Development of a data assimilation method for the calibration and continuous update of wind turbines digital twins.

Adrien Hirvoas

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : Mathématiques Appliquées

Arrêté ministériel : 25 mai 2016

Présentée par

Adrien HIRVOAS

Thèse dirigée par **Clémentine PRIEUR, Professeur, UGA/LJK**
et codirigée par **Elise, ARNAUD, Maître de conférences, UGA/LJK**

préparée au sein **du Laboratoire Jean Kuntzmann**
dans **l'École Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Développement d'une méthode d'assimilation de données pour la calibration et la mise à jour en continu de modèles fidèles d'éoliennes.

Development of a data assimilation method for the calibration and continuous update of wind turbines digital twins.

Thèse soutenue publiquement le **30 mars 2021**,
devant le jury composé de :

Nathalie BARTOLI
Professeur, ONERA/ISAE-SUPAERO, Rapporteur
Valérie MONBET
Professeur, Université Rennes 1, Rapporteur
Eric BLAYO
Professeur, Université Grenoble Alpes/LJK, Président
Anne CUZOL
Maître de conférences, Université Bretagne Sud, Examinateur
Nicolas GAYTON
Professeur, SIGMA Clermont, Examinateur
Clémentine PRIEUR
Professeur Université Grenoble Alpes/LJK, Directrice de thèse
Bruno SUDRET
Professeur, ETH Zurich, Examinateur
Elise, ARNAUD
Maître de conférences, Université Grenoble Alpes/LJK, Co-encadrante de thèse
Fabien CALEYRON
Ingénieur de recherche, IFP Energies nouvelles, Invité
Miguel MUNOZ ZUNIGA
Ingénieur de recherche, IFP Energies nouvelles, Invité

"To my family, my friends, and my godson."

Abstract

In the context of energy transition, wind power generation is developing rapidly. Meanwhile, in the framework of digitalization of the industry, the exploitation of collected data can be optimized by combination with numerical models. Such models can be complex and costly as they involve dynamic equations coupled with different physics. Furthermore, some of their input parameters related to the model properties as well as the external conditions can be badly known. These uncertainties affect the predictions obtained from model simulations and thus can impact the components' lifetime for example. Consequently, this dissertation focuses on quantifying and reducing the input parameter uncertainties involved in an aero-servo-elastic wind turbine model. Nevertheless, the widely used methods in uncertainty quantification are not suitable in the present industrial context because of the stochastic nature of the external solicitation and the time consuming behavior of the simulator. Our main contributions are twofold.

Firstly, we want to quantify the impact of the uncertainties on the fatigue behavior of a wind turbine. We propose a global sensitivity analysis (GSA) methodology, based on the so-called Sobol' indices, for stochastic computer simulations. Such techniques, which often refer to the probabilistic framework and Monte Carlo (MC) methods, require a lot of calls to the numerical model. The uncertain input parameters are modeled by independent random variables gathered into a random vector and characterized by their probability distribution function (pdf). Variance-based GSA for time consuming deterministic computer models is usually performed by approximating the model by a surrogate regression. Among the different surrogates, we focus on Gaussian process (GP) regression characterized by its mean and covariance functions. One advantage of the GP regression metamodeling is to provide both a prediction of the numerical model and the associated uncertainty. In order to take into account the inherent randomness from stochastic simulations, we propose as a surrogate for the mean of the output of interest a GP regression with heteroscedastic noise. Then, this surrogate model is used to perform a sensitivity analysis based on classical MC estimation procedure.

Secondly, we propose a Bayesian inference framework to carry out the calibration of influential input parameters from in situ measurements. It uses some observations to update some prior pdf on the unknown input parameters through the Bayes' theorem. Recent decades have been marked by a simultaneous development of sensor technologies and internet of things capabilities. Thus, our research efforts have been directed toward inference techniques where the data are sequentially processed when new observations

become available. In this context, model parameter inference can be carried out using data assimilation methods. We carry out the calibration using an ensemble Kalman filter (EnKF). Nevertheless, unlike the model properties having a static or slow time-variant behavior, the parameters related to the external conditions have a dynamic aspect. Thus, we propose to carry out the inference problem using an EnKF coupled with an analog forecasting strategy based on K-nearest neighbors to model the underlying dynamics. However, such problems can be solved assuming that several conditions of well-posedness and identifiability are achieved. We exploit the relationship between non-identifiability of input parameters and total Sobol' indices. Indeed, for each measured output, we compute total Sobol' indices associated to input parameters. If all the total Sobol' indices associated to a prescribed input parameter are "small", it means that this parameter is non-identifiable. Due to the functional nature of the measurements, we rely on a dimension reduction preliminary step through principal component analysis and then we compute an aggregated Sobol' index for each model parameter.

Keywords: Ensemble Kalman filter, Gaussian process regression model, K-nearest neighbors, aggregated Sobol' indices, wind turbine numerical models

Résumé

Dans un contexte énergétique en pleine transition, l'énergie d'origine éolienne se développe rapidement. Parallèlement, dans le cadre de la digitalisation de l'industrie, l'exploitation des données collectées peut être optimisée par combinaison avec des modèles numériques d'éoliennes. Ces modèles peuvent être complexes et coûteux car ils impliquent des équations dynamiques couplées à différentes physiques. De plus, certains de leurs paramètres d'entrée peuvent être mal ou peu connus. Ces incertitudes affectent les prédictions obtenues à partir de ces simulations et peuvent avoir un impact important sur la surveillance de l'état de la structure. Cette thèse se concentre sur la quantification et la réduction des incertitudes des paramètres d'entrée d'un modèle aéro-servo-élastique d'une éolienne. Néanmoins, les méthodes largement utilisées de quantification des incertitudes ne conviennent pas à notre contexte industriel du fait de la nature stochastique et du coût de chaque évaluation du simulateur. Nos principales contributions sont les suivantes.

Premièrement, nous quantifions l'impact des incertitudes sur le comportement en fatigue d'une éolienne. Nous proposons une méthodologie d'analyse de sensibilité globale (ASG) basée sur les indices de Sobol' dans le cadre de simulations numériques stochastiques. De telles techniques, qui font souvent référence au cadre probabiliste et aux méthodes de Monte Carlo (MC), nécessitent de nombreux appels au modèle. Les paramètres d'entrée incertains sont modélisés par des variables aléatoires indépendantes regroupées dans un vecteur aléatoire et caractérisées par leur loi de probabilité. De telles analyses pour des simulations déterministes coûteuses en temps de calcul sont en général réalisées en approchant le modèle par un métamodèle. Nous nous concentrons sur un métamodèle de type processus gaussien (PG) caractérisé par sa moyenne et sa fonction de covariance. Il présente l'avantage de fournir à la fois une prédiction du modèle numérique et l'incertitude associée. Cependant, l'ASG basée sur ce type de modèle de substitution ne tient pas compte du caractère aléatoire inhérent à la simulation stochastique. Ainsi, nous proposons de modéliser la moyenne de la sortie d'intérêt avec un modèle par un PG avec bruit hétéroscédastique. Ensuite, ce métamodèle est utilisé pour effectuer une ASG avec une procédure classique d'estimation MC.

Deuxièmement, nous proposons une procédure d'inférence bayésienne à partir de mesures in situ permettant de réduire les incertitudes qui entachent les paramètres d'entrée influents sur le comportement en fatigue de l'éolienne. Les dernières décennies ont été marquées par un développement simultané des technologies de capteurs et de l'internet

des objets. Ainsi, nos efforts de recherche ont été orientés vers des techniques d'inférence où les données sont traitées séquentiellement lorsque de nouvelles observations deviennent disponibles. Dans ce contexte, l'inférence des paramètres du modèle peut être effectuée à l'aide de méthodes d'assimilation de données. Nous nous focalisons tout particulièrement sur le filtre de Kalman d'ensemble (EnKF). Lorsque le modèle dynamique sous-jacent des paramètres d'entrée est inconnu, nous proposons d'utiliser une procédure d'inférence combinant un EnKF à une stratégie de prévision par analogues basée sur une méthode des plus proches voisins. Cependant, seule l'inférence des paramètres identifiables a du sens. Un paramètre n'ayant aucune influence sur les sorties mesurées n'est pas identifiable. Cette influence est quantifiée en estimant les indices de Sobol' totaux des sorties mesurées aux paramètres d'entrée. En raison de la nature fonctionnelle de ces observations, nous nous appuyons sur une réduction de dimension par analyse en composantes principales préalable à l'estimation d'un indice de Sobol' agrégé pour chaque sortie mesurée aux paramètres du modèle.

Mot-clés : Filtre de Kalman d'ensemble, modèle de regression par processus Gaussien, K-plus proches voisins, indices de Sobol' agrégés, modèles numériques d'éolienne

Remerciements

Je tiens à remercier l'ensemble de mon équipe encadrante pour m'avoir conseillé et accompagné tout au long de ces trois ans de thèse. D'une part, Clémentine Prieur et Elise Arnaud pour leur disponibilité, leur encadrement rigoureux, et leur soutien infaillible durant cette aventure scientifique. Je souhaite aussi remercier Miguel Munoz Zuniga pour sa disponibilité, pour m'avoir guidé sur les aspects théoriques et m'avoir apporté sa rigueur scientifique. Je tiens à présent à remercier Fabien Caleyron pour la confiance qu'il a su m'accorder sur ce sujet et l'autonomie qui m'a été laissée. Merci de m'avoir fait découvrir et partager tes connaissances dans le domaine de la simulation numérique d'éoliennes. Merci beaucoup pour ton soutien durant les périodes difficiles qu'une thèse peut apporter d'un point de vue professionnel ou personnel. Merci également pour ta bienveillance et tes conseils qui m'ont été très utiles. Pour conclure, merci à ces personnes pour l'encadrement offert m'ayant permis de mener à bien mes travaux de recherche et j'espère que nos chemins professionnels se croiseront à nouveau.

Je tiens par la présente à exprimer ma gratitude envers mes deux rapporteurs, Nathalie Bartoli et Valérie Monbet pour m'avoir fait l'honneur d'évaluer ce document et leurs pertinentes remarques qui m'ont permis d'améliorer mon travail. Je remercie également les autres membres du jury, qui ont accepté de s'intéresser à mes travaux de recherche et le temps qu'ils m'accordent.

Je souhaite maintenant remercier l'ensemble des personnes de l'IFP Energies Nouvelles que j'ai pu côtoyer durant ces dernières années. En particulier, l'équipe du département Mécanique Appliquée de Solaize pour les bons moments passés en leur compagnie et l'atmosphère de travail agréable. J'ai également une pensée pour les différents doctorants d'IFPEN qui ont su être présents lors de ces trois années qui peuvent être parfois difficiles : Constance, Mohammed et Thibaut.

Merci également à l'ensemble de mes « copains Breizhou » qui m'ont permis de survivre à ces derniers mois de thèse ! Merci à la famille Migeois pour votre soutien sans faille et ces week-ends montagnards qui m'ont permis de décompresser en creusant quelques tranchées ou en déménageant vos affaires. Hermine, merci pour notre voyage dans le sud avant le début de la rédaction de mon manuscrit cela m'a permis de recharger mes batteries avant de commencer cette dure épreuve ! Je n'oublierai pas ton aide de relecture cruciale qui sera récompensée dès que les restaurants auront ré-ouverts... Merci également Maud, Alexandre, et Lorraine pour ce dernier confinement ensemble, j'espère que vous n'êtes pas dégoutés de Jean-Luc Lahaye depuis... Merci à l'ensemble des membres de ce

groupe pour votre amitié sans faille !

Pour terminer, je tiens à adresser un grand merci à ma famille et en particulier à mes parents pour croire en moi et m'avoir soutenu jusqu'à aujourd'hui. Je pense que ces travaux de recherche n'auraient pas été possibles sans leur présence et leur encouragement. Mes pensées vont également à ma sœur et mon frère qui chacun à leur manière m'ont aidé durant cette période.

Contents

List of Figures

List of Tables

Overview

Ce chapitre présente le contexte dans lequel s'inscrivent les travaux de recherche menés lors de cette thèse et les solutions proposées pour répondre aux objectifs sous-jacents. Les incertitudes sont omniprésentes en particulier dans le domaine de la simulation numérique d'éolienne. La prise en compte de ces incertitudes et l'estimation de leur influence lors de l'analyse de structures en opération sont devenues des enjeux cruciaux. Le but de cette thèse est de proposer une procédure de quantification et de réduction des incertitudes affectant un modèle numérique aéro-servo-élastique d'une éolienne terrestre. Ce manuscrit est divisé en trois parties : la Partie I présente les principaux concepts pour la quantification d'incertitudes et la modélisation d'une éolienne ; la Partie II développe les différents outils méthodologiques pour l'analyse de sensibilité, la métamodélisation et la calibration utilisés dans la suite de cette thèse ; la Partie III présente les contributions applicatives de la thèse dans le domaine de la modélisation numérique d'éolienne où les simulations numériques sont coûteuses et stochastiques. Pour terminer, le dernier chapitre de ce manuscrit fait le point sur l'ensemble du travail de thèse et esquisse quelques perspectives.

Motivation

The last years of the 20th century were a turning point for awareness of the population on the use of fossil energy and its terrible impact on climate change. In that context, political decisions have been recently taken with regard to a reduction of greenhouse gas emissions and a sustainable mix of energy resources, such as for example the Kyoto agreement in 1997 or more recently the Paris agreement signed in 2015. The countries, which have ratified the Paris agreement, agreed to prevent irreversible damage from climate change by limiting global warming to 2°C compared to pre-industrial times. In this context, the long-term objective is to drastically reduce greenhouse gas emission by 80-95% by 2050. According to researchers, at least 32% of total energy consumption must stem from renewable energy sources to reach this target. Among the different renewable energy technologies, wind and solar ones have increased their production performances thanks to innovative breakthroughs. As a result, these technologies have experienced a rapid growth in the last decades. As an example, France's well adapted climatic conditions have led to a growing development of the onshore wind energy, as depicted in Figure 1. This continuous growth allows the creation of an industry working on the design, the operation and the maintenance of wind turbines.

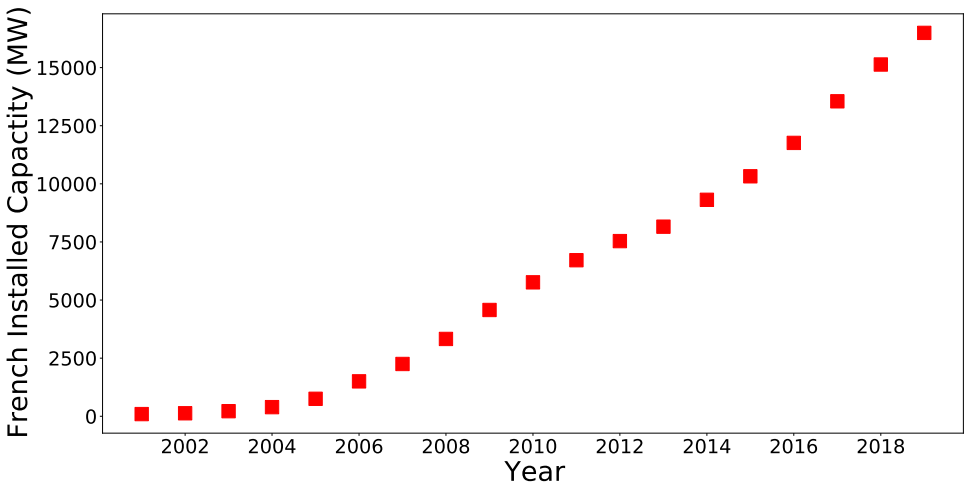


Figure 1 – French installed capacity development of onshore wind power production between 2001 and 2019 (source : RTE, Bilan électrique 2019).

The design procedures to testify the reliability of the wind turbine structures are given by standards. They prescribe various operating and environmental conditions that have to be taken into account when planning the development of such systems. These design procedures assess the structural integrity of the components by using aero-servo-elastic codes and by considering the probability of each environmental and operational condition. Nevertheless, operating wind turbines experience real external solicitations and operational conditions that can be different from the envelope defined in the design standard prescriptions. Moreover, the dynamic response of the structure and its lifetime can be affected by some uncertainties or evolution in the wind turbine properties. Consequently, it is mandatory to estimate the remaining life of an operating wind turbine by taking into account all the inherent uncertainties. In that context, the quantification and reduction of uncertainties involved in the aero-servo-elastic numerical

models used to determine the effective fatigue loads of the turbine have to be properly performed. Overall, aero-servo-elastic numerical simulations involve many uncertainties in the parameters of the wind turbine numerical model as well as in the external sollicitations. Thus, a first problematic of this research work is devoted to the quantification of these uncertainties on the variability of the model responses and can be formulated as : How to quantify the impact of the uncertainties affecting the input parameters of an aero-servo-elastic numerical model on the variability of the quantities of interest used in the estimation of the components' lifetime ?

When the uncertainties are quantified, a major challenge is to determine a procedure to reduce them. Recent decades have been marked by a simultaneous development of sensor technologies and internet of things capabilities allowing real-time monitoring of wind turbines. Nevertheless, current solutions do not take full advantage of the large amounts of data provided by sensors placed on modern operating wind turbines. In a context of digitalization of the industry, the exploitation of these collected observations can be optimized with numerical modeling technologies in order to refine the predictions of production, the estimation of remaining lifetime of the structure, and the planning of maintenance. The underlying industrial concept is the idea of a digital twin filling the gap between the modeling and the measurements. In our study, this strategy relies on a recursive procedure for parameter uncertainty reduction. This procedure is based on an adaptive update of model calibration from real-time observations, leading to a highly reliable model. We can formulate a second problematic as : How to recursively reduce the uncertainties affecting the influential parameters in terms of variability of the quantities of interest obtained from a wind turbine numerical model based on in situ observations ?

