THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Institut National Polytechnique de Toulouse

Présentée et soutenue par

Rem-Sophia MOURADI

Le 16 mars 2021

Modélisation non-linéaire de champs multidimensionnels guidée par la donnée : application aux écoulements côtiers hydro-morphodynamiques

Ecole doctorale : **SDU2E - Sciences de l'Univers, de l'Environnement et de l'Espace**

Spécialité : **Océan, Atmosphère, Climat**

Unité de recherche :
CECI - Climat, Environnement, Couplages et Incertitudes / CERFACS

Thèse dirigée par
Olivier THUAL et Cédric GOEURY

Jury

M. Olivier Le Maître, Rapporteur
M. Pierre-Olivier Malaterre, Rapporteur
Mme Clémentine Prieur, Rapporteure
M. Florent Lyard, Examinateur
Mme Hélène Roux, Examinatrice
Mme Christine Keribin, Examinatrice
M. Olivier THUAL, Directeur de thèse
M. Cédric Goeury, Directeur de thèse

M. Pablo Tassi, invité
M. Fabrice Zaoui, invité

Non-linear data-driven modelling on multidimensional fields: an application to hydro-morphodynamic coastal flows

Rem-Sophia MOURADI

January 13, 2021

Contents

List of Figures

List of Tables

Introduction

Au 21ème siècle, la recherche scientifique dans de nombreux domaines de la physique implique de traiter diverses sources de données. Les observations deviennent de plus en plus nombreuses et accessibles, grâce aux avancées technologiques de mesure, stockage de données, capacité de transmission et traitement. Par ailleurs, la puissante avancée historique des ressources computationnelles, à partir de la fin des années 90, a encouragé l'usage des modèles à base de processus physiques, permettant d'améliorer notre compréhension des phénomènes observés, et de générer de la donnée supplémentaire. C'est ainsi que les approches guidées par la donnée sont devenues une des pierres angulaires de la physique, allant de l'Assimilation de Données (AD) au Machine Learning (ML).

Cette thèse se concentre sur le ML interprétable, en utilisant les techniques de Réduction de Dimension (DR) et de régression probabiliste non-linéaire et multivariée, en combinant deux approches classiques : la Décomposition en modes Propres Orthogonaux (POD), et l'Expansion par Polynômes du Chaos (PCE). La méthodologie proposée est appliquée à différentes étapes de la modélisation guidée par la donnée : (i) l'apprentissage à base de données mesurées, (ii) la Quantification des Incertitudes (UQ) efficace et (iii) l'Assimilation de Données rapide et précise.

Les contributions présentées découlent d'investigations menées au sein de la communauté scientifique des géosciences, qui connaît par ailleurs une augmentation constante des sources de données. Les méthodologies proposées ont pour but de fournir un outil prédictif dans un contexte industriel, avec des défis sous-jacents, notamment concernant les contraintes liées au temps de calcul. Plus précisement, la modélisation de la morphodynamique dans un chenal bord-de-mer de centrale électrique est visée. Des données de surveillance du chenal, collectées durant plusieurs années dans le but d'optimiser sa gestion, ainsi qu'un modèle numérique hydro-morphodynamique, sont disponibles. L'objectif principal de cette thèse est donc d'établir une méthodologie de couplage optimal entre données de terrain et modélisation numérique, en utilisant des outils statistiques adaptés. La finalité est de prédire de manière rapide et précise l'élévation du lit sous-marin, aussi appelée bathymétrie. Cette méthodologie est appliquée dans une configuration côtière, avec pour objectif de mieux comprendre la morphodynamique (évolution des bathymétries). Cet aspect est crucial pour plusieurs applications, en particulier pour la prédiction de l'écoulement résultant, ce qui peut être d'intérêt socio-économique (par exemple pour prédire des inondations), ou d'intérêt industriel comme pour l'application proposée.

In the 21st century, scientific research in many physical fields involves dealing with a variety of data sources. Observations are becoming plentiful and accessible, due to technological advances in measurement devices, data storage, transmission and treatment capacities. The powerful historical jump of computational resources since late 90's has encouraged the use of process-based models, allowing to improve our understanding of physical phenomena, and to generate additional data. This is how data-driven approaches have become one of the cornerstones of physics, from Data Assimilation (DA) to Machine Learning (ML). As an example, ML has gained interest from classical physics (fluid mechanics [32, 116, 145, 174]; aerodynamics [265, 275]; plasma physics [83, 188], astrophysics and astronomy [115, 258]) to quantum physics (particle physics [4]; quantum mechanics [164]). Perhaps, physicists nourish a hope about exploring the "chasm of ignorance" using data-based techniques, by pushing the boundaries of classical approaches [105]. Although time has not yet come for drastic change [105], and believing that data may come with added value to previously established theories, one may ask the following: what are the optimal combinations between all information sources? and how to help physical modelling advances using statistical tools? These are open research questions that we do not pretend to solve in one thesis, but represent a guideline for the presented discussions.

The thesis work focuses on interpretable ML using Dimensionality Reduction (DR) and non-linear multivariate probabilistic regression, by combining two classical approaches: POD and Polynomial Chaos Expansion (PCE). The proposed methodology is applied at various steps of the data-driven modelling: (i) pure measurement based learning; (ii) efficient UQ and (iii) fast and accurate DA. In particular, the work takes place in the geosciences community, which also registers a constant increase in data sources [109]. Some interesting programs can be cited as the new SWOT satellite mission [166, 176], or the Sentinel satellite missions (Copernicus program) [66, 147]. Consequently, the community is also keen on data-based works, with increasingly represented contributions in ML [109, 205, 216], DA [39, 72], and UQ [19, 169]. These methods are of particular interest for example, in a context of climate change, where new data constantly need to be taken into account [204].

The proposed steps of this thesis aim at providing a predictive tool in an industrial context with inherent challenges, for example concerning computational time constraints. More precisely, the modelling of *morphodynamics* in a coastal power plant's water intake was targeted. Data were collected during many years of monitoring, in order to optimize the intake's management, and a *hydro-morphodynamic* numerical model is available. The main objective of this thesis is therefore to enable an optimal coupling methodology between field data and numerical modelling using appropriate statistical tools, for fast and accurate prediction of underwater topography (also called *bathymetry* or *bottom/bed* elevation). This methodology is applied in a coastal set-up, with the goal of better understanding sea bed temporal evolution, known as *morphodynamics*. This topic is crucial for many applications, in particular for the prediction of the resulting flow, which can be of socio-economical interest (e.g floods prediction) or industrial interest (e.g. current application).

Industrial context: power plants monitoring

Water intakes are a crucial component of power plants, as they ensure their cooling process via a pumping system. The plants are therefore constructed near to natural water