first modes ordered by their influence, using the GGWs in Equation 20. Also shown are the GWs calculated as in Equation 7

Références

[1] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad. State-of-the-art in artificial neural network applications : A survey. *Heliyon*, 4(11) :e00938, 2018. ISSN 2405-8440. doi : https://doi.org/10.1016/j.heliyon.2018.e00938. URL `http://www.sciencedirect.com/science/article/pii/S2405844018332067`.

[2] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau. The higgs boson machine learning challenge. In *Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning - Volume 42*, HEPML'14, page 19–55. JMLR.org, 2014.

[3] N. Akkari. *Mathematical study of the sensitivity of the POD method (Proper orthogonal decomposition)*. Theses, Université de La Rochelle, Dec. 2012. URL `https://tel.archives-ouvertes.fr/tel-01066073`.

[4] L. O. Amoudry and A. J. Souza. Deterministic coastal morphological and sediment transport modeling : a review and discussion. *Reviews of Geophysics*, 49(2), 2011. doi : 10.1029/2010RG000341. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010RG000341`.

[5] A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Mach. Learn.*, 14(1) : 115–133, Jan. 1994. ISSN 0885-6125. doi : 10.1023/A:1022650905902. URL https://doi.org/10.1023/A:1022650905902.

[6] G. Blatman. *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis.* PhD thesis, 2009.

[7] G. Blatman and B. Sudret. Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of Computational Physics*, 230(6) :2345 – 2367, 2011. ISSN 0021-9991. doi : https://doi.org/10.1016/j.jcp.2010.12.021. URL http://www.sciencedirect.com/science/article/pii/S0021999110006856.

[8] L. Bruno, C. Canuto, and D. Fransos. Stochastic aerodynamics and aeroelasticity of a flat plate via generalised polynomial chaos. *Journal of Fluids and Structures*, 25(7) :1158 – 1176, 2009. ISSN 0889-9746. doi : https://doi.org/10.1016/j.jfluidstructs.2009.06.001. URL http://www.sciencedirect.com/science/article/pii/S0889974609000693.

[9] S. L. Brunton, B. R. Noack, and P. Koumoutsakos. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 52(1) :477–508, 2020. doi : 10.1146/annurev-fluid-010719-060214. URL https://doi.org/10.1146/annurev-fluid-010719-060214.

[10] K. Campbell, M. D. McKay, and B. J. Williams. Sensitivity analysis when model outputs are functions. *Reliability Engineering & System Safety*, 91(10) :1468 – 1472, 2006. ISSN 0951-8320. doi : https://doi.org/10.1016/j.ress.2005.11.049. URL http://www.sciencedirect.com/science/article/pii/S0951832005002565. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004).

[11] A. Casas, G. Benito, V. Thorndycraft, and M. Rico. The topographic data source of digital terrain models as a key element in the accuracy of hydraulic flood modelling. *Earth Surface Processes and Landforms : The Journal of the British Geomorphological Research Group*, 31(4) :444–456, 2006.

[12] J. Castro, C. Mantas, and J. Benitez. Neural networks with a continuous squashing function in the output are universal approximators. *Neural Networks*, 13(6) :561 – 563, 2000. ISSN 0893-6080. doi : https://doi.org/10.1016/S0893-6080(00)00031-9. URL http://www.sciencedirect.com/science/article/pii/S0893608000000319.

[13] N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning : A tensor analysis. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 698–728, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL http://proceedings.mlr.press/v49/cohen16.html.

[14] L. Cordier and M. Bergmann. Proper orthogonal decomposition : an overview. In *Lecture series 2002-04, 2003-03 and 2008-01 on post-processing of experimental and numerical data, Von Karman Institute for Fluid Dynamics, 2008.*, page 46 pages. VKI, 2008. URL https://hal.archives-ouvertes.fr/hal-00417819.

[15] S. Costa, F. Gourmelon, C. Augris, P. Clabaut, and B. Latteux. Apport de l'approche systémique et pluridisciplinaire dans l'étude du domaine littoral et marin de la seine-maritime (france). *Norois. Environnement, aménagement, société*, (196) :91–108, 2005.

[16] M. Couplet. *Reduced-order POD-Galerkin modelling for the control of unsteady flows.* Theses, Université Paris-Nord - Paris XIII, Jan. 2005. URL https://tel.archives-ouvertes.fr/tel-00142745.