Approach of this thesis

In order to answer the first problematic mentioned previously, a global sensitivity analysis (GSA) methodology is investigated. Such statistical approaches focus on the investigation of how the uncertainty of a model can be allocated to the one in the input parameters. The main disadvantage of these techniques, which often refer to the probabilistic framework and Monte Carlo (MC) methods, is the requirement of a high number of calls to the numerical model. Among the different sensitivity measures, we focus on variance-based sensitivity measures called Sobol' indices. In this kind of analysis, the unknown input parameters are considered as independent random variables gathered into a random vector and characterized by their probability distribution. In many applications, the probability distributions of input parameters are often determined using expert knowledge. Nevertheless, variance-based GSA in the context of aero-servo-elastic wind turbines is a challenge because we are facing a time-consuming stochastic simulator. To take into account the inherent variability due to the stochastic simulator, we propose to model the mean of the output of interest with a Gaussian process with heteroscedastic noise. Then, this surrogate model is used to perform a sensitivity analysis based on usual MC estimation procedure. By performing such sensitivity analysis, we are able to determine on which important input parameters the calibration effort has to be made and which non influential ones can be fixed to nominal values without affecting model prediction accuracy.

Concerning the second question, recursive Bayesian inference can be used to reduce

the uncertainties in a wind turbine numerical model. This inverse problem approach uses some measurements to recursively update some prior probability distributions of the model parameters through the Bayes' theorem. Bayesian framework enables to tackle ill-posed problems where some (or all) targeted parameters cannot be identified based on available data. Nevertheless in this study, an approach involving a GSA based on total Sobol' indices is firstly performed to analyze the input parameter identifiability from in situ measurements. For the recursive inference problem, we propose to rely on an ensemble-based filtering method which is a Monte Carlo variant of the Kalman filter. This data assimilation procedure has been widely used to estimate the state of a given model based on partial observations. A major challenge in such kind of inference problems is when the underlying dynamic behavior of the state is not explicitly known. In our wind turbine application, we are facing such an issue when we are looking at the inference of the parameters related to the wind inflow solicitation. In this study, we rely on the substitution of the unknown dynamical model involved in the state-space formulation of the data assimilation problem by a data-driven representation of the state dynamics based on an analog forecasting strategy. The recursive inference procedure is then performed thanks to a so-called Analog Ensemble Kalman Filtering (AnEnKF) scheme.

Outline of the manuscript

This manuscript is divided into three parts and is organized as follows :

Part I :

- In Chapter 1, a non-exhaustive state of the art of concepts and methods used for uncertainty quantification is given. Section 1.1 describes some of the most common methods used for propagation of uncertainty. In Section 1.2, we give an overview of the sensitivity analysis procedures. Then, Section 1.3 addresses the procedure of metamodeling to emulate a time-consuming numerical model. Model calibration, also known as uncertainty reduction, is explained in Section 1.4. Lastly in Section 1.5, the approach of uncertainty quantification techniques in wind turbine applications is detailed.
- Chapter 2 details the different modeling aspects that have to be considered in aero-servo-elastic simulations. Section 2.1 details the modeling of turbulent full field winds. In section 2.2, we present the widely used blade-element momentum theory allowing to estimate the loads on wind turbine blades due to wind solicitation. The basics of a wind turbine control strategy are described in Section 2.3. The aero-servo-elastic dynamic analysis is detailed in Section 2.4. A brief description of fatigue analysis used to estimate the accumulated damage that the structure faces during its lifetime is given in Section 2.5.

Part II :

— Chapter 3 is devoted to variance-based global sensitivity analysis with a specific regard to Sobol' indices. The functional analysis of variance decomposition for deterministic models is presented in Section 3.1 . Sections 3.2 and 3.3 respectively define Sobol' indices and the associated estimation procedures in the framework of independent input parameters. Section 3.4 is devoted to the procedure to define and to estimate aggregated Sobol' indices when dealing with a multivariate output. Lastly in Section 3.5, we present a space-filling strategy called Latin Hypercube Sampling (LHS).

— Chapter 4 presents the concept of metamodeling based on Gaussian process

regression and its application in the estimation of Sobol' indices. Section 4.1 explains theoretically the metamodeling procedure using Gaussian process regression. Then, in Section 4.2 we present a kriging-based sensitivity analysis relying on the estimation of Sobol' indices. Finally in Section 4.3, Gaussian process metamodel framework with heteroscedastic noise is proposed in order to estimate Sobol' indices in the context of a stochastic numerical model.

— Chapter 5 is related to the model calibration formulation. In Section 5.1, we give an overview of the Bayesian inference paradigm. In Section 5.2 we present two data assimilation strategies : the Kalman filter and its Monte Carlo approximation named ensemble Kalman filter. Section 5.3 is dedicated to the extension of ensemble Kalman filter for model calibration.

Part III :

— Chapter 6 proposes a complete framework for quantifying and reducing the uncertainties affecting the properties of an aero-servo-elastic wind turbine numerical model. The global sensitivity analysis methodology in the context of stochastic time-consuming numerical model is introduced in Section 6.1. In Section 6.2, we explore how the ensemble Kalman filter can be employed in model calibration problems. Section 6.3 is devoted to present the wind turbine numerical model, its uncertain input parameters and the selected output quantities used for quantifying and reducing the uncertainties. In Section 6.4, a wind turbine numerical case study is used to illustrate the proposed framework and its capabilities in calibrating parameters with noisy pseudo-experimental output data.

— Chapter 7 extends the previous framework to the uncertainties affecting the wind field by relying on a data-driven data assimilation method. Section 7.1 describes the different uncertainties considered in this study. In Section 7.2, the theoretical framework of data-driven data assimilation is detailed with a specific interest in the ensemble Kalman filtering scheme coupled with the analog forecasting strategy. Finally, results of the application of this complete procedure of uncertainty quantification and reduction to a wind turbine model are presented in Section 7.3.

Finally, the last chapter exposes some conclusions on this thesis work and draws some perspectives for future research work.

Première partie

Introduction

Il n'est pas certain que tout soit incertain.

Blaise Pascal

1 Uncertainty Quantification

Ce chapitre présente quelques concepts et méthodes associés à la quantification des incertitudes. Ce domaine regroupe de nombreux outils statistiques, tels que la propagation des incertitudes, l'analyse de sensibilité, la construction d'un modèle de substitution et la calibration de modèle. Dans la Section 1.1, nous aborderons le principe de la propagation des incertitudes à travers un modèle numérique. Puis dans la Section 1.2, nous détaillerons quelques méthodes d'analyse de sensibilité. Ces méthodes sont des outils utiles lors de l'exploitation de modèles numériques permettant notamment de caractériser quels sont les paramètres d'entrée du modèle qui contribuent le plus à la variabilité d'une quantité d'intérêt obtenue en sortie, quels sont ceux qui n'ont pas d'influence et quels sont ceux qui interagissent entre eux. Néanmoins, ces approches mathématiques présentent une contrainte computationnelle élevée. Ainsi, la Section 1.3 présente l'approche qui consiste à ajuster un modèle de substitution pour s'affranchir du coût de calcul prohibitif du modèle numérique initial. Le modèle de substitution est ajusté à partir d'un nombre d'appels limité au modèle numérique initial. Dans la Section 1.4, nous nous concentrons sur la description de méthodes de calibration permettant de mettre à jour la valeur des paramètres d'entrée à partir de mesures in situ. Ces approches pour la calibration de code numérique couplées au développement des technologies de capteurs permettent d'envisager la mise à jour en continu des modèles numériques grâce aux mesures effectuées sur le système en fonctionnement. Dans ce contexte, les méthodes provenant du domaine de l'assimilation de données peuvent apporter une réponse. En particulier, les techniques de filtrage qui ont la faculté d'être facilement exécutables en parallèle, ce qui est d'un grand intérêt pour les codes numériques coûteux en temps de calcul que nous introduirons au Chapitre 5. Pour terminer dans la Section 1.5, nous proposons une revue bibliographique des méthodes de quantification des incertitudes dans le domaine de l'énergie éolienne.

Introduction

During the last decades, numerical models have been widely used to substitute physical experiments which are considered costly and difficult to perform [Santner et al., 2003]. In our industrial context, a numerical model, also known as a simulator, includes mathematical equations describing the physics of the system, discretization techniques, and algorithms used to solve the discretized equations. Nevertheless, in practice, such a simulator never fully reproduces the real phenomena of interest. Indeed, they always rely upon physical simplifications, numerical approximations, and the estimation of the model input parameters can be inaccurate, i.e., uncertain. Consequently, one has to assess the impact of these uncertainties on the model response, in particular the ones due to the input parameters. In the literature, two categories of uncertainties are considered. On the one hand, the aleatoric uncertainties, also called statistical uncertainties, representing the inherent variability of the inputs, e.g., intrinsic random fluctuations of some environmental variables. On the other hand, the epistemic uncertainties which result from a lack of knowledge on the input parameters [Walker et al., 2003]. In many cases, both types of uncertainties may be present in the system of interest [Smith, 2013, Soize, 2017] and it is challenging to describe and quantify the uncertainty [Der Kiureghian and Ditlevsen, 2009]. Hereafter, we will focus on the epistemic uncertainty which refers to a lack of knowledge and can thus be considered as reducible.

Uncertainty quantification (UQ) aims at taking into account uncertainties affecting input parameters of numerical models and studying their impact onto the model response [De Rocquigny et al., 2008, Smith, 2013, Ghanem et al., 2017]. The UQ field entails many statistical tools, such as, uncertainty propagation, surrogate modeling or parameter inference [Sullivan, 2015]. Throughout this chapter, we will suppose that the physical system of interest is represented by a model denoted by f and defined by the deterministic function:

$$f : \begin{cases} \mathcal{P} \subset \mathbb{R}^p & \to \mathbb{R} \\ \mathbf{x} = (x_1, ..., x_p)^T & \longmapsto y = f(\mathbf{x}) \end{cases}, \tag{1.1}$$

where, the input parameters of the model are gathered into a vector denoted $\mathbf{x}$ and the model response scalar output (also referred to as quantity of interest in the rest of this chapter) is y.

The input parameters usually represent physical constants which describe the mathematical formulation of the system of interest. Nevertheless, in many simulators, one can also consider in this input parameter vector, some tuning parameters, which have no physical interpretation. They have to be properly defined by the user in order to make the numerical model mimics the underlined physical phenomena. As mentioned by Sudret [2007], in the field of mechanical engineering, such as wind turbine modeling, the input parameters of the simulator can encompass:

- parameters modeling the loading of the system of interest;
- parameters specifying the geometry of the system, such as thickness, length, cross-sections, etc.;
- parameters describing the material constitutive laws, such as stiffness parameter, damping's ratios, Young's modulus, etc.

At the opposite, the quantities of interest obtained from the numerical model are usually:

- displacements (or velocities), e.g., vector of nodal displacements (or velocities) in

the context of finite element simulation;

- damage equivalent load, electricity production, rotor velocity, temperature, etc., in case of wind turbine analysis;
- stress quantities, such as stress intensity, equivalent Von Mises stresses, etc.

By taking the formalism described by Baudin et al. [2017], four main steps can be considered to perform UQ analysis and are summarized in Figure 1.1.

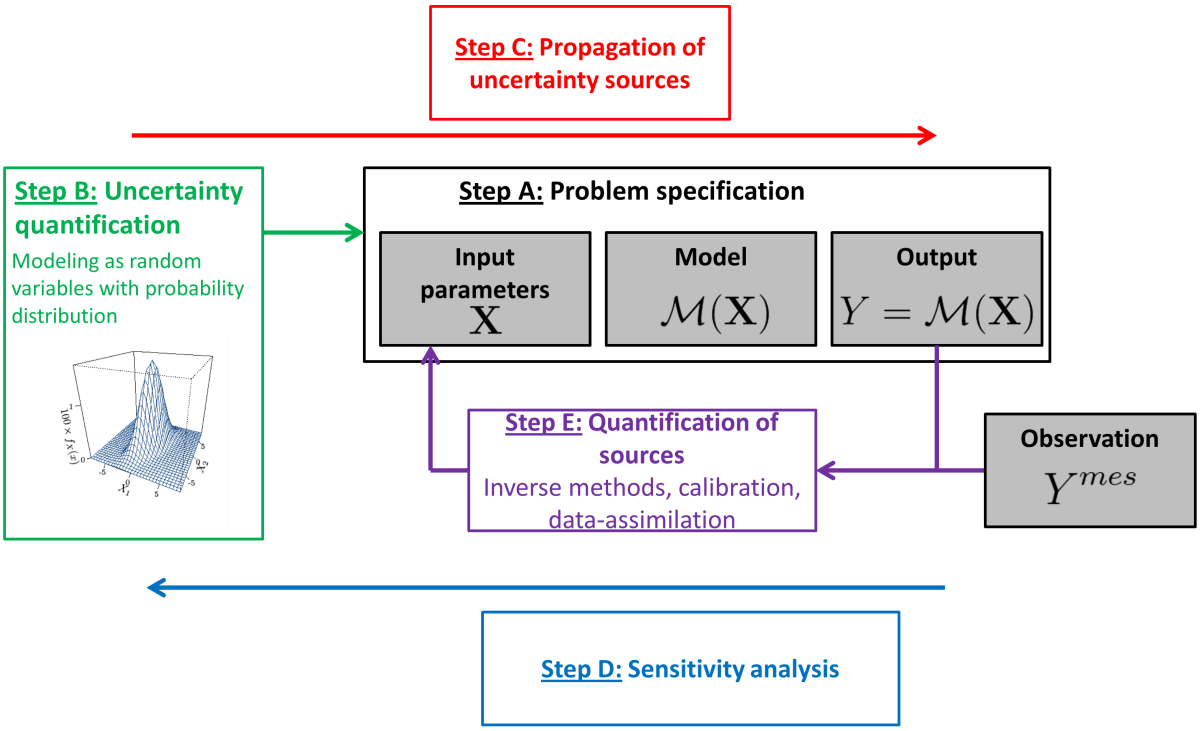


Figure 1.1 – Sketch representing the different steps that compose an uncertainty quantification procedure adapted from [Baudin et al., 2017].

According to the authors, these main steps can be detailed as follows:

Step A consists in defining the model and identifying the quantity of interest but also the input parameters that should be used to assess the physical system of interest.

Step B consists in quantifying all the sources of uncertainties by determining the input parameters considered as poorly or not known. During this step, their variability are established by modeling them as a random vector with distribution selected by statistical fitting or expert knowledge.

Step C consists in propagating uncertainties from the input parameters through the numerical model in order to estimate the distribution of the quantity of interest defined in Step A.

Step D consists in analyzing the sensitivity of the different random input parameters on the randomness of the quantity of interest. This step allows to hierarchize the uncertainty sources and is known as sensitivity analysis (SA).

Step E aims to estimate the values or the posterior probability distribution of the unknown input parameters by employing methods that use real observations of the model response.

In this chapter, we propose a brief state-of-the-art of the different strategies used for uncertainty quantification analysis. In Section 1.1, some of the methods for propagation of uncertainty existing in the literature are formulated. Then, in Section 1.2 a review of sensitivity analysis procedures is presented. Section 1.3 proposes to address the subject of metamodeling to approximate a time-consuming numerical model. Uncertainty reduction, also known as model calibration, is explained in Section 1.4. Lastly in Section 1.5, we propose a review of uncertainty quantification techniques for wind turbine applications.

1.1 Uncertainty Propagation

Uncertainty Propagation (UP) attempts to pass on uncertainties from the input parameters throughout the numerical model of interest, also known as the simulator [Ghanem et al., 2017]. These uncertainties can be described by a probability distribution function (pdf). Then propagating input parameter uncertainties through the numerical model consists in determining the pdf of the model response. Let us suppose that such probabilistic description of the input parameters has been defined from Step B, see Figure 1.1, in terms of a random vector denoted by $\mathbf{X} \in \mathcal{P} \subset \mathbb{R}^p$ with mutually independent components. Then, the random model response vector is defined as:

$$Y = f(\mathbf{X}) \cdot$$

As mentioned previously, the purpose of UP is to investigate the probabilistic content of the random model response Y by studying its probability distribution function. Figure 1.2 is a graphical representation of the classical UP procedure. Nevertheless, computing the model response distribution is most of the time not a straightforward task. Indeed, such pdf depends on the joint pdf of the unknown input parameters and on the model functional representation f. Consequently, methods to perform UP have been developed, [see Lee and Chen, 2009].

In the literrature, a classification of methods for UP is proposed as follow, [see Sudret and Der Kiureghian, 2000] :

- *second moment* methods provide an estimate of the response variability by giving the first two statistical moments of the quantity of interest, i.e., the mean value μ_Y, and the standard deviation σ_Y;
- *spectral* methods deal with the complete pdf by doing an expansion of the quantity of interest onto a suitable basis;
- *structural reliability* methods investigate the tails of the pdf of the response of interest. These methods rely on the computation of the probability of exceeding a predefined threshold.

We focus our interest in this section on second moment methods, i.e., methods allowing to investigate the mean and standard deviation values of a quantity of interest $Y = f(\mathbf{X})$, such as :

$$\mu_Y = \mathbb{E}_{\mathbf{X}}[f(\mathbf{X})] = \int_{\mathcal{P}} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \,,$$

and the variance,

$$\sigma_Y^2 = Var_{\mathbf{X}}[f(\mathbf{X})] = \mathbb{E}_{\mathbf{X}}[f(\mathbf{X})^2] - \mathbb{E}_{\mathbf{X}}[f(\mathbf{X})]^2 \cdot$$

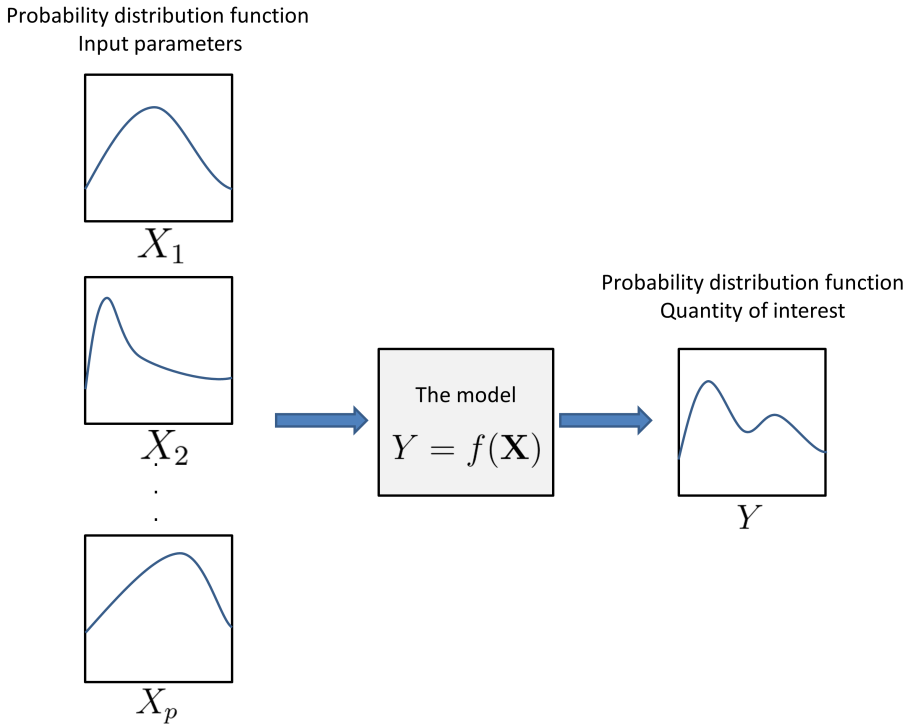


Figure 1.2 – Graphical representation of the uncertainty propagation procedure.

In this context, Monte Carlo (MC) simulation can be used to estimate respectively the mean value and the variance previously defined. MC approach is a sampling method, first introduced by Von Neumann and Ulam [1951], which relies on the law of large numbers. Assuming that a s sample of input parameter vector has been constructed, denoted by $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(s)}\}$, the estimators of those two first statistical moments are defined as:

$$\widehat{\mu_Y} = \frac{1}{s}\sum_{i=1}^{s} f(\mathbf{x}^{(i)})\,,$$
$$\widehat{\sigma_Y^2} = \frac{1}{s-1}\sum_{i=1}^{s}\left(f(\mathbf{x}^{(i)}) - \widehat{\mu}_Y\right)\,.$$

This sampling strategy is widely employed in the probabilistic UP framework due mainly to its non-intrusive property. Nevertheless, a major drawback is its convergence rate in $o(s^{-1/2})$ which makes it infeasible when we are dealing with computationally expensive numerical model. In order to outperform this convergence rate, different methodologies based on variance reduction techniques have been proposed [Rubinstein and Kroese, 2016], e.g., importance sampling [Hastings, 1970, Gilks and Berzuini, 2001, Robert and Casella, 2004], quasi-Monte Carlo sampling [Niederreiter, 1978, Caflisch, 1998] or Latin Hypercube Sampling (LHS) see Section 3.5. Another well-known approach consists in approximating the time-consuming numerical code by a reduced model, called a metamodel (also known as a surrogate model), which can be simulated with acceptable calculation time, see Section 1.3.

1.2 Sensitivity Analysis

Sensitivity Analysis (SA) aims to determine the contribution of different sources of uncertainty in the inputs on the output obtained from a complex system [Saltelli et al., 2004]. In the SA literature, two major procedures are presented: local and global SA methods.

Local SA methods quantify the impact of an input parameter on the numerical model around a nominal value. In other words, local SA methods indicate how fast the output increases or decreases locally around given values of the input parameters. Such kind of SA is mainly done with partial derivatives. Let us assume that the response is modeled by the mathematical function $f : \mathcal{P} \subset \mathbb{R}^p \to \mathbb{R}$, described in Equation (1.1). The first-order local sensitivity index of the i-th parameter is defined as the partial derivative of the output of interest with respect to the input parameter, at a given nominal value $\mathbf{x}^0$:

$$S_i^{local} = \left. \frac{\partial f(\mathbf{x})}{\partial x_i} \right|_{\mathbf{x}=\mathbf{x}^0}.$$

Such so-called local techniques rely mainly on the computation of gradient of the quantity of interest (QoI) with respect to its parameters around a given nominal value. In order to estimate efficiently this gradient, numerous techniques have been proposed such as adjoint differentiation methods or finite-difference approximation methods [Cacuci, 2003]. A major drawback of this definition of sensitivity lies in its local property. Indeed, if the function of interest f is highly nonlinear with respect to x_i, then the computed partial derivative will vary depending on the considered nominal value. In other words, derivatives are only informative at the point where they are estimated. Nevertheless, local approaches are commonly employed when dealing with industrial numerical model with a large number of parameters.

At the opposite, global methods consist in studying the impact of the input parameter variation on the variability of some output by considering their whole variation space [Saltelli et al., 2007]. Three different techniques can be considered in order to perform global sensitivity analysis (GSA): screening approaches, methods measuring the effect of a parameter distibution on the output distribution, and other measures. In order to properly choose the most suitable GSA method, Saltelli et al. [2004] emphasize the need to specify the objectives of the study and propose the following ones:

— factor prioritization: identification of the input parameters for which uncertainty reduction would allow to obtain the greatest reduction of the uncertainty impacting the quantity of interest;

— factor fixing: identification of the non-influential parameters in order to set them to nominal values without influencing the quantity of interest obtained from the numerical model;

— factor mapping: mapping the output behavior in function of the input parameters by focusing on a specific domain of parameters if necessary;

— variance cutting: determination of the input parameters to be fixed in order to obtain a given tolerance in the uncertainty affecting the quantity of interest.

1.2.1 Screening methods

Firstly, the screening methods are based on the amplitude of the variations of the QoI obtained considering different input parameter values. They have been developed in order to quickly explore the variability of a QoI induced by the variation of each input parameter. Such characterization of the importance of each parameter on a quantity of interest can be performed by computing the derivatives of the output with respect to an input and considering the others as constant.

In the literature, such characterization can be done by using elementary effects mainly developed in Morris method [Morris, 1991]. Hereafter, let us consider the model f with p independent input parameters $\mathbf{x} = (x_1, \ldots, x_p)^T$ varying in the p-dimensional unit cube, $\mathcal{P} = [0,1]^p$. Then, the input space $\mathcal{P}$ is discretized into a q-level grid, such as $D = [0, \frac{1}{q-1}, \frac{2}{q-1}, \ldots, 1]^p$. The elementary effect of input parameter X_i, with $i \in \{1, \ldots, p\}$, is hereafter denoted by EE_{X_i} and defined as follows:

$$EE_{X_i}^{(k)} = \frac{f(\mathbf{x}^{(k)} + e_i\delta) - f(\mathbf{x}^{(k)})}{e_i\delta}\,, \tag{1.2}$$

where δ, known as the grid jump, is a value in $\{\frac{1}{q-1}, \ldots, 1 - \frac{1}{q-1}\}$, q is the number of levels, e_i is a vector of the canonical base, and $\mathbf{x} = [x_1^{(k)}, \ldots, x_p^{(k)}]^T$ a randomly selected value on the grid D such as $\mathbf{X} + e_i\delta$ is still in the domain of parameter space.

In a nutshell, Morris screening method consists in moving each parameter one-at-a-time between each model evaluation (one-at-a-time method). The distribution of the effect associated to the i-th input parameter in Equation (1.2) is obtained by randomly sampling N different $\mathbf{x}$ from D which induces to a total number of calls to the function f of $N \times (p+1)$. The sensitivity measures proposed by Morris, are the estimates of the mean and the standard deviation describing the distribution previously mentioned. The estimate of the elementary effect for each input parameter describes its overall influence on the output of interest and is defined as:

$$\hat{\mu}_i = \frac{1}{N}\sum_{k=1}^{N} EE_{X_i}^{(k)}\,.$$

In the context of non-monotonic model, an estimated mean of the elementary effects near zero does not mean that the input parameter has no impact on the output variability. Indeed, in the case of non-monotonic model, a variation of the input parameter value can change the sign of the output and consequently results in a cancellation of its effect as described in Equation (1.2). To circumvent this issue, authors in [Campolongo et al., 2011] propose to estimate the mean of the absolute elementary effect :

$$\hat{\mu}_i^* = \frac{1}{N}\sum_{k=1}^{N} \mid EE_{X_i}^{(k)} \mid\,. \tag{1.3}$$

As said previously, the second index is the estimate of the standard deviation of the elementary effect. This measure gives an indication of the presence of interactions and/or nonlinearity between the i-th input parameter and the other ones. It is defined as:

$$\hat{\sigma}_i = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(EE_{X_i}^{(k)} - \hat{\mu}_i)^2}\,. \tag{1.4}$$

By calculating the indices defined in Equations (1.3) and (1.4), the method can lead to the identification of three types of input parameters: non influential input parameters, influential input parameters with linear effects, and influential input parameters with non-linear and/or interaction effects [Iooss and Lemaître, 2015]. Figure 1.3 gives a graphical representation example of a trajectory design in the 2-dimensional input parameter space with $N = 3$ random samples. The input parameter space is uniformly divided into 9 levels and the grid jump is $\delta = \frac{1}{8}$.

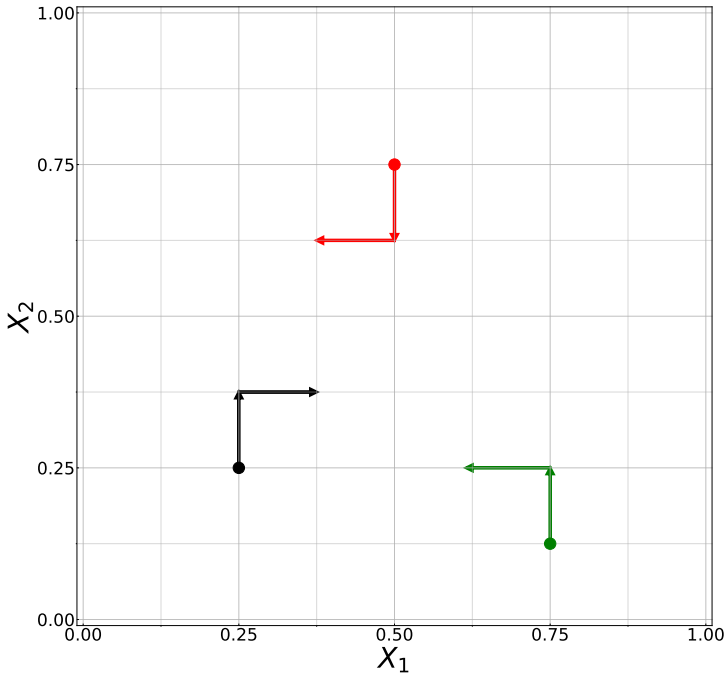


Figure 1.3 – Graphical representation of Morris screening method with $N = 3$, $q = 9$ and $\delta = \frac{1}{8}$. The filled circles are the random nominal sample points from which the random perturbation based on the grid jump is carried out one-at-a-time.

An extension of this method has been proposed by Sobol' and Kucherenko [2009]. The described method, called derivative-based global sensitivity measures (DGSM) allows to quantify any little variation of the quantity of interest due to input parameters by using the second moment of model derivatives as importance measure [Lamboni et al., 2013]. DGSM might require some regularity assumptions on the numerical model and in general it is not recommended for highly nonlinear model [Sobol' and Kucherenko, 2009]. An inequality relation between total Sobol' indices (see Chapter 3) and DGSM has been established by Sobol' and Kucherenko [2009] and further extended for any input parameter having a Boltzmann probability measure in [Lamboni et al., 2013].

1.2.2 Methods measuring the effect of a parameter distibution on the output distribution

In order to classify by order of importance the different parameter influences, other sensitivity methods measuring the effect of a parameter distribution on the output distribution must be used. These approaches are based on sensitivity indices allowing to define a measure of the influence of an input parameter, taken singly or in combination with

others, on the variability of the quantity of interest [Saltelli et al., 2007]. When dealing with nonlinear and non-monotonic model, a classical measure of importance is the variance. Variance-based methods consist in fixing one parameter (or several parameters), and studying the impact on the output variability. The first works based on this idea were done by Hora and Iman, [see Hora and Iman, 1986], with the introduction of a measure of importance for the variable X_i given by:

$$I_i = \sqrt{Var[Y] - \mathbb{E}[Var_{X_i}[Y|X_i = x_i]]}\,. \tag{1.5}$$

Nevertheless, as proven by the authors, the measure in Equation (1.5) is not robust and can be highly influenced by outliers. Thus, Iman and Hora have proposed a new index, [see Iman and Hora, 1990], defined as:

$$\frac{Var_{X_i}[\mathbb{E}[log(Y)|X_i = x_i]]}{Var[log(Y)]}\,,$$

where, $\mathbb{E}[log(Y)|X_i = x_i]$ is estimated by a linear regression. Then, McKay [1997] uses a classical regression model, the decomposition of the total variance, and considers as measure of sensitivity the correlation ratio, defined as:

$$\eta_i^2 = \frac{Var_{X_i}[\mathbb{E}[Y|X_i = x_i]]}{Var[Y]}\,. \tag{1.6}$$

This index is a straightforward measure of the impact of X_i on the output Y. The conditional expectation $\mathbb{E}[Y|X_i = x_i]$ is the best approximation of Y by a function of only X_i [Saltelli et al., 2007]. By considering the variance of this conditional expectation, one can obtain the spread of Y due to the variation of X_i. Then, the ratio presented in Equation (1.6) allows to compare this dispersion with the overall one $Var[Y]$. This index is also known as the first-order sensitivity index and has been generalized by Sobol' based on the functional analysis of the variance (FANOVA), see Chapter 3.

As detailed in Section 3.3, Monte Carlo sampling based on pick-freeze procedures can be used in order to estimate Sobol' indices [Homma and Saltelli, 1996, Sobol', 2001, Saltelli, 2002]. To reduce the computational cost, as proposed by Tissot and Prieur [2015], one can rely on a pick-freeze scheme based on replicated sampling. Cukier et al. [1978] propose another strategy which relies on a multi-dimensional Fourier transform, called the Fast Amplitude Sensitivity Test (FAST). Later, Tarantola et al. [2006] have coupled the FAST technique with a random balance design. The previous mentioned estimation procedures are based on structured sample designs. To overcome sampling constraints, estimation procedures based on given data sets have been developed [Plischke, 2010, Plischke et al., 2013].

In some applications, variance-based methods are not useful to properly estimate the global effect of parameters. Indeed, variance can sometimes poorly represents the variability of the output distribution. In this context, Borgonovo [2007] proposes a new sensitivity measure which allows to analyze the impact of parameter uncertainty on the overall output distribution. This moment independent sensitivity measure has been further developed by Da Veiga [2015], based on measures of dissimilarity and dependence. Other moment independent measures have been proposed in the literature, such as the entropy-based sensitivity indices [Krzykacz-Hausmann, 2001, Liu et al., 2006, Auder and Iooss, 2008].

1.2.3 Other measures

Another approach in GSA consists in fitting a linear model in order to explain the behavior of Y given the values of the parameters $\mathbf{X}$ and then to compute sensitivity measures such as Pearson correlation or partial correlation coefficients. These methods based on the analysis of linear models are often referred to as the so-called sampling-based global sensitivity analysis method [Helton and Davis, 2003]. However, they require linear and/or monotony assumptions. These hypotheses have to be validated thanks to statistical techniques, such as the predictivity coefficient and the coefficient of determination [Iooss and Lemaître, 2015]. Nevertheless, they often limit the application of the sampling-based GSA approaches.

SA for dependent input parameters have been also investigated. In this context, one can rely on Shapley effects [Owen and Prieur, 2017]. This sensitivity measure, described in [Owen, 2014], is based on the cooperative game theory concept of Shapley value [Shapley, 1953].

1.3 Surrogate modeling

The large majority of uncertainty quantification procedures require a large number of calls to the numerical model, which quickly goes beyond the limits of available resources when you are dealing with long-running computational code [Iooss et al., 2010, Iooss and Saltelli, 2016, Lamboni et al., 2011, Le Maître and Knio, 2010, Saltelli et al., 2007, Storlie et al., 2009]. Consequently, to perform such analyses, the time-consuming numerical model has to be replaced by a mathematical approximation (often referred to as metamodel) which relies on an acceptable number of output simulations. A metamodel, also known as surrogate model, is a generalization of the response surface methodology proposed by Box and Draper [1987].

In uncertainty quantification, several surrogate models are commonly used, such as reduced bases [Janon et al., 2013] and reference therein, polynomial chaos expansion [Sudret, 2008], neural networks [Alam et al., 2004, Fang et al., 2005] or Gaussian process regression, see Section 4.1. Some of them allow to obtain Sobol' index analytical expressions. In other words, by using some metamodels, we can directly estimate the sensitivity indices without any additional cost.

Sudret [2008] has proven that Sobol' indices are obtained as a direct byproduct of the polynomial chaos (PC) decomposition. PC methodology consists in building a polynomial response surface to model the dependency of the output as a function of the uncertain input parameters. Assuming that the output variable has a finite variance and that the parameters are independent [Soize and Ghanem, 2004]:

$$Y = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} a_{\boldsymbol{\alpha}} \, \Psi_{\boldsymbol{\alpha}}(\mathbf{X}) \, ,$$

where $\Psi_{\boldsymbol{\alpha}}(\mathbf{X})$ are multivariate orthonormal polynomials defined according to the distribution of the input parameters and $a_{\boldsymbol{\alpha}}$ are coefficients to be estimated. PC metamodeling has been used in many fields, such as structural mechanics [Dubreuil et al., 2014, Berveiller, 2005], wildfires [Rochoux et al., 2014], hydraulics [Liang et al., 2008] or hydrology Deman et al. [2015].

The formulation of the Gaussian process (GP) method provides also analytical formulae for the Sobol' indices [Chen et al., 2004, Oakley and O'Hagan, 2004]. GP method is related to the kriging approach in geostatistics [Krige, 1951], developed for spatial interpolation. Then, Sacks et al. [1989] applied this method to numerical models by considering the correlation between two simulated responses depending on the distance between input parameters. GP model treats the numerical model response as a realization of a Gaussian stochastic process characterized by its mean and covariance functions, see Section 4.1.