[17] G. Cruciani, M. Baroni, S. Clementi, G. Costantino, D. Riganelli, and B. Skagerberg. Predictive ability of regression models. part i : Standard deviation of prediction errors (sdep). *Journal of Chemometrics*, 6(6) : 335–346, 1992.

[18] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4) :303–314, Dec 1989. ISSN 1435-568X. doi : 10.1007/BF02551274. URL https://doi.org/10.1007/BF02551274.

[19] R. G. Dean and R. A. Dalrymple. *Coastal processes with engineering applications.* Cambridge University Press, 2004.

[20] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986.

[21] R. Eldan and O. Shamir. The power of depth for feedforward neural networks. *Conference on Learning Theory*, 12 2015.

[22] N. Faber and R. Rajkô. How to avoid over-fitting in multivariate calibration—the conventional validation approach and an alternative. *Analytica Chimica Acta*, 595(1) :98 – 106, 2007. ISSN 0003-2670. doi : https://doi.org/10.1016/j.aca.2007.05.030. URL `http://www.sciencedirect.com/science/article/pii/S0003267007009129`. Papers presented at the 10th International Conference on Chemometrics in Analytical Chemistry.

[23] K.-I. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3) :183 – 192, 1989. ISSN 0893-6080. doi : https://doi.org/10.1016/0893-6080(89)90003-8. URL `http://www.sciencedirect.com/science/article/pii/0893608089900038`.

[24] O. Garcia-Cabrejo and A. Valocchi. Global sensitivity analysis for multivariate output using polynomial chaos expansion. *Reliability Engineering & System Safety*, 126 :25–36, 2014.

[25] M. Gevrey, I. Dimopoulos, and S. Lek. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling*, 160(3) :249 – 264, 2003. ISSN 0304-3800. doi : https://doi.org/10.1016/S0304-3800(02)00257-0. URL `http://www.sciencedirect.com/science/article/pii/S0304380002002570`. Modelling the structure of acquatic communities : concepts, methods and problems.

[26] M. Ghil, P. Yiou, S. Hallegatte, B. Malamud, P. Naveau, A. Soloviev, V. Keilis-Borok, D. Kondrashov, V. Kossobokov, O. Mestre, et al. Extreme events : dynamics, statistics and prediction. *Nonlinear Processes in Geophysics*, 18(3) :295, 2011.

[27] E. B. Goldstein, G. Coco, and N. G. Plant. A review of machine learning applications to coastal sediment transport and morphodynamics. *Earth-Science Reviews*, 194 :97 – 108, 2019. ISSN 0012-8252. doi : https://doi.org/10.1016/j.earscirev.2019.04.022. URL `http://www.sciencedirect.com/science/article/pii/S001282521830391X`.

[28] A. Gonoskov, E. Wallin, A. Polovinkin, and I. Meyerov. Employing machine learning for theory validation and identification of experimental conditions in laserplasma physics. *Scientific Reports*, 9 :7043, 2019. doi : 10.1038/s41598-019-43465-3.

[29] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2018.

[30] S. Gordeyev and F. Thomas. Temporal proper decomposition (tpod) for closed-loop flow control. *Exp Fluids*, 54, 2013.

[31] A. Guillaume. *VAG-Modele de prevision de l'etat de la mer en eau profonde*. Dir. de la Meteorologie Nationale, 1987.

[32] M. Guo and J. S. Hesthaven. Reduced order modeling for nonlinear structural analysis using gaussian process regression. *Computer Methods in Applied Mechanics and Engineering*, 341 :807 – 826, 2018. ISSN 0045-7825. doi : https://doi.org/10.1016/j.cma.2018.07.017. URL `http://www.sciencedirect.com/science/article/pii/S0045782518303487`.

[33] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3(null) : 1157–1182, Mar. 2003. ISSN 1532-4435.

[34] B. Hanin. Universal function approximation by deep neural nets with bounded width and relu activations. *Mathematics*, 7(10), 2019. ISSN 2227-7390. doi : 10.3390/math7100992. URL `https://www.mdpi.com/2227-7390/7/10/992`.

[35] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. 02 2009. ISBN 0387848576.

[36] L. Hawchar, C.-P. E. Soueidy, and F. Schoefs. Principal component analysis and polynomial chaos expansion for time-variant reliability problems. *Reliability Engineering & System Safety*, 167 :406 – 416, 2017. ISSN 0951-8320. doi : https://doi.org/10.1016/j.ress.2017.06.024. URL http://www.sciencedirect.com/science/article/pii/S0951832016302587. Special Section : Applications of Probabilistic Graphical Models in Dependability, Diagnosis and Prognosis.