These two metamodeling approaches have been widely used in the uncertainty quantification field [Soize and Ghanem, 2004, Le Maître et al., 2004, Choi et al., 2004, Le Gratiet et al., 2017, Lockwood and Anitescu, 2012, Marrel et al., 2015b]. They have been compared for sensitivity analysis studies [Le Gratiet et al., 2017, Owen et al., 2017, Schöbi et al., 2015]. In this work, we will focus on Gaussian process regression metamodeling to approximate the behavior of the numerical model of interest. Indeed we are interested in the GP metamodeling principle which allows a direct estimation of the predictive error, useful when estimating Sobol' indices [Oakley and O'Hagan, 2004, Marrel et al., 2009, Le Gratiet et al., 2013]. In Chapter 4, we briefly explain the Gaussian process metamodeling approach and its use for Sobol' index estimation.

1.4 Model Calibration – Uncertainty reduction

Model calibration aims to answer the question of how measured data can inform about the model input parameters and reduce their uncertainty. Model calibration is also known in the literature as inverse problems [Tarantola, 2005]. These problems aim to infer the most likely combination(s) of parameters which cannot be observed directly by using measurements. In this scientific field, two paradigms are facing [Smith, 2013]. Firstly, there are the frequentist methods in which parameters are assumed to be deterministic but unknown. An estimation of these unknown parameters can be done through a statistical minimization of the error between the measurements and model predictions. Nevertheless, as highlighted by Hadamard, Jacques [1902], these inverse problems can be considered as ill-posed, i.e., problems where some (or all) targeted parameters cannot be identified based on available data. To overcome this ill-conditioned aspect, one can rely on Bayesian approaches which allow to model the uncertainty associated to the inferred set of parameters by a probability distribution [Muto and Beck, 2008]. These calibration techniques use measurements to update some prior probability distribution to a posterior one [Kennedy and O'Hagan, 2001, Tarantola, 2005]. These methods are based on Bayes' theorem which can be written as following:

$$\underbrace{p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}_{obs})}_{\text{posterior}} = \frac{\overbrace{p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}_{obs}|\mathbf{x})}^{likelihood} . \overbrace{p_{\mathbf{X}}(\mathbf{x})}^{prior}}{\underbrace{p_{\mathbf{Y}}(\mathbf{y}_{obs})}_{\text{evidence}}},$$

where $\mathbf{y}_{obs}$ is a vector of n noisy observations.

In the literature, Bayesian model calibration has been mainly applied in a batch manner, i.e., offline. These procedures rely on a batch of observations for performing model calibration by typically using Markov chain Monte Carlo methods (MCMC). A well-known strategy among the different methods is the Metropolis-Hastings algorithm

[Hastings, 1970]. It relies on the sequential construction of a Markov chain by applying acceptance-rejection. However, this approach can require thousands of sampling points before convergence of the posterior is achieved. This requirement can quickly be expensive due to the fact that each sampling point corresponds to an evaluation of the numerical model. To overcome this computational burden, a possible solution is to use a surrogate model to replace the numerical model, [see Marzouk et al., 2007, Yan and Zhou, 2019].

Recent decades have been marked by a simultaneous development of sensor technologies and internet of things capabilities. Thus, our research efforts have been mostly directed toward online techniques, i.e., the data are sequentially processed when new observations become available. In this context, the model parameter inference can be carried out using a parameter estimation algorithm based on sequential Bayesian updating techniques. In geosciences, these techniques are called data assimilation methods [Blayo et al., 2014]. These inverse methods have found numerous applications in other fields such as oceanography, weather forecasting, seismology or finance [Ghil and Malanotte-Rizzoli, 1991, Emerick and Reynolds, 2012]. Historically, they have been used to monitor the latent state of a system by combining a model with real observations. The rise and the generalization of these methods have occurred in the middle of the 20-th century with the works of Kalman [Kalman et al., 1960]. These works have led to the Kalman filter, providing the best linear unbiased predictor under linearity assumption of the model and Gaussianity. In the extended Kalman filter the restriction of linearity is relaxed thanks to a linearization of the model and the observation operator. Then, Evensen [2009] introduced a Monte Carlo variant of the Kalman filter, called Ensemble Kalman Filter (EnKF). EnKF, seen as an online Bayesian calibration method, is computationally efficient which is a crucial advantage for real-time monitoring applications. However, it is still based on a Gaussianity assumption of the prior and posterior probability distributions. In some engineering practical cases, the Gaussianity assumption might be inexact. Thus, other online Bayesian inference methods have to be used in order to identify parameters. These approaches are called sequential Monte Carlo or particle filter methods, [see Chopin, 2002]. Nevertheless, such methods suffer from the so-called curse-of-dimensionality, i.e., the required number of particles, in order to approximate the probability distribution, increases with the dimension of the system under study. Some examples of model calibration using a sequential Bayesian inference can be found in the literature, see e.g., Conte et al. [2015] for input parameter inference of nonlinear finite element (FE) models using the extended Kalman filter, iterated Extended Kalman filter, and unscented Kalman filter and Tekieli and Słoński [2013] for an experimental validation of a parameter inference of a FE model of a structure.

1.5 Uncertainty quantification in wind energy

Uncertainty quantification studies of fully aero-servo-elastic numerical models are not frequent. However, a few authors have done some work to identify the sources of uncertainty in wind turbine numerical models. Negro et al. [2014] study which parameters, characterizing the design of support structures and foundations of offshore wind turbine, have to be considered. The authors mainly investigate the uncertainties in the structural model. Sørensen and Toft [2010] deal with uncertainties in material properties constituting wind turbine structural model based on expert knowledge. In the literature the research work was mainly devoted to the identification of sources of uncertainty in the

external solicitations. Dimitrov et al. [2015] use high-frequency wind velocity observations from two test stations in order to deduce a probabilistic model for the wind shear. An overview of uncertainties affecting a wind turbine is given by Veldkamp [2006, 2008], where the author proposes to categorize them in five groups: wind climate, sea climate, aerodynamics, structural model, and material fatigue properties.

After assessing the sources, the objective is to propagate these uncertainties through the wind turbine numerical model. Graf et al. [2016] propose to use a Monte Carlo simulation to propagate the uncertainties affecting five parameters modeling the inflow conditions to the lifetime equivalent fatigue loads of a floating offshore wind turbine. Nevertheless, due to the time-consuming behavior of these models, research efforts have been mainly devoted to the use of surrogate models. Murcia et al. [2018] use polynomial chaos expansion surrogate models in order to propagate the uncertainties affecting the environmental conditions in the context of fatigue estimation and energy production of a wind turbine. In Toft et al. [2016b,a], the authors analyze the influence of uncertain wind solicitation parameters on fatigue loads using a quadratic response surface technique based on a circular central composite design. Morató et al. [2019] fit a Gaussian process regression model, also known as kriging model, to approximate the stresses and moments of an offshore wind turbine and thereby obtain its reliability. In [Teixeira et al., 2017], the authors fit a Gaussian process regression model to highlight the importance of different wind and wave inflow parameters. Abdallah et al. [2019] develop a multi-fidelity surrogate modeling approach based on hierarchical kriging for multiple aero-servo-elastic numerical models of varying complexity in order to simulate the extreme flapwise bending moments at the blade root. Clifton et al. [2013, 2014] propose to use different regression-tree surrogates for modeling the power production and equivalent fatigue loads as a function of wind speed, turbulence intensity and shear exponent.

Sensitivity analyses to determine the most influent input parameters are quite rare in the field of wind energy. Robertson et al. [2019a] propose to estimate elementary effects to perform a global sensitivity analysis of an aero-servo-elastic numerical model. The authors in [Murcia et al., 2018] use polynomial chaos expansion metamodeling to estimate Sobol' indices. Their study shows that the turbulent inflow realization has a major influence on the distribution of equivalent fatigue loads in comparison with the shear coefficient or yaw missalignment. In a similar manner, Rinker [2016] builds a four-dimensional polynomial surface response to estimate Sobol' sensitivity indices of turbine load response to turbulence parameters. Recently, Hübler et al. [2017] propose a hierarchical four-step global sensitivity analysis of offshore wind turbines based on aero-elastic time domain simulations. The authors rely on a quantification of the sources of uncertainty based on expert knowledge in the first step, an one-at-a-time sampling strategy in the second one, a linear regression in the third step, and finally, a variance-based analysis. However in this study, the framework used for performing global sensitivity analysis neglects the inherent stochasticity of the aero-servo-elastic numerical models.

After identification of the most influential parameters on the variability of some output of interest, one can focus on model calibration techniques. Model calibration of fully coupled aero-servo-elastic numerical codes is extremely rare. In the specific context of component models, Van Buren et al. [2013] propose to perform the calibration of a simplified finite element model for a wind turbine blade based on Bayesian inference. The authors first rely on a global sensitivity analysis based on ANOVA to determine the most influential parameters on blade vibrations. Then, these parameters are inferred using

Markov chain Monte Carlo methods. Due to the computational cost of the approach, they propose to approximate the finite element model by a Gaussian regression model. For online inference problems, Kalman filtering has been mainly employed for the estimation of inflow wind speed [Østergaard et al., 2007, Soltani et al., 2013], yaw misalignment [Simley and Pao, 2016], and wind shear [Bottasso et al., 2010]. Nevertheless, these applications rely on the representation of the wind turbine numerical model as a simplified linearized state-space model. These methods are referred to in the literature as digital concepts [Branlard et al., 2020].

In this work, we propose a complete framework to quantify and reduce the uncertainties affecting a wind turbine numerical model. It employs a global sensitivity analysis based on the Sobol' index estimation and a recursive Bayesian inference procedure to reduce the uncertainties. In order to alleviate the computational cost of the index estimation during the sensitivity analysis of the fatigue loads, we propose to replace the aero-servo-elastic time-consuming numerical model by a surrogate. Nevertheless, the major challenge in building such mathematical approximation is the fact that the turbulent wind inflow realization causes variations in the wind turbine model quantity of interest. Hence, we propose to use a noisy heteroscedastic Gaussian process regression model to take into account the variability on the turbine response induced by different turbulent wind fields. Then, the recursive Bayesian inference strategy is based on the ensemble Kalman filter. This sequential data assimilation is computationally efficient with high-performance computing tools which is crucial for online calibration of time-consuming codes, such as aero-servo-elastic wind turbine model. Lastly, the presented methodology is extended to the recursive reduction of the uncertainties affecting the turbulent synthetic wind field parameters by relying on a combination of K-nearest neighbors method with ensemble Kalman filtering approach.

Conclusion

This first chapter introduced some of the main techniques in uncertainty quantification for numerical models. In the context of uncertainty propagation, Monte Carlo approaches are popular, but computationally inefficient when dealing with time-consuming computer codes. Then, an invaluable group of techniques in uncertainty quantification is sensitivity analysis. These techniques allow to determine how the uncertainty in the quantity of interest obtained from a model can be apportioned to different sources of uncertainty in the input parameters. Nevertheless, these techniques can be computationally too demanding when dealing with nonlinear models. Consequently, to perform such uncertainty quantification analyses the model has to be approximated by a metamodel. This mathematical approximation simulates the behavior of the time-consuming computer code within a negligible computational cost. Lastly, model calibration can be used to reduce the uncertainties based on the combination of model predictions and observed data. A relevant calibration technique is called the Bayesian model calibration, which uses measurements to update some prior probability distribution of the parameters to a posterior one. One can use sequential Bayesian updating techniques, where the data are sequentially processed when new observations become available.

In wind energy fields, uncertainties are ubiquitous. There is a growing awareness of taking into account these uncertainties in the wind energy community. However, research efforts have been mainly devoted to the quantification of uncertainties in low-fidelity

numerical models. We propose in our work to quantify and reduce the uncertainties of a high-fidelity wind turbine numerical model. Due to the simultaneous development of sensor technologies and internet of things capabilities, we propose to recursively reduce the uncertainties thanks to data assimilation techniques. The suggested framework is equivalent to the concept of digital twin, which combines measured data from the structure and a numerical model to build a digital equivalent of the real-world wind turbine.

2

Wind turbine modeling

Dans de nombreux secteurs industriels, la mise en place d'études expérimentales s'avère être très chronophage et coûteuse. Ceci est notamment vrai dans le secteur éolien où les structures sont de grandes dimensions et soumises à des chargements complexes. En outre, l'évolution de la puissance de calcul des ordinateurs, associée à la plus grande complexité des systèmes étudiés, a conduit à l'émergence de simulations numériques. Elles reposent sur la modélisation du système physique sous la forme d'équations mathématiques et de leurs résolutions numériques. Ainsi, les simulations numériques complètent l'expérience physique et permettent de réduire les contraintes qui en découlent. Les outils de simulation utilisés dans la conception des éoliennes visent à prédire les chargements dynamiques et la réponse de l'ensemble du système. Dans l'étude d'une éolienne terrestre, les principales contraintes sont issues du caractère rotatif du système et de la sollicitation externe agissant sur ses différents composants ; par exemple la tour, la nacelle et les pales, qui sont alors soumises à des déformations élastiques. Afin de garantir une utilisation optimale et dans des conditions sûres, ces structures sont régulées par des stratégies de contrôle. Dans ce contexte, les outils numériques utilisés pour simuler et concevoir un système éolien sont appelés codes aéro-servo-élastiques. Ils permettent d'avoir un environnement de simulation global permettant de coupler différents modèles physiques décrivant l'aérodynamisme, le contrôle et la dynamique structurelle. Dans ce chapitre, une présentation des différentes modélisations physiques utilisées dans ces outils aéro-servo-élastiques est réalisée. La Section 2.1 décrit la modélisation stochastique d'un champ de vent turbulent synthétique par approche spectrale. Dans la Section 2.2 , la théorie des éléments de pale (Blade-Element Momentum) permettant d'obtenir les chargements aérodynamiques grâce au profil aérodynamique des pales est décrite. La Section 2.3 est quant à elle dédiée à la description de la stratégie de contrôle d'une éolienne. La Section 2.4 s'attarde sur la description de l'analyse aéro-servo-élastique dynamique. Pour terminer, dans la Section 2.5, la modélisation de la fatigue d'une éolienne soumise au chargement dynamique et aléatoire du vent est évoquée.

Introduction

The amount of renewable energies in the global energy production is currently increasing continuously in order to reduce greenhouse gas emissions. In this context of energy transition, wind power generation is developing very rapidly in France and worldwide. In the last decade, the trend has been towards the development of larger wind turbine structures. Indeed, the size of wind turbines is a crucial aspect for economic profitability. A major challenge consists in lowering the cost of wind energy by finding the optimal structural design without affecting its safety. The design standard IEC 61400-1 [IEC, 2005], published by the International Electrotechnical Commission (IEC), provides recommendations for modeling the external conditions and designing the structure, control system and mechanical systems. In particular, it prescribes a set of environmental and operational specifications, gathered in a number of design load case (DLC), in order to ensure the structural integrity over the turbine's entire lifetime.

For design validation, two major categories of limit states have to be properly represented by the DLC simulations. Firstly, the ultimate limit states which allow to estimate the maximal mechanical loads in the turbine's components due to external environmental solicitations and operating conditions. The second category gathers the fatigue limit states, which consider the damage accumulation due to fluctuating loading from environmental solicitations and gravity. Wind turbines are facing a high number of cycles (between $1e7$ and $1e8$ for 20-25 years of operation), and consequently, an accurate investigation has to be performed in order to estimate the damages at several critical locations.

DLC simulations consist first in splitting the domain of variation of the environmental conditions in several bins, in relying on dynamic simulations to determine some quantity of interest, and then in incorporating the probability of occurrence of each bin [Ragan and Manuel, 2007, Freudenreich and Argyriadis, 2008]. The dynamic simulation of the wind turbine system has to take into account all the phenomena that can affect its behavior such as aerodynamics, structural dynamics, and control actions [IEC, 2005, DNV GL, 2010]. These physical effects are mutually influenced and the involved numerical model has to evaluate them in a coupled manner. In the literature, such numerical models are referred to as aero-servo-elastic simulators. They are composed of different sub-models representing all the physics that contribute to the turbine dynamics, such as the external conditions, the aerodynamics, the structural dynamics, and also the wind turbine control strategy.

Most of fatigue DLC correspond to stochastic wind fields that have to be properly generated in order to reflect the real-world conditions and avoid over- or under-designed structures. In this context, the generation of synthetic turbulent wind-inflows is a crucial aspect for wind turbine designs. There are two main models for such synthetic wind solicitations: the Mann turbulence model [Mann, 1998] and the Kaimal spectrum coupled with an exponential spatial coherence method [Kaimal et al., 1972]. These approaches are stochastic spectral methods using an inverse fast Fourier transform to construct the field in the time domain. In our work, we will focus on the stochastic method based on the Kaimal spectrum due to its computationally efficiency and its implementation in the open-source turbulent wind simulator TurbSim developed by the National Renewable Energy Laboratory (NREL) [Jonkman, 2009].

In this chapter, we propose a brief overview of the different modeling aspects that have to be considered in aero-servo-elastic simulations. In Section 2.1, we propose to

focus on the modeling of turbulent full field winds. Section 2.2 details the Blade-Element Momentum (BEM) theory which allows to obtain the loads on turbine blades due to wind solicitation. Section 2.3 describes the basics of a wind turbine control strategy. In Section 2.4, the aero-servo-elastic dynamic analysis is described. Finally, Section 2.5 gives a brief description of fatigue analysis in order to estimate the accumulated damage that the structure is supposed to face during its overall lifetime.

2.1 Modeling of synthetic wind

Wind turbines are subjected to fluctuating loads from the wind. This environmental solicitation is by nature random and in order to characterize it, one can define a statistical distribution by making some assumptions. A common consideration is to assume, for 10-minute periods, that the wind-inflow is an ergodic Gaussian process represented by a mean wind speed $\overline{u}$ and a standard deviation σ_u (usually considered at the hub height of the wind turbine) [Burton et al., 2001]. The 10-minute simulations length is a consequence of the wind spectrum, where the high frequency range refers to the turbulence, as depicted in Figure 2.1.

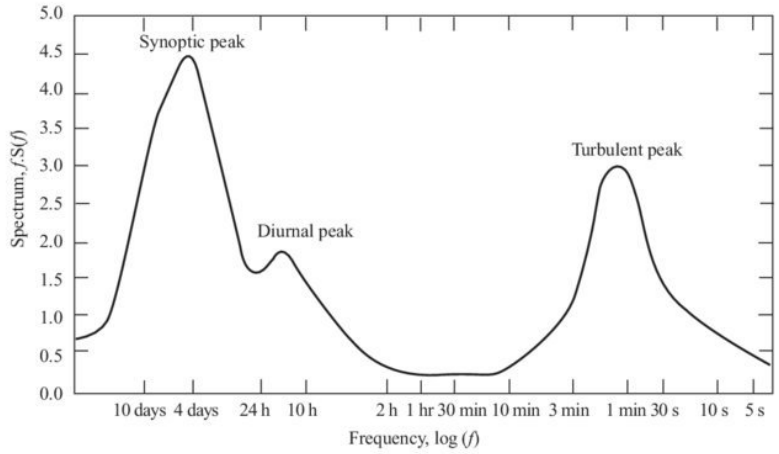


Figure 2.1 – Wind spectrum [Burton et al., 2001]

In wind energy application, the wind speed variations is often referred to as the turbulence intensity defined for the 10-minute period of time as:

$$I = \frac{\sigma_u}{\overline{u}} \cdot$$

Modeling wind field requires the construction of three-dimensional wind velocity vectors that are non-uniform in space and unsteady in time as illustrated in Figure 2.2.

The mean wind speed varies according to the height above the ground level. In IEC standards, two possible distributions respectively named logarithmic and power-law have been proposed to model this evolution due to ground roughness. The power law of the mean wind profile is defined as:

$$\overline{u}(z) = \overline{u} \left(\frac{z}{z_{hub}} \right)^{\alpha}, \tag{2.1}$$

where $\overline{u}$ is the prescribed hub-height mean wind velocity, z is the vertical distance from the ground surface, and z_{hub} is the hub height, and α is the vertical wind shear coefficient.

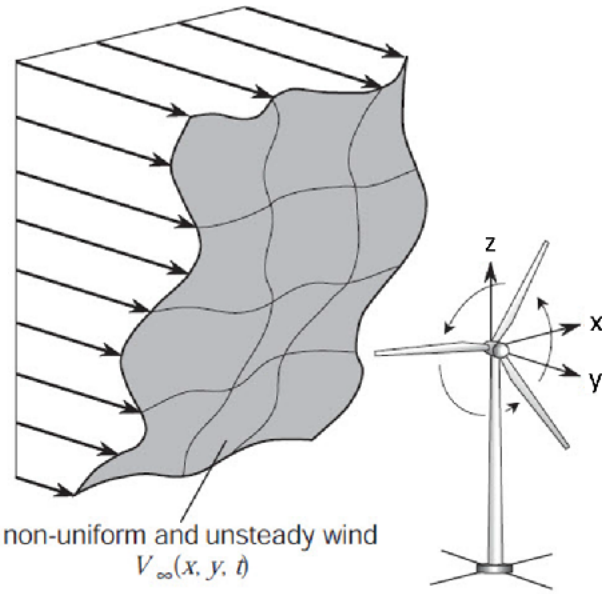


Figure 2.2 – Spatial distribution of the wind inflow upstream of the wind turbine, source [Hasegawa et al., 2004].

The generation of the short-term variations of the wind speed is performed thanks to turbulence spectral models depending on the wind mean speed and the standard deviation. According to the IEC 61400-1 standard [IEC, 2005], two main methods are recommended to model synthetic turbulent wind-inflow for the structural design of wind turbines. The first method relies on the Mann turbulence model [Mann, 1998] while the second one is based on the Kaimal spectrum with an exponent coherence model [Kaimal et al., 1972]. Nevertheless, other techniques (not detailed hereafter) for producing turbulent wind inflow are available in the literature, e.g., Hilbert spectral analyses or wavelets [Gurley and Kareem, 1999, Wang and Kareem, 2005]. The Kaimal spectrum, mainly used in wind energy application, has a one-sided power spectral density defined as:

$$S_k(f) = \frac{4\sigma_k^2 \frac{L_k}{\overline{u}}}{(1 + 6f\frac{L_k}{\overline{u}})^{\frac{5}{3}}}, \tag{2.2}$$

where f is the frequency, the subscript k represents the turbulent longitudinal, crosswise or vertical components (respectively denoted by u, v, and w), L_k is the Kaimal length scale, $\overline{u}$ is the longitudinal mean wind speed at hub height, and σ_k is the standard deviation of the wind velocity. The IEC design standard [IEC, 2005] recommends the following relationships:

$$L_k = \begin{cases} 8.10\Lambda_u, \ k = u \\ 2.70\Lambda_u, \ k = v \\ 0.66\Lambda_u, \ k = w \end{cases},$$

where, Λ_u is the longitudinal turbulence length scale parameter. According to the design standard IEC 61400-1 [IEC, 2005], this scale parameter is given as:

$$\Lambda_u = \begin{cases} 0.7z_{hub}, \ z_{hub} < 60 \text{ meters} \\ 42, \ z_{hub} \geq 60 \text{ meters} \end{cases}.$$

Nevertheless, the wind inflow at hub height is not sufficient to properly simulate the wind turbine behavior. Indeed, the wind inflow over the swept area has to be properly generated based on a grid of points, see Figure 2.3. The different power spectra of each point of the grid have to be correlated thanks to a coherence function:

$$|S_{ij}(f)| = \mathrm{coh}_{i,j}(f)\sqrt{S_{ii}(f)\ S_{jj}(f)}\,,$$

where f is the frequency, S_{ij} is the cross-spectral density of points i and j, and S_{ii} and S_{jj} are respectively the discrete spectrum described in Equation (2.2) at the points i and j.

By considering an exponential spatial coherence method, this function for the longitudinal wind component of two distinct points i and j separated by a distance Δr on a plan perpendicular to the wind direction is defined as:

$$\mathrm{coh}_{i,j}(f) = \exp\left(-a_u\left(\frac{\Delta r}{z_m}\right)^{\gamma}\sqrt{\left(\frac{f\,\Delta r}{\overline{u}_m}\right)^2+\left(\frac{b'_u\,\Delta r}{L_u}\right)^2}\right)\,, \tag{2.3}$$

where z_m is the mean height of the two points, $\overline{u}_m$ is the mean of the wind speeds of the two points, a_u and b'_u are respectively the input coherence decrement and offset parameter, and γ is the coherence exponent.

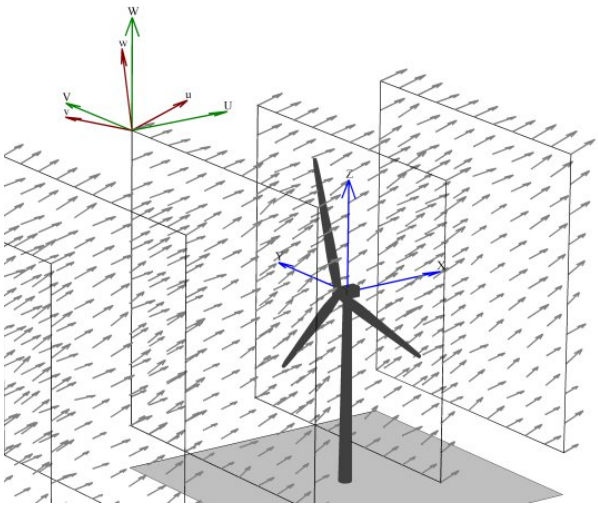


Figure 2.3 – Example of grids used by TurbSim to generate wind inflow with vertical angle set to 8 ° and the horizontal one to 15° source [Jonkman, 2009]

Finally, the Veers method (also known as Sandia method), relying on an inverse Fourier transform with random phases sampled from a quasi-random generator, is applied to the Kaimal spectrum in order to model a turbulent time-series for each of the wind components independently [Veers, 1988]. TurbSim is a full-field, turbulent-wind simulator developed by the National Renewable Energy Laboratory (NREL). It allows to generate realistic three-dimensional wind field vectors, describing the longitudinal, crosswise and vertical components of the wind [Jonkman, 2009]. As an illustration, Figure 2.4 presents the temporal evolution of the longitudinal wind speed at hub height over a period of 600 seconds with a wind mean speed $\overline{u} = 12$ m/s and a turbulence standard deviation $\sigma_u = 2.04$ m/s.

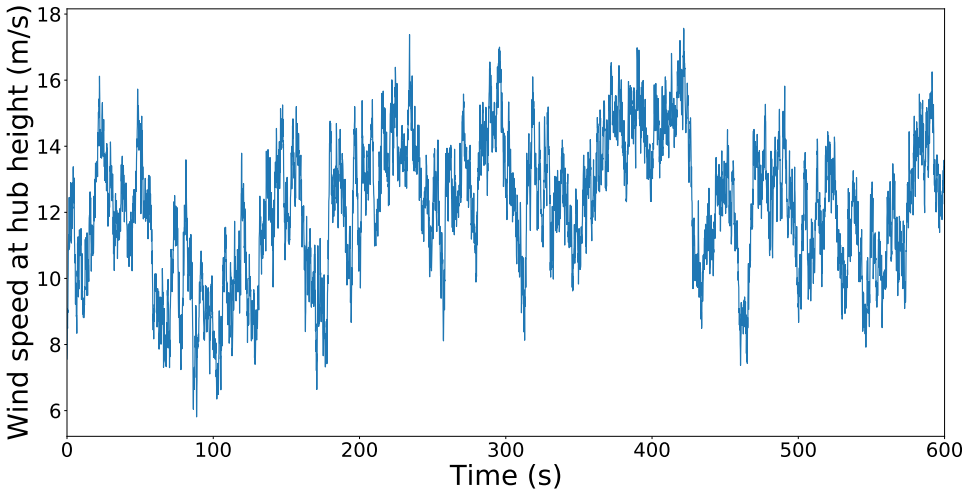


Figure 2.4 – Representation of the random generation of wind inflow using TurbSim [Jonkman, 2009]. The 10-minute mean wind speed is fixed to 12 m/s with a turbulence standard deviation $\sigma_u = 2.04$ m/s.

Lastly, TurbSim allows to specify the mean flow angle in the vertical or horizontal direction across the entire grid. These angles respectively denoted ϕ_v and in ϕ_h define the direction of the mean velocity vector with respect to the wind turbine reference coordinates system [Jonkman, 2009]. Figure 2.3 pictures the wind components generated by TurbSim across the entire grid by considering 15° horizontal and 8° vertical mean flow angles.

2.2 Aerodynamic Load computation

The structural response of a wind turbine is strongly influenced by the aerodynamic loads caused by the wind action on the blades. A popular engineering tool to approximate these aerodynamic loads in wind energy application is called the Blade-Element Momentum (BEM) theory [Leishman, 2000]. The methodology relies on the combination of the momentum and the blade element theories. The approach makes the assumption that each blade of the wind turbine is divided along the span into a finite number of sections of length dr.

Figure 2.5 gives a schematic representation of actuator disc and stream tube concepts used in the momentum theory. The application of this theory allows to obtain the axial forces dF_a and torque dQ:

$$dF_a = 4\pi\rho U_\infty^2 a(1-a)rdr\,, \tag{2.4}$$

$$dQ = 4\pi\rho U_\infty \Omega a^{'}(1-a)r^3 dr\,, \tag{2.5}$$

where a and $a^{'}$ are the axial and tangential flow induction factors, r and Ω are respectively the blade element radius and rotational speed, U_∞ is the undisturbed wind speed inflow, and ρ is the air density.

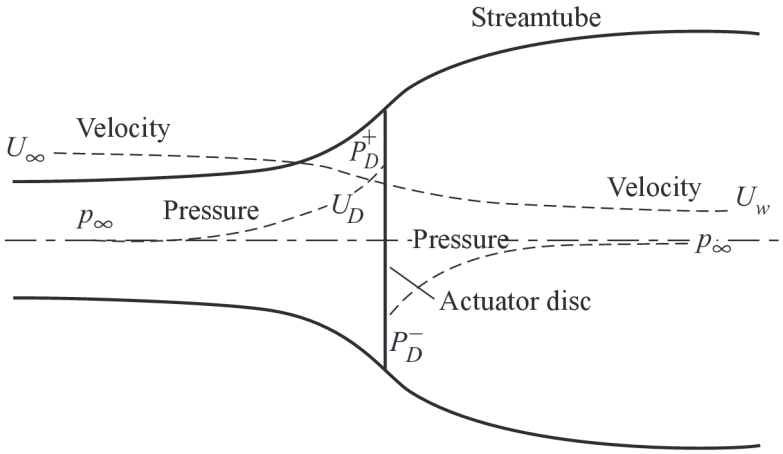


Figure 2.5 – Actuator disc and stream tube concept for the Blade-Element Momentum theory as proposed by Burton et al. [2001].

This methodology relies on some assumptions, such as that the aerodynamic interactions between different blade elements are neglected and the forces on the blade elements are only determined by the lift and drag coefficients [Kulunk and Yilmaz, 2009, Manwell et al., 2010]. By applying the blade element theory based on a geometrical analysis, see Figure 2.6, the thrust and torque quantities can be derived as:

$$dF_a = B\frac{\rho c}{2}W^2(C_L\cos\phi + C_D\sin\phi)dr\,, \tag{2.6}$$

$$dQ = rdF_t = B\frac{\rho c}{2}W^2(C_L\sin\phi - C_D\cos\phi)rdr\,, \tag{2.7}$$

where B is the number of blades, C_L and C_D are respectively lift and drag coefficients depending on the angle of attack α and the blade profile, c is the profile chord, and W is the resultant relative velocity at the blade.

By combining Equation (2.4) and Equation. (2.6), Equation (2.5) and Equation (2.7), we can develop two equations of equilibrium as:

$$4\pi\rho U_\infty^2 a(1-a)rdr = B\frac{\rho c}{2}W^2(C_L\cos\phi + C_D\sin\phi)dr\,, \tag{2.8}$$

$$4\pi\rho U_\infty \Omega a'(1-a)r^3dr = B\frac{\rho c}{2}W^2(C_L\sin\phi - C_D\cos\phi)dr\cdot \tag{2.9}$$

One of the main challenges when using this theory, is to properly determine the induction factors (a and a') which represent the momentum loss due to the presence of the rotor. A common approach is based on iteration methods [Duran, 2005, Dai et al., 2011]. After the determination of the induction factors for each section, the axial forces and torque can be computed for the entire rotor-blade system, such as:

$$\begin{aligned} F_a &= \int_{r_r}^{r_t} dF_a\,, \\ Q_{\text{aero}} &= \int_{r_r}^{r_t} dQ\,, \end{aligned} \tag{2.10}$$

where, r_r and r_t are respectively the root and tip radius of the blade.

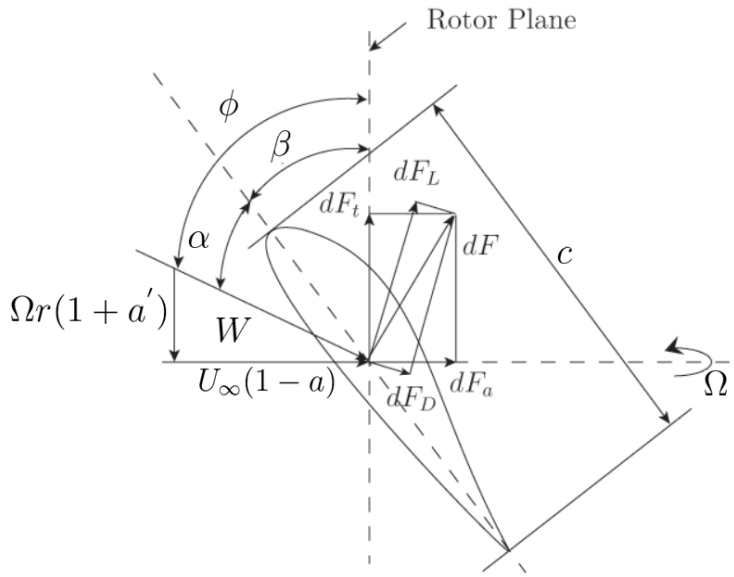


Figure 2.6 – Blade element approach with velocities and forces on a blade element adapted from [Wang et al., 2014].

In industrial codes, most of the aerodynamic solvers are based on BEM theory because of its simplicity [Robertson et al., 2013]. Nevertheless, this method faces some limitations which can have a real impact for the simulation of modern large wind turbines or depending on the operating conditions (production, idling, etc...) [see Moriarty and Hansen, 2005]. Consequently, several corrections are usually made in order to apply this methodology. They can include Glauert correction, hub- and tip-loss models, tower shadowing models, dynamic inflow correction, dynamic stall models, as well as skewed wake corrections [Blondel et al., 2016].

2.3 Control strategy

From the rotor torque defined in Equation (2.10), it is possible to estimate the power extracted from the wind such as:

$$P = \Omega Q_{\text{aero}} \cdot \tag{2.11}$$

The power extracted by a wind turbine is a function of a variable called the blade tip-speed ratio [Burton et al., 2001]. It is defined as the ratio of the rotational speed of the blade tip over the undisturbed longitudinal wind speed:

$$\lambda = \frac{\Omega r_t}{U_\infty} \cdot \tag{2.12}$$

Then, the power coefficient, denoted by C_p, allows to quantify the efficiency of the conversion of the wind energy by a rotor system. It is defined as the ratio of the extractable power to the available wind power:

$$C_p = \frac{\Omega Q_{\text{aero}}(\lambda)}{\frac{1}{2}\rho A U_\infty^3}, \tag{2.13}$$

where A is the total area swept by the blades. According to the Betz' law [Van Kuik, 2007], the entire inflow power cannot be converted by the rotor system and the power coefficient is limited to 59.3 %. Moreover, this coefficient is dependent on the tip-speed ratio and can reach a unique maximum value for a specific ratio.