[37] A. Hekmati, D. Ricot, and P. Druault. About the convergence of pod and epod modes computed from cfd simulation. *Computers & Fluids*, 50(1) :60 – 71, 2011. ISSN 0045-7930. doi : https://doi.org/10.1016/j.compfluid.2011.06.018. URL http://www.sciencedirect.com/science/article/pii/S0045793011002064.

[38] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2) :251 – 257, 1991. ISSN 0893-6080. doi : https://doi.org/10.1016/0893-6080(91)90009-T. URL http://www.sciencedirect.com/science/article/pii/089360809190009T.

[39] O. Ibrahim. A comparison of methods for assessing the relative importance of input variables in artificial neural networks. *Journal of applied sciences research*, 9(11) :5692–5700, 2013.

[40] B. Iooss and P. Lemaître. *A Review on Global Sensitivity Analysis Methods*, pages 101–122. Springer US, Boston, MA, 2015. ISBN 978-1-4899-7547-8. doi : 10.1007/978-1-4899-7547-8_5. URL https://doi.org/10.1007/978-1-4899-7547-8_5.

[41] R. Iten, T. Metger, H. Wilming, L. del Rio, and R. Renner. Discovering physical concepts with neural networks. *Phys. Rev. Lett.*, 124 :010508, Jan 2020. doi : 10.1103/PhysRevLett.124.010508. URL https://link.aps.org/doi/10.1103/PhysRevLett.124.010508.

[42] M. Janocko, M. Cartigny, W. Nemec, and E. Hansen. Turbidity current hydraulics and sediment deposition in erodible sinuous channels : laboratory experiments and numerical simulations. *Marine and Petroleum Geology*, 41 :222–249, 2013.

[43] I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486) :682–693, 2009.

[44] I. Jolliffe. *Principal Component Analysis*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi : 10.1007/978-3-642-04898-2_455. URL https://doi.org/10.1007/978-3-642-04898-2_455.

[45] B. A. Jones and A. Doostan. Satellite collision probability estimation using polynomial chaos expansions. *Advances in Space Research*, 52(11) :1860 – 1875, 2013. ISSN 0273-1177. doi : https://doi.org/10.1016/j.asr.2013.08.027. URL http://www.sciencedirect.com/science/article/pii/S0273117713005413.

[46] A. Karpatne, I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar. Machine learning for the geosciences : Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31(8) :1544–1554, Aug 2019. ISSN 2326-3865. doi : 10.1109/TKDE.2018.2861006.

[47] G. Kerschen and J. Golinval. Physical interpretation of the proper orthogonal modes using the singular value decomposition. *Journal of Sound and Vibration*, 249(5) :849 – 865, 2002. ISSN 0022-460X. doi : https://doi.org/10.1006/jsvi.2001.3930. URL http://www.sciencedirect.com/science/article/pii/S0022460X01939306.

[48] G. Kerschen, J. Golinval, V. A.F., and B. L.A. The method of proper orthogonal decomposition for dynamical characterization and order reduction of mechanical systems : an overview. *Nonlinear Dynamis*, 41 :147, 2002. ISSN 1573-269X. doi : https://doi.org/10.1007/s11071-005-2803-2.

[49] O. M. Knio and O. P. Le Maître. Uncertainty propagation in CFD using polynomial chaos decomposition. *Fluid Dynamics Research*, 38(9) :616–640, sep 2006. doi : 10.1016/j.fluiddyn.2005.12.003. URL https://doi.org/10.1016%2Fj.fluiddyn.2005.12.003.

[50] J. Kremer, K. Stensbo-Smidt, F. Gieseke, K. S. Pedersen, and C. Igel. Big universe, big data : Machine learning and image analysis for astronomy. *IEEE Intelligent Systems*, 32(2) :16–22, Mar 2017. ISSN 1941-1294. doi : 10.1109/MIS.2017.40.

[51] J. N. Kutz. Deep learning in fluid dynamics. *Journal of Fluid Mechanics*, 814 :1–4, 2017. doi : 10.1017/jfm.2016.803.

[52] M. Lamboni, H. Monod, and D. Makowski. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety*, 96(4) :450 – 459, 2011. ISSN 0951-8320. doi : https://doi.org/10.1016/j.ress.2010.12.002. URL http://www.sciencedirect.com/science/article/pii/S0951832010002504.