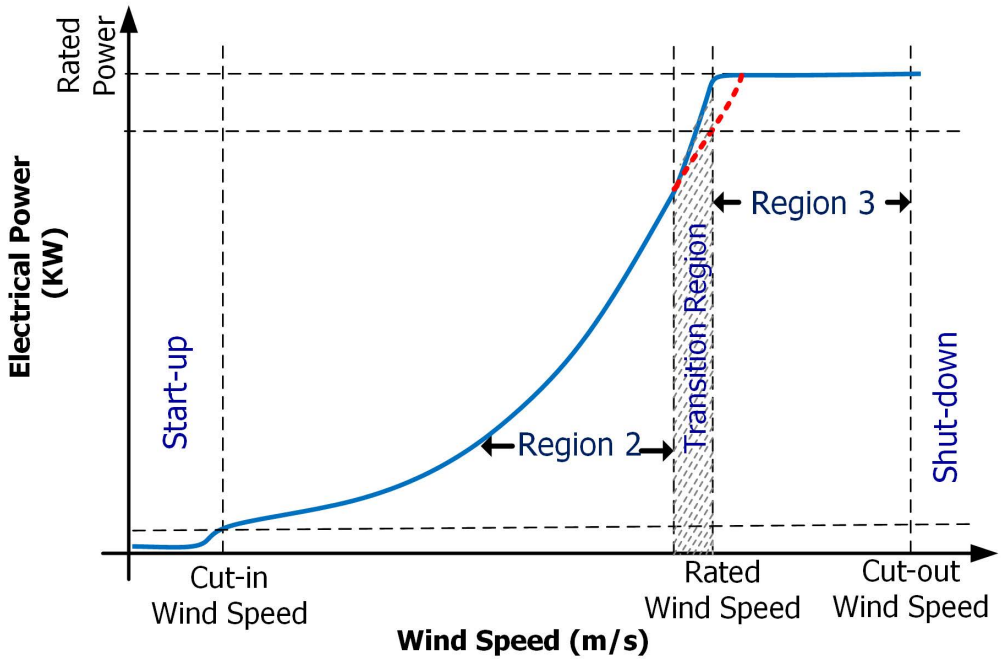


Figure 2.7 – Wind turbine operating regions [Tofighi et al., 2015]

A wind turbine operates differently depending on the wind speed value. As depicted in Figure 2.7, one can consider four different regimes. When the wind speed value is under a cut-in speed, the speed is not high enough to create a consistent rotation of the blades. Then, the turbine does not operate and the rotor is usually parked or idling to prevent an effect on the fatigue life of the blades. In this operating region, the control strategy of the turbine involves the analysis of the wind condition in order to estimate a suitable start.

When the wind speed is higher than the cut-in value, the turbine starts to rotate and the generator produces electricity. The torque control strategy is activated in order to maintain the optimum tip-speed ratio of the rotor defined in Equation (2.12).

At the rated wind speed, the turbine reaches its nominal production power according to the converter capacities. Once the rated power is reached, it has to be regulated in order not to overcome the capacity of the generator. Two main methods are used, called respectively passive and active regulation methods. In the first technology, the extracted power is controlled by using the aerodynamic stall properties of the blades. The blade geometry is designed to present stall initialization corresponding to the rated wind speed. Nevertheless, this economical technology creates aerodynamic perturbations, such as vortices, which lead to important mechanical loads and structural vibrations. In this context, active pitch control has been developed and relies on the control of the angle of attack of the blades in function of the wind speed. It uses an actuator to rotate each blades along their principal axis. Therefore, in order to regulate the rotational speed, the blade orientation is changed according to the wind solicitation.

The last regime is reached when the wind speed is higher than a cut-out speed. The rotation is mechanically stopped by a brake system and the blades are set to a parking or idling state in order to prevent structural damaging.

Most of the modern wind turbines contain a controller that regulates the blade pitch, and a variable speed generator. In our work, we consider a variable-speed generator-torque controller and proportional-integral (PI) collective blade pitch controller developed by Jonkman [2007]. In Region 2 previously described, this baseline controller uses a 2-D lookup table to determine the generator torque in order to maintain an optimal tip-speed ratio. Then in Region 3, the algorithm uses a PI feedback on low-pass filtered high speed shaft rotational speed measurements, and the control gain is rescheduled as a function of the collective pitch angle. The parameters inside the formulation of the controller strategy have been estimated for our wind turbine of interest using the procedure described in [Jonkman, 2007].

2.4 Aero-servo-elastic dynamic simulations

Wind turbines are structures subjected to complex dynamic behavior. In this context, the interaction of aerodynamics with control system and nonlinear structural reactions has to be properly taken into account. Therefore, wind turbine models tend to be sophisticated and to rely on coupled aero-servo-elastic simulations. This means that the turbine is not split up into several component models being solved independently, but dependencies and interactions are considered by employing a global model, i.e., aero-servo-elastic model.

Over the last decades, various servo-aero-elastic numerical codes have been developed, such as FAST (Fatigue, aerodynamics, structures, and turbulence) [Jonkman and Jonkman, 2016], Bladed [DNV GL, 2013], HAWCK2 (Horizontal axis wind turbine code 2nd generation) [Larsen and Hansen, 2007] or Deeplines Wind™ [Principia]. Böker [2010] proposes an overview of the most used aero-servo-elastic numerical codes in the context of offshore wind turbines. Moreover, several code cross-verification studies have been performed [see Schepers et al., 2002, Jonkman and Musial, 2010, Popko et al., 2012]. During these verification studies, the specific structure used to compare these numerical codes is mainly the NREL 5MW reference wind turbine [Jonkman et al., 2009].

In order to model and simulate the nonlinear structural response of wind turbines under external solicitations, these numerical codes mainly rely on a combination of rigid body parts (nacelle, hub, generator) and flexible components (blades, tower, shaft). Most of the time, beam finite elements are used to model blades and tower due to the fact that they are slender structures with specific bending, tension, and torsion properties. In our study we mainly rely on the open-source aero-servo-elastic code Fast and the proprietary software Deeplines Wind™ developed by PRINCIPIA in collaboration with IFP Energies Nouvelles.

Firstly, Deeplines Wind™ [Principia] is an aero-servo-elastic simulator based on a full finite element discretization. The flexible components of the model are discretized through finite elements interconnected at points called nodes. Each finite element has physical properties related to the component of interest, such as mass, inertia, and stiffness. Under material and geometrical linearity assumptions, a finite elements model can be described by the following system of equations:

$$\mathbf{M}\ddot{\mathbf{q}} + \boldsymbol{C}\dot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{F}\,, \tag{2.14}$$

where $\mathbf{q}$, $\dot{\mathbf{q}}$ and $\ddot{\mathbf{q}} \in \mathbb{R}^n$ are respectively the nodal displacement, velocity and acceleration vectors of the degrees of freedom from the finite element discretization, $\mathbf{F}$ is the external force vector, such as the action of wind on the structure, and $\mathbf{M} \in \mathbb{R}^{n\times n}$ is the mass matrix, $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ is the stiffness matrix, and $\mathbf{K} \in \mathbb{R}^{n\times n}$ is the damping matrix. These global matrices are obtained from the assembly of the local matrices of each finite element of the model. In Deeplines Wind ™ software, beam finite elements use Mindlin formulation also known as the thick beam theory as it takes into account the transverse shear of the elements. The theory is coded in the nonlinear framework of large rotations, large displacements and large deformations [see Fargues, 1995].

Due to the fact that external solicitation is time varying, the temporal discretization of Equation (2.14) in Deeplines Wind™ is performed thanks to a Newmark scheme that we briefly describe hereafter. The reader is referred to the work of Newmark [1959] for further details. Let us suppose that the complete solution of the system is known at iteration t_n, i.e., displacement, velocity, and acceleration vectors corresponding to the degrees of freedom have been estimated. The Newmark method states that the velocity and displacement vectors in the equation of motion at iteration t_{n+1} can be expressed as:

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \Delta t \dot{\mathbf{q}}_n + \frac{\Delta t^2}{2}[(1-2\beta)\ddot{\mathbf{q}}_n + 2\beta\ddot{\mathbf{q}}_{n+1}]\,, \tag{2.15}$$

$$\dot{\mathbf{q}}_{n+1} = \dot{\mathbf{q}}_n + \Delta t[(1-\gamma)\ddot{\mathbf{q}}_n + \gamma\ddot{\mathbf{q}}_{n+1}]\,, \tag{2.16}$$

where γ and β are the Newmark time integrators chosen according to the applications, and Δt is the iteration time step.

From Equation (2.15), one can express the acceleration vector at time step t_{n+1} only considering known terms and the displacement vector at iteration t_{n+1}:

$$\ddot{\mathbf{q}}_{n+1} = \frac{\mathbf{q}_{n+1} - \mathbf{q}_n}{\beta\Delta t^2} - \frac{\dot{\mathbf{q}}_n}{\beta\Delta t} - \left(\frac{1}{2\beta} - 1\right)\ddot{\mathbf{q}}_n\,. \tag{2.17}$$

By substituting Equation (2.15) in Equation (2.16), one can obtain the expression of the velocity vector at time t_{n+1} only as a function of known terms and of the displacement vector at time t_{n+1}:

$$\dot{\mathbf{q}}_{n+1} = \left(1 - \frac{\gamma}{\beta}\right)\dot{\mathbf{q}}_n + \Delta t\left((1-\gamma) - (\frac{\gamma}{2\beta} - \gamma)\right)\ddot{\mathbf{q}}_n + \frac{\gamma}{\Delta t\beta}(\mathbf{q}_{n+1} - \mathbf{q}_n)\,. \tag{2.18}$$

Then at iteration t_{n+1}, one can apply the Newton-Raphson iterative scheme to solve the system, defined in Equation (2.14), for determining $\mathbf{q}_{n+1}$. Finally, velocity and acceleration vectors are updated thanks to Equations (2.17) and (2.18).

Nevertheless, high-fidelity wind turbine models based on full finite element discretization, as the ones used in Deeplines Wind™, can involve many degrees of freedom (DoFs), e.g., 564 DoFs for our considered onshore wind turbine. Therefore, a reduction of the DoFs is performed in some cases to deal with affordable computing times. In that context, the second aero-servo-elastic software, called FAST [Jonkman and Jonkman, 2016], is based on a combined modal and multibody dynamics formulation [Jonkman et al., 2005]. The model uses rigid bodies and flexible components to describe the structure. The flexible components are described by using a linear modal transformation that assumes small deflection, small rotations, and small deformations. In FAST this transformation, proposed by Craig Jr. and Bampton [1968], relies on the Ritz transformation:

$$\Phi^T\mathbf{M}\Phi\ddot{\mathbf{q}} + \Phi^T\boldsymbol{C}\Phi\dot{\mathbf{q}} + \Phi^T\mathbf{K}\Phi\mathbf{q} = \Phi^T\mathbf{F}\,, \tag{2.19}$$

where Φ is a matrix containing the modes to be included in the model. This reduction technique allows to decrease drastically the number of DoFs and hence the size of the system to be solved. For example in FAST, the structural model of an onshore wind turbine considers only 20 degrees of freedom. It considers four DoFs for the tower, i.e., longitudinal and lateral first and second tower bending modes. Two flapwise and one edgewise bending-mode DOFs are considered per blade. The generator azimuth angle is modeled by one DoF as well as the torsional flexibility in the drivetrain. Nevertheless, this reduction is constraint by the requirement of the linearity of the equations of motion. In FAST, the temporal discretization is performed thanks to a constant-time-step Adams-Bashforth-Adams-Moulton predictor-corrector integration scheme [Jonkman, 2003].

Both FAST and Deeplines Wind™ aero-servo-elastic simulators are based on the Blade-Element Momentum theory, and use the same control system in the form of an external Dynamic-Link library (DLL). In Deeplines Wind™, the exchange of information between the mechanical solver and the external aerodynamic or control libraries is performed through application programming interfaces, at the start of each iteration time step, see Figure 2.8. For a fully detailed study of the differences between the two software, the reader is referred to the work of Le Cunff et al. [2013].

A specific remark has to be made on the use of these models for time-domain simulations. Indeed, the initial part of aero-servo-elastic numerical simulations is often characterized by a start-up transient period. The transient period is due to the application of the gravity and rotor rotation on the model assumed to be initially at rest or in a stationary state [Jonkman and Jonkman, 2016]. However, these start-up behaviors disappear over time due to damping. Consequently, one has to testify that the response statistics are truly representative of the structural responses of the wind turbine before any post-processing. IEC 61400-1 [IEC, 2005] recommends to remove the first 5 seconds or longer from the simulation statistics to reduce the impact of start-up transient periods. Haid et al. [2013] suggest that with proper initial conditions the 60 first seconds from the simulation has to be removed. After an autocorrelation study of the fatigue loads, we have decided to remove the 250 first seconds from the dynamic simulations.

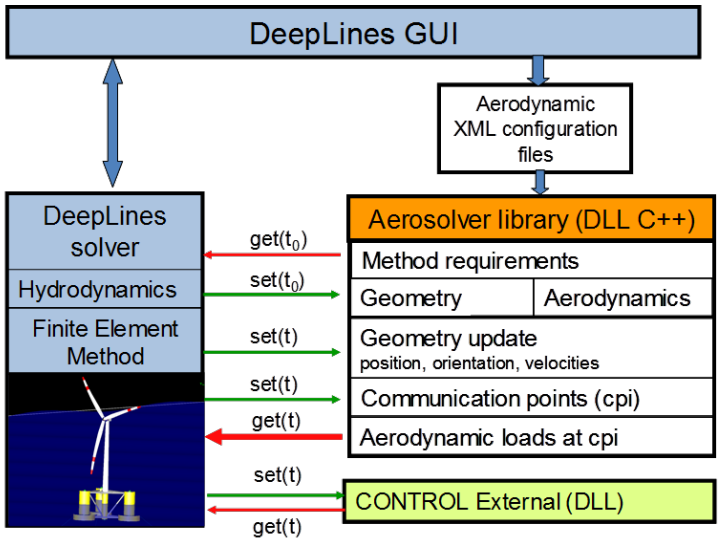


Figure 2.8 – Deeplines Wind™ architecture

2.5 Fatigue assessment

As highlighted by Matha [2010], the design of wind turbine components is generally not governed by ultimate loads but mainly by fatigue loads. In this context, the design assessment of a wind turbine relies on the estimation of the accumulated fatigue damage that the structure is expected to face during its overall lifetime due to the fluctuating loadings. The fatigue accumulation at a location of the structure is represented by a damage variable D computed by using a rainflow cycles counting algorithm and by applying the Palmgren-Miner linear damage law with the Whöler curve of the considered material.

The rainflow counting algorithm [Cosack, 2011] consists in analyzing the fatigue cycles contained in the load time-series by counting and sorting them depending on their amplitude range. As depicted in Figure 2.9, the algorithm is divided into two main steps. First, a reduction of the time-series is performed by transforming it into a sequence of peaks and valleys. Then, from the obtained sequence, a cycle counting procedure is carried out in order to extract the cycles. Among the different counting algorithms, Algorithm 1 presents the Pagoda roof cycle counting procedure.

Once the rainflow counting algorithm has determined the number n_i of cycles for each bin of stress range S_i in the 10-minute time-series of structural loading response, one can estimate the total damage using the Palmgren-Miner linear damage rule [Miner et al., 1945]. It relies on the assumption that the damage is only dependent on the different cycles, and on the linear damage hypothesis , i.e., the fatigue damage is the combined sum of every different fatigue cycles contribution. By definition, when the obtained total damage is equal to one, failure occurs. The Palmgren-Miner rule is defined as:

$$D = \sum_{i=1}^{N_c} \frac{n_i}{N_i}\,, \tag{2.20}$$

where $i = 1, \ldots, N_c$ corresponds to each range bin, n_i is the rainflow counting number of cycles for the i-th bin, and N_i is the number of cycles to failure for bin i. This value is determined with the Whöler curve of the material , also known as S-N curve, which gives the relation between the number of cycles to failure N_i and the cycle range value S_i. This relation is mathematically defined as:

$$N_i = \left(\frac{S_0}{S_i}\right)^m\,, \tag{2.21}$$

where the empirical coefficients S_0 and m, depending on the mechanical characteristics of the material, are respectively the critical stress level and the negative inverse slope. By combining Equations (2.21) and (2.20), the fatigue damage can be written as:

$$D = S_0^{-m} \sum_{i=1}^{N_c} n_i S_i^m\,.$$

The notion of damage equivalent load (DEL) [Veldkamp, 2006] is often used and is defined as a virtual stress amplitude that would create the damage D in N_{ref} regular

cycles.

$$D = \frac{N_{ref} DEL^m}{S_0^m} = \frac{1}{S_0^m} \sum_{i=1}^{N_c} S_i^m \; n_i \,,$$
$$\rightarrow \quad DEL = \left(\frac{\sum_{i=1}^{N_c} S_i^m \; n_i}{N_{ref}} \right)^{\frac{1}{m}} \,, \tag{2.22}$$

where N_{ref} is the reference number of cycles usually set to an arbitrary value.

Algorithm 1: Pagoda roof cycle counting procedure.

1. Consider that the sequence of peaks and valleys is a template for a rigid sheet, and turn it clockwise 90°.
2. Each peak is seen as a point where the water starts to drip down.
3. Count the half-cycles by searching for interruption in the flow happening when either:
   - — the flow reaches the end of the sequence;
   - — the opposite peak has a bigger magnitude; or
   - — the flow merges with another one having started at an earlier peak.
4. Repeat step 3. for valleys.
5. Estimate a magnitude to each half-cycle by using the magnitude difference between its start and termination.
6. Combine all the half-cycles of the same magnitude in order to count the number of whole cycles.

Conclusion

In this chapter we have discussed some of the main components in aero-servo-elastic simulation of wind turbine structures. It starts with a description of the modeling of stochastic synthetic wind fields based on a turbulence spectral model depending on the wind speed and the standard deviation. The mentioned approach relies on the Kaimal spectrum with an exponent coherence model. Then, from these synthetic wind-inflows, one can approximate the obtained aerodynamic loads by using the blade element momentum theory. We can further investigate the structural dynamic behavior of the wind turbine by taking into account the interaction of aerodynamics with control system, and nonlinear structural reactions. In this context, aero-servo-elastic numerical codes have been developed. In particular, they allow to compute the fatigue damage that the turbine structure is supposed to face during its entire lifetime due to the fluctuating loadings.