[53] M. Larson, M. Capobianco, M. Jansen, G. Różyński, H. Southgate, M. Stive, K. Wijnberg, and S. Hulscher. Analysis and modeling of field data on coastal morphological evolution over yearly and decadal time scales. part 1 : Background and linear techniques. *Journal of Coastal Research*, 19, 09 2003.

[54] C. Lataniotis, S. Marelli, and B. Sudret. Extending classical surrogate modelling to ultrahigh dimensional problems through supervised dimensionality reduction : a data-driven approach. 12 2018.

[55] A. Laudani, G. M. Lozito, F. R. Fulginei, and A. Salvini. On training efficiency and computational costs of a feed forward neural network : A review. *Computational Intelligence and Neuroscience*, 2015. doi : 10.1155/2015/818243.

[56] S. Le Bot, R. Lafite, M. Fournier, A. Baltzer, and M. Desprez. Morphological and sedimentary impacts and recovery on a mixed sandy to pebbly seabed exposed to marine aggregate extraction (eastern english channel, france). *Estuarine, Coastal and Shelf Science*, 89(3) :221–233, 2010.

[57] O. P. Le Maitre, O. M. Knio, H. N. Najm, and R. G. Ghanem. A stochastic projection method for fluid flow : I. basic formulation. *Journal of Computational Physics*, 173(2) :481 – 511, 2001. ISSN 0021-9991. doi : https://doi.org/10.1006/jcph.2001.6889. URL http://www.sciencedirect.com/science/article/pii/S0021999101968895.

[58] O. P. Le Maitre, M. T. Reagan, H. N. Najm, R. G. Ghanem, and O. M. Knio. A stochastic projection method for fluid flow : Ii. random process. *Journal of Computational Physics*, 181(1) :9 – 44, 2002. ISSN 0021-9991. doi : https://doi.org/10.1006/jcph.2002.7104. URL http://www.sciencedirect.com/science/article/pii/S0021999102971044.

[59] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521 :436–444, 2015. doi : https://doi.org/10.1038/nature14539.

[60] C. J. Legleiter, P. C. Kyriakidis, R. R. McDonald, and J. M. Nelson. Effects of uncertain topographic input data on two-dimensional flow modeling in a gravel-bed river. *Water Resources Research*, 47(3), 2011. doi : 10.1029/2010WR009618. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010WR009618.

[61] Y. Liang, H. Lee, S. Lim, W. Lin, K. Lee, and C. WU. Proper orthogonal decomposition and its applications — part i : Theory. *Journal of Sound and Vibration*, 252(3) :527 – 544, 2002. ISSN 0022-460X. doi : https://doi.org/10.1006/jsvi.2001.4041. URL http://www.sciencedirect.com/science/article/pii/S0022460X01940416.

[62] Y. Liang, W. Lin, H. Lee, S. Lim, K. Lee, and H. Sun. Proper orthogonal decomposition and its applications — part ii : Model reduction for mems dynamical analysis. *Journal of Sound and Vibration*, 252(3) :527 – 544, 2002. doi : https://doi.org/10.1006/jsvi.2002.5007.

[63] T. E. Lovett, F. Ponci, and A. Monti. A polynomial chaos approach to measurement uncertainty. *IEEE Transactions on Instrumentation and Measurement*, 55(3) :729–736, June 2006. ISSN 1557-9662. doi : 10.1109/TIM.2006.873807.

[64] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks : A view from the width. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6231–6239. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7203-the-expressive-power-of-neural-networks-a-view-from-the-width.pdf.

[65] D. Lucor, C.-H. Su, and G. E. Karniadakis. Generalized polynomial chaos and random oscillators. *International Journal for Numerical Methods in Engineering*, 60(3) :571–596, 2004. doi : 10.1002/nme.976. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.976.

[66] J. L. Lumley. The structure of inhomogeneous turbulent flows. *Atmospheric Turbulence and Radio Wave Propagation*, 1967.

[67] M. Ma, J. Lu, and G. Tryggvason. Using statistical learning to close two-fluid multiphase flow equations for bubbly flows in vertical channels. *International Journal of Multiphase Flow*, 85 :336 – 347, 2016. ISSN 0301-9322. doi : https://doi.org/10.1016/j.ijmultiphaseflow.2016.06.021. URL http://www.sciencedirect.com/science/article/pii/S0301932215302226.

[68] C. Michel, S. Le Bot, F. Druine, S. Costa, F. Levoy, C. Dubrulle-Brunaud, and R. Lafite. Stages of sedimentary infilling in a hypertidal bay using a combination of sedimentological, morphological and dynamic criteria (bay of somme, france). *Journal of Maps*, 13(2) :858–865, 2017.

[69] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch. Kernel pca and de-noising in feature spaces. In *Advances in neural information processing systems*, pages 536–542, 1999.

[70] K. Mills, M. Spanner, and I. Tamblyn. Deep learning and the Schrödinger equation. *Phys. Rev. A*, 96 :042113, Oct 2017. doi : 10.1103/PhysRevA.96.042113. URL https://link.aps.org/doi/10.1103/PhysRevA.96.042113.

[71] R. Morrow, L.-L. Fu, F. Ardhuin, M. Benkiran, B. Chapron, E. Cosme, F. d'Ovidio, J. T. Farrar, S. T. Gille, G. Lapeyre, P.-Y. Le Traon, A. Pascual, A. Ponte, B. Qiu, N. Rascle, C. Ubelmann, J. Wang, and E. D. Zaron. Global observations of fine-scale ocean surface topography with the surface water and ocean topography (swot) mission. *Frontiers in Marine Science*, 6 :232, 2019. ISSN 2296-7745. doi : 10.3389/fmars.2019.00232. URL https://www.frontiersin.org/article/10.3389/fmars.2019.00232.

[72] A. Mosavi, S. Shamshirband, E. Salwana, K.-w. Chau, and J. H. Tah. Prediction of multi-inputs bubble column reactor using a novel hybrid model of computational fluid dynamics and machine learning. *Engineering Applications of Computational Fluid Mechanics*, 13(1) :482–492, 2019.

[73] M. Muller. *On the POD method : an abstract investigation with applications to reduced-order modeling and suboptimal control.* PhD thesis, 2008.

[74] E. Muravleva, I. Oseledets, and D. Koroteev. Application of machine learning to viscoplastic flow modeling. *Physics of Fluids*, 30(10) :103102, 2018. doi : 10.1063/1.5058127. URL https://doi.org/10.1063/1.5058127.

[75] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Interpretable machine learning : definitions, methods, and applications. *arXiv preprint arXiv :1901.04592*, 2019.

[76] H. N. Southgate, K. Wijnberg, M. Larson, M. Capobianco, and H. Jansen. Analysis of field data of coastal morphological evolution over yearly and decadal timescales. part 2 : Non-linear techniques. *Journal of Coastal Research*, 19, 09 2003.

[77] J. B. Nagel, J. Rieckermann, and B. Sudret. Principal component analysis and sparse polynomial chaos expansions for global sensitivity analysis and model calibration : Application to urban drainage simulation. *Reliability Engineering & System Safety*, 195 :106737, 2020. ISSN 0951-8320. doi : https://doi.org/10.1016/j.ress.2019.106737. URL http://www.sciencedirect.com/science/article/pii/S0951832019301747.

[78] NASA (National Aeronautics and Space Administration) and CNES (Centre National d'Etudes Spatiales) in partnership with CSA (Canadian Space Agency) and UKSA (UK Space Agency). Surface water and ocean topography. https://swot.jpl.nasa.gov/home.htm, 2021.

[79] A. Ng. Machine learning lecture notes. http://cs229.stanford.edu/materials.html. Stanford Univ. TR, Stanford, CA, 2014.

[80] B. T. Nguyen, A. Samimi, and J. J. Simpson. A polynomial chaos approach for em uncertainty propagation in 3d-fdtd magnetized cold plasma. In *2015 IEEE Symposium on Electromagnetic Compatibility and Signal Integrity*, pages 356–360, March 2015. doi : 10.1109/EMCSI.2015.7107714.

[81] R. Noori, A. Karbassi, A. Moghaddamnia, D. Han, M. Zokaei-Ashtiani, A. Farokhnia, and M. G. Gousheh. Assessment of input variables determination on the svm model performance using pca, gamma test, and forward selection techniques for monthly stream flow prediction. *Journal of Hydrology*, 401(3) :177 – 189, 2011. ISSN 0022-1694. doi : https://doi.org/10.1016/j.jhydrol.2011.02.021. URL http://www.sciencedirect.com/science/article/pii/S0022169411001363.