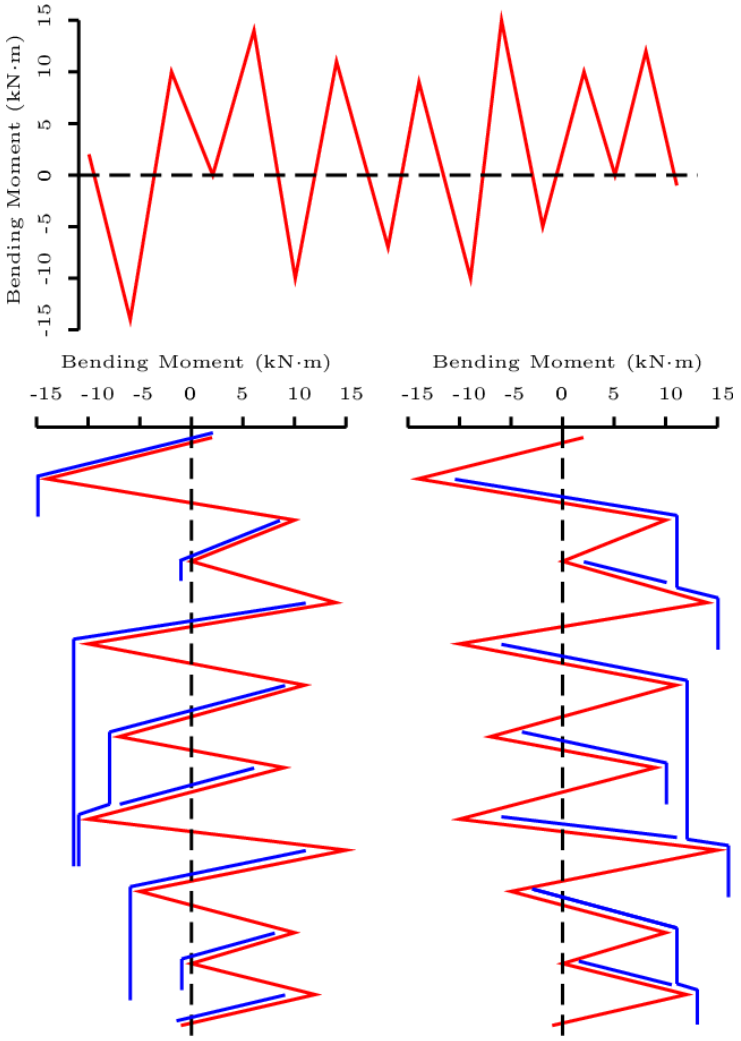


Figure 2.9 – Illustration of the rainflow cycle counting method [Si, 2015]

Part II

Methodological tools

Measure what is measurable, and make measurable what is not so.

Galileo Galilei

3

Variance-Based Sensitivity Analysis - Sobol' indices

Comme évoqué dans la partie précédente, l'un des intérêts de l'analyse de sensibilité est de déterminer les paramètres d'entrée ayant une grande importance sur la variabilité d'une sortie d'un modèle numérique [Saltelli et al., 2000]. Ce chapitre est consacré aux méthodes d'analyse de sensibilité globales sous l'hypothèse d'indépendance des variables d'entrée. Nous nous focaliserons sur les méthodes basées sur le calcul des indices de sensibilité de Sobol' permettant de quantifier l'influence de chaque paramètre et d'identifier l'existence de potentielles interactions entre ces différents paramètres. La Section 3.1 présente l'analyse fonctionnelle de la variance (FANOVA). Dans les Sections 3.2 et 3.3, nous présentons les indices de Sobol' et leurs estimateurs de type Monte Carlo. Nous ne traiterons pas des autres méthodes d'estimation disponibles dans la littérature comme par exemple celles reposant sur une décomposition spectrale de la sortie d'intérêt du modèle numérique [Cukier et al., 1973, Tarantola et al., 2006, Ghanem and Spanos, 2003, Prieur and Tarantola, 2017]. Dans la Section 3.4, le cas spécifique d'étude de sensibilité sur des séries temporelles est présenté par la méthode des indices de Sobol' généralisés, appelés également indices agrégés [Lamboni et al., 2008], basée sur une réduction de la dimension par analyse en composantes principales (ACP) [Anderson, 1958, Jolliffe, 1986, Besse, 1992]. Afin de mener de telles analyses de sensibilité, un certain nombre de simulations issues du modèle numérique est nécessaire. Pour mener ces simulations numériques, nous nous basons généralement sur des plans d'expériences. Dans la Section 3.5, nous nous intéressons tout particulièrement à l'Hypercube Latin qui est couramment utilisé en simulation numérique du fait de ses propriétés de remplissage de l'espace.

Introduction

As previously discussed, numerous numerical models, such as aero-servo-elastic models, involve input parameters, which are not precisely known. Global sensitivity analysis (GSA) techniques aim to identify the inputs whose uncertainty has the largest impact on the variability of an output of the model, also known as quantity of interest (QoI) [Saltelli et al., 2008]. One widely used statistical tool in order to quantify the influence of each input parameter on a QoI is based on the Sobol' sensitivity indices [Sobol', 1993]. These sensitivity indices measure the part of the QoI variance due to one or a set of input parameters. This approach refers to the variance-based method and consists in decomposing the model of interest, denoted by f, into a finite hierarchical expansion of functions. This decomposition is called functional analysis of variance (FANOVA) and is also known as the high-dimensional model representation (HDMR) technique [Hoeffding, 1948]. If the inputs are independent, each of these partial variance terms quantifies the uncertainty on the output induced by an individual input or a group of inputs. Among all Sobol' indices one can distinguish first-order and total effect indices. The first ones measure the effect of a single input, while the second ones quantify the effect of a single input plus all its interactions with the other inputs. Then, as proposed in [Saltelli et al., 2004], the closed $|\mathbf{u}|$-th order indices can be considered and allow to quantify up to $|\mathbf{u}|$-th order interactions in addition to the main effect of each of the $\mathbf{u}$-tuple of input parameters. Nevertheless, variance indicator can sometimes poorly represents the variability of a distribution. In this context, other methods not detailed in this chapter have been developed in the literature, e.g., distribution based sensitivity indices [Borgonovo, 2007, Borgonovo et al., 2011] or entropy-based sensitivity measures [Krzykacz-Hausmann, 2001, Liu et al., 2006, Auder and Iooss, 2008]. The reader is referred to the article of Borgonovo and Plischke [2016] for a complete review of sensitivity methods.

When facing complex numerical models, the analytical expressions of Sobol' indices are most of the time inaccessible. Indeed, the complexity of the function describing the numerical model of interest causes the solution of index integrals to be intractable. In such situations, one can rather estimate these sensitivity indices. The estimation of these quantities is often performed through Monte Carlo or quasi-Monte Carlo methods [Helton et al., 2006]. However, the numerical computation of such index estimators requires a large number of numerical model calls, i.e., the computational cost, to estimate all first-order and total Sobol' indices, depends linearly on the dimension of the input space. The original estimation procedure to estimate first-order Sobol' indices was developed by Sobol' [Saltelli et al., 1993, Sobol', 2001]. Then, Saltelli [2002] proposed combinatorial strategies to estimate sets of Sobol' indices, i.e., first-order and total effect indices at once. In this work, we mention two Sobol' index estimators proposed respectively by Homma and Saltelli [1996] and by Monod et al. [2006] then further studied in [Janon et al., 2014]. Moreover, Monte Carlo procedures are necessarily tainted by a sampling error. We estimate this sampling error by using a bootstrap resampling technique [Efron and Tibshirani, 1993, Archer et al., 1997]. This method allows to produce confidence intervals of the Sobol' indices with a moderate numerical cost.

Sobol' indices, for GSA, have been developed in the context of univariate output. Nevertheless in many cases such as wind turbine modeling, numerical models produce functional or multivariate output. When dealing with multivariate output, such as discretized functional output, a straightforward practice is to consider each component as

distinct output variable in order to perform a GSA [Marrel et al., 2015a]. However, this basic implementation of sensitivity indices does not take into account the functional aspect of the output especially the high level of redundancy between close components [Campbell et al., 2006]. In this paper, a generalization of the Sobol' indices for multivariate output based on the so-called aggregated indices is considered. The method relies on principal component analysis (PCA) and on FANOVA decomposition allowing to estimate the influence of each parameter, or set of parameters, on the whole multivariate output. Functional PCA, also known as Karhunen-Loève decomposition, consists in projecting the output on a new basis so that most information is concentrated in the first few components [Anderson, 1958, Jolliffe, 1986, Besse, 1992]. After decomposing the output in an orthogonal basis using PCA, generalized Sobol' sensitivity indices are computed on the coefficients of the expansion and then aggregated in an index, [see Lamboni et al., 2008].

In variance-based sensitivity analysis, Monte Carlo based procedures for Sobol' indices estimation consist in an approximating of multidimensional integrals thanks to sampling designs. A classic sampling scheme consists in considering independent and identically distributed designs, as in a crude Monte Carlo sample. The major drawback of this estimation procedure is the large number of model calls to compute reliable sensitivity indices. When dealing with time-consuming numerical models, using an effective sampling strategy is mandatory to compute the most accurate Sobol' indices with the fewest model calls. In this context, we can rely on space-filling sampling strategies, e.g., Latin Hypercube Sample [McKay et al., 1979].

This chapter is organized as follows. Section 3.1 presents the FANOVA decomposition for deterministic models. Sections 3.2 and 3.3 respectively define Sobol' indices and the associated estimation procedures in the framework of independent input parameters. Section 3.4 is devoted to the procedure to define and to estimate aggregated Sobol' indices when dealing with multivariate output. Lastly in Section 3.5, we present a space-filling strategy called Latin Hypercube Sampling (LHS).

3.1 Functional analysis of variance decomposition

A review of the functional analysis of variance (FANOVA) decomposition can be found in [Tissot, 2012]. Hereafter, we will briefly present the FANOVA decomposition by considering a deterministic numerical model f depending on p mutually independent random inputs. This decomposition was first introduced by Hoeffding [1948] and then was properly formalized in the work of Efron and Stein [1981]. The FANOVA decomposition can be defined with the formalism described by Sobol' [1993], such as:

Definition 1. *Let $\mathbf{x} = (x_1, \ldots, x_p)$, $p \geq 1$, be the vector of input parameters of f. We assume that $f \in \mathbb{L}^2([0,1]^p)$ where $f(\mathbf{x})$ is defined for all $\mathbf{x} \in [0,1]^p$. Let us decompose $f(\mathbf{x}) = f(x_1, \ldots, x_p)$ as the sum of increasing dimension functions:*

$$f(\mathbf{x}) = f_{\varnothing} + \sum_{i=1}^{p} f_i(x_i) + \sum_{i=1}^{p-1} \sum_{j>i}^{p} f_{i,j}(x_i, x_j) + \ldots + f_{1,\ldots,p}(x_1, \ldots, x_p)\,. \tag{3.1}$$

Let us denote $\mathbf{u}$ a subset of $\mathcal{P} = [0,1]^p$, $-\mathbf{u}$ its complement, and $|\mathbf{u}|$ its cardinality. The

decomposition described by Equation (3.1) can be rewritten as:

$$f(\mathbf{x}) = f_{\varnothing} + \sum_{\mathbf{u}\subseteq\mathcal{P},\mathbf{u}\neq\varnothing} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})\,.$$

This decomposition is unique under the following set of constraints: $f_{\varnothing}$ is a constant and $\int_{[0,1]} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})dx_i = 0\,, \forall i \in \mathbf{u}, \forall \mathbf{u} \subseteq \{1,\dots,p\}$. It is then known as Sobol'-Hoeffding or FANOVA decomposition.

As a consequence, the summands are orthogonal to each other:

$$\int_{[0,1]^p} f_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}}) f_{\mathbf{v}}(\mathbf{x}_{\mathbf{v}}) d\mathbf{x} = 0\,, \forall \mathbf{u}, \mathbf{v} \subseteq \{1,\dots,p\}\,, \mathbf{u} \neq \mathbf{v}\,. \tag{3.2}$$

Moreover, a second outcome is that the constant $f_{\varnothing}$ is equal to the mean value of the function:

$$f_{\varnothing} = \int_{[0,1]^p} f(\mathbf{x}) d\mathbf{x}\,.$$

Lastly, the terms in the FANOVA decomposition can be derived analytically:

$$f_{\mathbf{u}}(\mathbf{x}) = \int_{[0,1]^{|\mathbf{u}|}} f(\mathbf{x}) d\mathbf{x}_{-\mathbf{u}} - \sum_{\mathbf{v}\subsetneq\mathbf{u}} f_{\mathbf{v}}(\mathbf{x})\,.$$

The FANOVA decomposition is the key stone to define Sobol' indices in the framework of independent input parameters.

3.2 Definition of Sobol' indices

Let us consider that the uncertainty on $\mathbf{x}$ is modeled by a random vector that we suppose uniformly distributed on $[0,1]^p$.

$$\mathbf{X} = (X_1,\dots,X_p) \sim \mathcal{U}[0,1]^p\,.$$

Thus, the quantity of interest (QoI) $Y = f(\mathbf{X})$ is a random variable. Due to orthogonality constraints in Equation (3.2), it is possible to decompose the variance of the QoI as:

$$\begin{aligned} Var[Y] &= Var[f(\mathbf{X})] \\ &= \mathbb{E}[(f(\mathbf{X}) - f_{\emptyset})^2] = \sum_{\mathbf{u}\subseteq\mathcal{P},\mathbf{u}\neq\varnothing} Var[f_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}})]\,. \end{aligned} \tag{3.3}$$

For any $j \in \{1,\dots,p\}$, $Var[f_j(X_j)]$ represents the variance of the output due to the main effect of the parameter X_j. For any $j,k \in \{1,\dots,p\}, j < k$, the term $Var[f_{j,k}(X_j,X_k)]$ represents the joint effect of the parameters X_j and X_k on the output Y. And so on for partial variances of higher order. We then define the Sobol' indices:

(i) The Sobol' index of order $|\mathbf{u}|$ associated to the input vector $\mathbf{X}_{\mathbf{u}}$ is defined as:

$$S_{\mathbf{u}} = \frac{Var[f_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}})]}{Var[Y]}\,,\ \mathbf{u} \subseteq \{1,\dots,p\}, \mathbf{u} \neq \varnothing\,. \tag{3.4}$$

In the context we are only interested in a single input, i.e., $|\mathbf{u}| = 1$, the index in Equation (3.4) is called first-order sensitivity index. It measures the contribution of a single input to the output variance. If $|\mathbf{u}| > 1$, $S_{\mathbf{u}}$ evaluates the importance of the interaction of order $|\mathbf{u}|$ between the inputs X_j, $j \in \mathbf{u}$ with respect to the QoI Y.

(ii) The closed $|\mathbf{u}|$-order Sobol' index for the input vector $\mathbf{X_u}$ is defined as

$$\tilde{S}_{\mathbf{u}} = \frac{\underline{\tau}^2_{\mathbf{u}}}{Var[Y]}\,,\ \mathbf{u} \subseteq \{1,\dots,p\},\ \mathbf{u} \neq \varnothing\,, \tag{3.5}$$

where:

$$\underline{\tau}^2_{\mathbf{u}} = \sum_{\mathbf{v}\subseteq\mathbf{u}} Var[f_{\mathbf{v}}(\mathbf{X_v})]\cdot$$

The closed $|\mathbf{u}|$-order Sobol' index in Equation (3.5) quantifies the main effect of $\mathbf{X_u}$ and the effect of all interactions between variables in $\mathbf{X_u}$ on Y.

(iii) The total effect Sobol' index of order $|\mathbf{u}|$ associated to the input vector $\mathbf{X_u}$ is defined as:

$$\overline{S}_{\mathbf{u}} = \frac{\overline{\tau}^2_{\mathbf{u}}}{Var[Y]}\,,\ \mathbf{u} \subseteq \{1,\dots,p\},\ \mathbf{u} \neq \varnothing\,,$$

where:

$$\overline{\tau}^2_{\mathbf{u}} = \sum_{\mathbf{v}\cap\mathbf{u}\neq\varnothing} Var[f_{\mathbf{v}}(\mathbf{X_v})]\,,\mathbf{v} \subseteq \{1,\dots,p\}\cdot$$

The total effect Sobol' index of order $|\mathbf{u}|$ aims to quantify the main effect of $\mathbf{X_u}$ and the effect of all interactions between variables in $\mathbf{X_u}$ and variables in $\mathbf{X_{-u}}$ on Y.

By definition, according to Equation (3.3), these Sobol' indices satisfy:

$$1 = \sum_{j=1}^{p} S_i + \sum_{1\leq j<k\leq p} S_{j,k} + \ldots + S_{1,\ldots,p}\cdot$$

Example 1. *Let us consider a three-factor model f with X_1, X_2 and X_3, then:*

$$\tilde{S}_{\{1,2\}} = S_1 + S_2 + S_{\{1,2\}}\,,$$

and,

$$\overline{S}_1 = S_1 + S_{\{1,2\}} + S_{\{1,3\}} + S_{\{1,2,3\}}\cdot$$

3.3 Estimation of Sobol' indices

Usually the complexity of the numerical model of interest causes the analytical estimation of these Sobol' indices to be intractable. In such situations, one can rather estimate these sensitivity indices. The estimation of these quantities is usually based on Monte Carlo type procedures, [see Helton et al., 2006].

As in the previous sections, f is a deterministic model defined on $\mathcal{P} \subset \mathbb{R}^p$ and valued in $\mathbb{R}$. Let us describe the so-called pick and freeze method (see, e.g., Janon et al. [2014]). Let us consider $\mathbf{X}$ and $\mathbf{X}'$ two independent random vectors distributed according to the input vector. Let $\mathbf{u} \subseteq \{1,\dots,p\}$, $\mathbf{u} \neq \varnothing$, from Lemma 1.2 in [Janon et al., 2014], the closed $|\mathbf{u}|$-order Sobol' index can be expressed using covariances:

$$\tilde{S}_{\mathbf{u}} = \frac{Cov(Y, Y^{\mathbf{u}})}{Var[Y]}\,, \tag{3.6}$$

with $Y = f(\mathbf{X})$ and $Y^{\mathbf{u}} = f(\mathbf{X^u})$, where $\mathbf{X^u} = (X_j^{\mathbf{u}})_{1\leq j\leq p}$ with $\begin{cases} X_j^{\mathbf{u}} = X_j & \text{if } j \in \mathbf{u} \\ X_j^{\mathbf{u}} = X_j' & \text{otherwise} \end{cases}$.