[82] A. T. N. Papanicolaou, M. Elhakeem, G. Krallis, S. Prakash, and J. Edinger. Sediment transport modeling review&#x2014 ;current and future developments. *Journal of Hydraulic Engineering*, 134(1) :1–14, 2008. doi : 10.1061/(ASCE)0733-9429(2008)134:1(1). URL https://ascelibrary.org/doi/abs/10.1061/%28ASCE%290733-9429%282008%29134%3A1%281%29.

[83] M. S. Parsons. Interpretation of machine-learning-based disruption models for plasma control. *Plasma Physics and Controlled Fusion*, 59(8) :085001, jun 2017. doi : 10.1088/1361-6587/aa72a3. URL https://doi.org/10.1088%2F1361-6587%2Faa72a3.

[84] S. Paul and M. K. Verma. *Proper Orthogonal Decomposition vs. Fourier Analysis for Extraction of Large-Scale Structures of Thermal Convection*, pages 433–441. 2017.

[85] REFMAR. Réseaux de Référence des observations MARégraphiques, 2020.

[86] A. Rigos, G. E. Tsekouras, A. Chatzipavlis, and A. F. Velegrakis. Modeling Beach Rotation Using a Novel Legendre Polynomial Feedforward Neural Network Trained by Nonlinear Constrained Optimization. In L. Iliadis and I. Maglogiannis, editors, *Artificial Intelligence Applications and Innovations*, pages 167–179, Cham, 2016. Springer International Publishing. ISBN 978-3-319-44944-9.

[87] D. Rolnick and M. Tegmark. The power of deeper networks for expressing natural functions, 2017.

[88] B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, and P. A. Johnson. Machine learning predicts laboratory earthquakes. *Geophysical Research Letters*, 44(18) :9276–9282, 2017. doi : 10.1002/2017GL074677. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL074677.

[89] P. Saini, C. M. Arndt, and A. M. Steinberg. Development and evaluation of gappy-pod as a data reconstruction technique for noisy piv measurements in gas turbine combustors. *Experiments in Fluids*, 57(7) :122, 2016.

[90] P. J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656 :5–28, 2010. doi : 10.1017/S0022112010001217.

[91] J. Schmidhuber. Deep learning in neural networks : An overview. *Neural Networks*, 61 :85 – 117, 2015. ISSN 0893-6080. doi : https://doi.org/10.1016/j.neunet.2014.09.003. URL http://www.sciencedirect.com/science/article/pii/S0893608014002135.

[92] P. D. Sclavounos, Y. Ma, P. D. Sclavounos, and Y. Ma. Artificial intelligence machine learning in marine hydrodynamics. In *Proceedings of the ASME 2018 37th International Conference on Ocean, Offshore and Arctic Engineering,17-22 June, Madrid, Spain, ASME, 2018*, 06 2018. doi : 10.1115/OMAE2018-77599. URL https://doi.org/10.1115/OMAE2018-77599.

[93] J. Senent-Aparicio, P. Jimeno-Sáez, A. Bueno-Crespo, J. Pérez-Sánchez, and D. Pulido-Velázquez. Coupling machine-learning techniques with swat model for instantaneous peak flow prediction. *Biosystems Engineering*, 177 :67 – 77, 2019. ISSN 1537-5110. doi : https://doi.org/10.1016/j.biosystemseng.2018.04.022. URL http://www.sciencedirect.com/science/article/pii/S1537511017311686. Intelligent Systems for Environmental Applications.

[94] S. Sengupta, S. Basak, P. Saikia, S. Paul, V. Tsalavoutis, F. Atiah, V. Ravi, and A. Peters. A review of deep learning with special emphasis on architectures, applications and recent trends, 2019.

[95] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning : From Theory to Algorithms*. Cambridge University Press, 2014.

[96] L. Sirovich. Turbulence and the dynamics of coherent structures : I, ii and iii. *Quarterly Applied Mathematics*, 45 :561, 1987.

[97] C. Soize and R. Ghanem. Physical systems with random uncertainties : chaos representations with arbitrary probability measure. *SIAM J. Sci. Comput.*, pages 26(2), 395–410, 2004.

[98] T. Sruthi, K. Ranjith, and V. Chandra. Control of sediment entry into an intake canal by using submerged vanes. In *AIP Conference Proceedings*, volume 1875, page 030007. AIP Publishing LLC, 2017.

[99] B. Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7) :964 – 979, 2008. ISSN 0951-8320. doi : https://doi.org/10.1016/j.ress.2007.04.002. URL http://www.sciencedirect.com/science/article/pii/S0951832007001329. Bayesian Networks in Dependability.

[100] B. Sudret. *Polynomial chaos expansions and stochastic finite element methods*, page 624. CRC PressEditors : Kok-Kwang Phoon, Jianye Ching, 12 2014.

[101] K. Taira, S. L. Brunton, S. T. M. Dawson, C. W. Rowley, T. Colonius, B. J. McKeon, O. T. Schmidt, S. Gordeyev, V. Theofilis, and L. S. Ukeiley. Modal analysis of fluid flows : An overview. *AIAA Journal*, 55 (12) :4013–4041, 2017. doi : 10.2514/1.J056060. URL https://doi.org/10.2514/1.J056060.

[102] A. Tarakanov and A. H. Elsheikh. Regression-based sparse polynomial chaos for uncertainty quantification of subsurface flow models. *Journal of Computational Physics*, 399 :108909, 2019. ISSN 0021-9991. doi : https://doi.org/10.1016/j.jcp.2019.108909. URL http://www.sciencedirect.com/science/article/pii/S002199911930614X.

[103] E. Torre, S. Marelli, P. Embrechts, and B. Sudret. Data-driven polynomial chaos expansion for machine learning regression. *Journal of Computational Physics*, 2019.

[104] M. Tsang, D. Cheng, and Y. Liu. Detecting statistical interactions from neural network weights. *arXiv preprint arXiv :1705.04977*, 2017.

[105] L. C. van Rijn. Unified view of sediment transport by currents and waves. i : Initiation of motion, bed roughness, and bed-load transport. *Journal of Hydraulic Engineering*, 133(6) :649–667, 2007. doi : 10.1061/(ASCE)0733-9429(2007)133:6(649). URL https://ascelibrary.org/doi/abs/10.1061/%28ASCE%290733-9429%282007%29133%3A6%28649%29.

[106] J. VanderPlas, A. J. Connolly, Z. Ivezic, and A. Gray. Introduction to astroml : Machine learning for astrophysics. In *2012 Conference on Intelligent Data Understanding*, pages 47–54, Oct 2012. doi : 10.1109/CIDU.2012.6382200.

[107] X. Wan and G. Karniadakis. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. *Journal of Computational Physics*, 2006.

[108] Q. Wang, J. S. Hesthaven, and D. Ray. Non-intrusive reduced order modeling of unsteady flows using artificial neural networks with application to a combustion problem. *Journal of Computational Physics*, 384 :289 – 307, 2019. ISSN 0021-9991. doi : https://doi.org/10.1016/j.jcp.2019.01.031. URL http://www.sciencedirect.com/science/article/pii/S0021999119300828.

[109] N. Wiener. The homogeneous chaos. *American Journal of Mathematics*, pages 60, 897–936, 1938.

[110] S. Wilkinson, S. Hanna, L. Hesselgren, and V. Mueller. Inductive aerodynamics. In *Proceedings of eCAADe 2013 : Computation and Performance. pp.39-48.*, 09 2013.

[111] J. A. Witteveen and H. Bijl. *Modeling Arbitrary Uncertainties Using Gram-Schmidt Polynomial Chaos*. 2006. doi : 10.2514/6.2006-896. URL https://arc.aiaa.org/doi/abs/10.2514/6.2006-896.

[112] D. Xiu and G. Karniadakis. Modelling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Comput. Methods Appl. Mec. Engrg*, pages 191(43), 4927–4948, 2003.

[113] D. Xiu and G. E. Karniadakis. The wiener–askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2) :619–644, 2002. doi : 10.1137/S1064827501387826. URL https://doi.org/10.1137/S1064827501387826.

[114] D. Xiu and G. E. Karniadakis. Modeling uncertainty in flow simulations via generalized polynomial chaos. *Journal of Computational Physics*, 187(1) :137 – 167, 2003. ISSN 0021-9991. doi : https://doi.org/10.1016/S0021-9991(03)00092-5. URL http://www.sciencedirect.com/science/article/pii/S0021999103000925.

[115] L. Zhu, W. Zhang, J. Kou, and Y. Liu. Machine learning methods for turbulence modeling in subsonic flows around airfoils. *Physics of Fluids*, 31(1) :015105, 2019. doi : 10.1063/1.5061693. URL https://doi.org/10.1063/1.5061693.