An insightful estimator consists in considering the empirical estimators of the covariance and variance used in Equation (3.6). By taking the formalism of Sobol' in [Sobol', 1993], let us consider $\mathbf{P}$ and $\mathbf{P}'$ two designs of size s, such as:

$$\mathbf{P} = \{\mathbf{x}_i\}_{i=1}^s,\ \mathbf{P}^{'} = \{\mathbf{x}_i^{'}\}_{i=1}^s\,.$$

Each row of the design is a point $\mathbf{x}_i$ in $\mathcal{P}$, the j-th column of the design refers to a sample of X_j and for $\mathbf{u} \subseteq \{1,\dots,p\}$, $\mathbf{P^u} = \{\mathbf{x}_i^{\mathbf{u}}\}_{i=1}^s$ with $\forall j \in [1,p]$ $\begin{cases} x_j^{\mathbf{u}} = x_j & \text{if } j \in \mathbf{u} \\ x_j^{\mathbf{u}} = x_j^{'} & \text{otherwise} \end{cases}$.
Then, let denote:

$$\forall i = 1,\dots,s,\ y_i = f(\mathbf{x}_i) \text{ and } y_i^{\mathbf{u}} = f(\mathbf{x}_i^{\mathbf{u}})\,.$$

An estimator of $\tilde{S}_{\mathbf{u}}$ is then:

$$\widehat{\tilde{S}}_{\mathbf{u}} = \frac{\frac{1}{s}\sum_{i=1}^s y_i\ y_i^{\mathbf{u}} - (\frac{1}{s}\sum_{i=1}^s y_i)\ (\frac{1}{s}\sum_{i=1}^s y_i^{\mathbf{u}})}{\frac{1}{s}\sum_{i=1}^s y_i^2 - (\frac{1}{s}\sum_{i=1}^s y_i)^2}\,. \tag{3.7}$$

Moreover, due to the fact that $\mathbb{E}[Y] = \mathbb{E}[Y^{\mathbf{u}}]$, the empirical estimation of $\mathbb{E}[Y^{\mathbf{u}}]$ can be replaced by the one of $\mathbb{E}[Y]$ and then:

$$\widehat{\tilde{S}}_{\mathbf{u}} = \frac{\frac{1}{s}\sum_{i=1}^s y_i\ y_i^{\mathbf{u}} - \left(\frac{1}{s}\sum_{i=1}^s y_i\right)^2}{\frac{1}{s}\sum_{i=1}^s y_i^2 - (\frac{1}{s}\sum_{i=1}^s\ y_i)^2}\,.$$

This estimator has been developed in [Homma and Saltelli, 1996]. As highlighted by Monod et al. [2006], a second estimator can be introduced due to the fact that Y and $Y^{\mathbf{u}}$ have the same distribution. This new estimator, hereafter denoted by $\widehat{\tilde{T}}_{\mathbf{u}}$, has theoretical guarantees given in [Janon et al., 2014] and is defined as:

$$\widehat{\tilde{T}}_{\mathbf{u}} = \frac{\frac{1}{s}\sum_{i=1}^s y_i\ y_i^{\mathbf{u}} - \left(\frac{1}{s}\sum_{i=1}^s \frac{y_i+y_i^{\mathbf{u}}}{2}\right)^2}{\frac{1}{s}\sum_{i=1}^s \frac{y_i^2+(y_i^{\mathbf{u}})^2}{2} - \left(\frac{1}{s}\sum_{i=1}^s \frac{y_i+y_i^{\mathbf{u}}}{2}\right)^2}\,.$$

In [Janon et al., 2014], the authors prove indeed that $(\widehat{\tilde{T}}_{\mathbf{u}})_s$ is asymptotically normal, with variance σ_2^2/N, where:

$$\sigma_2^2 = \frac{Var\left[\left((Y-\mathbb{E}[Y])(Y^{\mathbf{u}}-\mathbb{E}[Y]) - \tilde{S}_{\mathbf{u}}/2\,((Y-\mathbb{E}[Y])^2 + (Y^{\mathbf{u}}-\mathbb{E}[Y])^2)\right)\right]}{(Var[Y])^2}\,,$$

and that this asymptotic variance is minimal in comparison to other regular estimators.

Nevertheless, the main drawback of this classical Monte Carlo procedure is its cost in terms of number of calls to the numerical model. Indeed, in order to compute all first-order Sobol' indices by using the estimator in Equation (3.7), the strategy requires $s(p+1)$ evaluations for an accuracy of order $s^{-\frac{1}{2}}$ (due to the central limit theorem). In this context, the solution is not feasible in case of large input space dimension unless the numerical model f is cheap. Note that the sampling error coming from the Monte Carlo evaluation of the variances can be estimated either by random repetitions [Iooss et al., 2006], asymptotic formulae [Janon et al., 2014] or bootstrap methods [Archer et al., 1997]. A formulation of the estimation of the confidence intervals based on a bootstrap procedure is described in Appendix A. In the literature other Monte Carlo based estimation formulae can be found such as Jansen formula for estimating total Sobol' indices [Jansen, 1999].

3.4 Sobol' indices with functional output

As in the previous sections, let us consider f, the function representing a deterministic numerical model which takes as input parameters the vector $\mathbf{X} = (X_1, \ldots, X_p)$. In this section, the QoI of the numerical model f is a discretized functional output which can be considered as a d multivariate vector $\mathbf{Y} = (Y_1, \ldots, Y_d)^T$, such as:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_d \end{pmatrix} = f(\mathbf{X}) = f(X_1, \ldots, X_p) \cdot$$

Assuming that $X_1, \ldots, X_p$ are independent and that the mapping function f is a square-integrable function, i.e., $E\left(\|\mathbf{Y}\|^2\right) < \infty$. In order to quantify the influence of each input parameter X_i on the multivariate vector $\mathbf{Y}$, Gamboa et al. [2013] propose to compute the aggregated Sobol' index denoted by GS_i. This sensitive index is defined as:

$$\forall i \in \{1, \ldots, p\},\ GS_i = \frac{\sum_{j=1}^{d} Var[Y_j] S_i^j}{\sum_{j=1}^{d} Var[Y_j]},$$

where S_i^j refers to the first-order Sobol' index of the scalar output Y_j with respect to the input parameter X_i. In a similar manner to the Sobol' indices in a scalar context, see Equation (3.4), a high value of this new sensitivity index indicates that the input parameter is influent and at the opposite a value close to zero designates it is not.

Campbell et al. [2006] propose to reduce the output dimension by decomposing the discretized functional QoI on a complete orthogonal basis and finally to compute sensitivity indices on each component of the decomposition. The orthogonal basis used can be estimated by using a principal component analysis from a collection of model outputs computed using different combinations of parameter values. Lastly, Lamboni et al. [2011] suggest in addition to the sensitivity indices on each principal component to compute a synthetic sensitivity index defined by Equation (3.8).

Functional principal component analysis, also known as Karhunen-Loeve decomposition, aims to project the discretized functional output on a new reduced space so that the variance can be explained by a small number of principal components, [see Anderson, 1958, Jolliffe, 1986, Besse, 1992]. The principal component analysis is used to decompose the variance-covariance matrix of the output, denoted by $\mathbf{C}$, based on the eigenvalues and eigenvectors, as:

$$\mathbf{C} = \sum_{j=1}^{d} \lambda_j \mathbf{u}_j \mathbf{u}_j^T,$$

where, $\lambda_1, \ldots, \lambda_d$ denote the eigenvalues in a decreasing order and $\mathbf{u}_1, \ldots, \mathbf{u}_d$ are orthonormal eigenvectors of $\mathbf{C}$. The discretized functional output can be decomposed such as:

$$\begin{aligned} \mathbf{Y} &= \mathbb{E}[\mathbf{Y}] + \sum_{j=1}^{d} ((\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T \mathbf{u}_j) \mathbf{u}_j \\ &= \mathbb{E}[\mathbf{Y}] + \sum_{j=1}^{d} h_j \mathbf{u}_j, \end{aligned}$$

where h_j is the j-th principal component of $\mathbf{Y}$. The multivariate output can be approximated using the $q \leq d$ first components which capture the major part of the output variance:

$$\mathbf{Y} \approx \mathbb{E}[\mathbf{Y}] + \sum_{j=1}^{q} h_j \mathbf{u}_j \cdot$$

Then, the generalized Sobol' indices can be computed based on the Sobol' indices of the first q principal components. The first-order aggregated Sobol' index for input parameter X_i is defined as:

$$GS_i \approx \frac{\sum_{j=1}^{q} \lambda_j S_i(h_j)}{\sum_{j=1}^{q} \lambda_j}, \tag{3.8}$$

where $S_i(h_j)$ denotes the Sobol' index of the j-th principal component h_j with respect to input parameter X_i.

Each one of these Sobol' indices linked to a principal component is computed using a so-called Pick and Freeze sampling scheme, [see Sobol', 1993, Janon et al., 2014]. The sensitivity indices in Equation (3.8) summarize the information on the importance of each input parameter on the functional output.

3.5 Latin Hypercube Sampling

As said previously, for computer experiments, especially global sensitivity analysis, one of the main interests is to figure out the variation of a quantity of interest with respect to the variation of some inputs [Sacks et al., 1989]. In this context, design of experiments have been developed in order to better understand the physical mechanisms governing the problem of interest by efficiently exploring the input space [Saltelli et al., 2008]. Contrary to the crude Monte Carlo sampling, which consists of s independently and identically distributed samples, the Latin Hypercube Sampling (LHS) consists in dividing the domain of each input variable in s equiprobable strata, and in sampling once from each stratum [McKay et al., 1979]. Let us consider the LHS of a random vector $\mathbf{X} = (X_1, \ldots, X_p) \in \mathcal{P} \subset \mathbb{R}^p$ and denoted by $\mathbf{P} = \{\mathbf{x}_i\}_{i=1}^{s}$. Then the forward numerical model, hereafter denoted by f, can be called for each sample in $\mathbf{P}$, such that:

$$(\mathbf{P},\ \mathbf{Y}) = (\mathbf{x}_i,\ y_i = f(\mathbf{x}_i))_{1 \leq i \leq s} \cdot$$

This stratified design allows to estimate a sample mean $m = \frac{1}{s}\sum_{i=1}^{s} y_i$ for the output $\mathbf{Y}$ with a smaller variance than the sample mean of a crude Monte Carlo sampling strategy [Stein, 1987].

Definition 2. *Let us consider the unit-hypercube $\mathcal{P} = [0,1]^p$. LHS technique consists, for $1 \leq i \leq s$, in sampling $\mathbf{x}_i = (x_{i,j})_{1 \leq j \leq p}$ as:*

$$x_{i,j} = \frac{l_{i,j} - u_{i,j}}{s},$$

where, $\mathbf{L} = (l_{i,j})_{1 \leq i \leq s,\, 1 \leq j \leq p}$ is an array where the j-th column contains a random permutation of the integers $[1, \ldots, s]$ and $U = (u_{i,j})_{1 \leq i \leq s,\, 1 \leq j \leq p}$ is an array where the j-th column contains a random vector of size s sampled from a uniform distribution on $[0,\ 1]$.

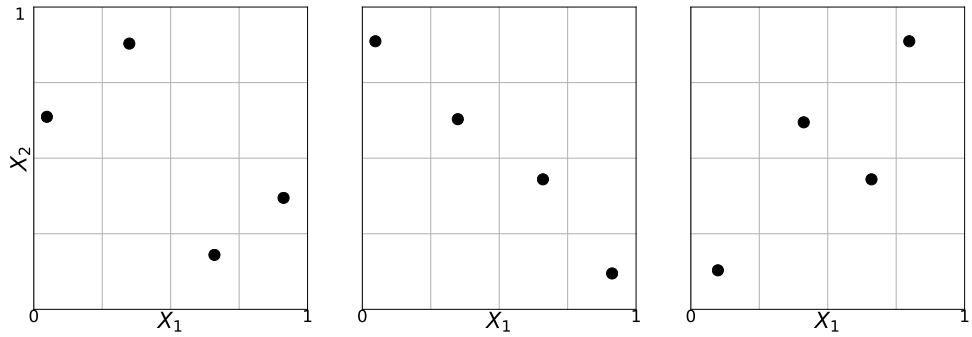


Figure 3.1 – Three examples of Latin Hypercube Sampling design of size $s = 4$ over $[0; 1]^2$, each circle represents a sample.

For example, the first LHS of Figure 3.1 (left-panel) is derived from $L = \begin{bmatrix} 4 & 2 & 3 & 1 \\ 2 & 4 & 1 & 3 \end{bmatrix}^T$. Nevertheless, LHS design can only guarantee good repartitions for one-dimensional projections and not for the other dimensions of projection, and consequently this design is not sufficient for complete space filling [Park, 1993]. Indeed, LHS technique does not always cover the input space properly, as it can be noticed with the second design (middle-panel) of Figure 3.1 which is almost diagonal. Consequently, LHS design does not reach the smallest possible variance for the estimated sample mean. To circumvent this poorly performances, Johnson et al. [1990] propose two optimality criteria based on the distance between two points. These geometrical criteria are respectively called minimax and maximin. Firstly, minimax criterion, denoted by ϕ_{mM}, minimizes the maximal distance between a point of the input space domain and the points of the design, such as:

$$\phi_{mM}(\mathbf{P}) = \max_{\mathbf{x}\in\mathcal{P}} \min_{i=1,\dots,s} ||\mathbf{x} - \mathbf{x}_i||_{l^2} \ , \tag{3.9}$$

where $||\cdot||_{l^2}$ is the Euclidean norm. LHS design of experiments which minimizes the criterion in Equation (3.9) is called minimax LHS.

The maximin LHS design is the design which maximizes the criterion called maximin and denoted by ϕ_{Mm}. It consists in maximizing the minimal distance between all points, such that:

$$\phi_{Mm}(\mathbf{P}) = \min_{1\leq i,j\leq s} ||\mathbf{x}_i - \mathbf{x}_j||_{l^2} \ .$$

In many situations, we are interested in conducting additional numerical model runs, e.g., for the validation of surrogate models such as Gaussian process regression (see Section 4.1). In this context, design augmentation generates a new design that combines with an existing one in a way that the new points maximize the space-filling properties of the combined design. A procedure to augment the number of points in a Latin Hypercube Sampling design, while preserving its Latin structure, has been developed by Carnell [2012]. The proposed method increases a LHS design by maximizing the mean distance from each point to all other points. The procedure identifies the positions of the inherited points in the new design space, find the intervals of each variable that are not represented by the inherited points, generate new points, and then fill in those underrepresented intervals for each input variable [Wang, 2003]. As an illustration, in Figure 3.2, a two dimensional maximin LHS design is augmented.

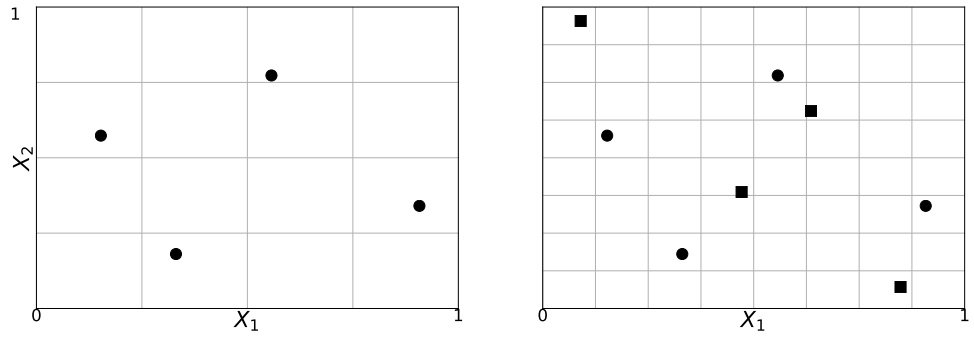


Figure 3.2 – Augmented Latin Hypercube Sampling design. The symbols circle and square represent respectively the original points obtained from a maximin LHS design and the new points based on the procedure developed by Carnell [2012].

Conclusion

In this chapter, we introduced the definition of Sobol' indices and their estimation using Monte Carlo sampling strategies. Sobol' indices rely on the decomposition of the variance of the output of a numerical model into partial variances representing the fraction of the output's variance induced by an individual input or a group of inputs. These sensitivity indices allow quantifying the influence of each input parameter on the output and also detecting any interactions between the inputs. The chapter provided the definitions of first-order, closed-order, and total sensitivity indices. Moreover, a generalization of Sobol' indices was presented in order to deal with computer codes producing a discretized functional output. The sensitivity analysis consists then in decomposing the multivariate output into some non-correlated principal components and in computing sensitivity indices on each one of the components. Then, an aggregated sensitivity index is defined in order to summarize the overall effect of each parameter (or set of input parameters) on the discretized functional output. As highlighted in this chapter, the Monte Carlo approaches for Sobol' index estimation need a high number of simulations to get accurate estimators due to the central limit theorem. In this context, we propose a space-filling strategy, based on Latin Hypercube Sampling, in order to get a better coverage of the input parameter space.

4

Gaussian process regression for global sensitivity analysis

L'analyse de sensibilité globale, basée sur l'estimation des indices de Sobol', nécessite de nombreux appels au modèle. De ce fait, elle est souvent impraticable pour les modèles coûteux en temps de calcul malgré l'utilisation de plans d'expériences avec de bonnes propriétés de remplissage de l'espace. Pour pallier ce problème computationnel, une approche couramment utilisée dans le domaine de la quantification d'incertitudes consiste à remplacer la relation entrée/sortie du modèle numérique par une approximation mathématique, appelée méta-modèle ou surface de réponse [Box and Draper, 1987]. Ce méta-modèle, dont le temps de calcul pour évaluer une réponse est négligeable, est construit à partir de quelques simulations du modèle numérique issues de différents jeux de valeurs des paramètres. Plusieurs techniques existent dans la littérature pour construire de telles approximations mathématiques, e.g., les polynômes, modèles linéaires ou additifs généralisés, processus Gaussien, réseaux de neurones, boosting d'arbres de régression, SVM [Soize and Ghanem, 2004, Smola and Schölkopf, 2004, Dreyfus, 2005, Krige, 1951, Simpson et al., 2001]. Dans ce chapitre nous nous intéressons tout particulièrement aux méta-modèles basés sur une régression par processus Gaussien [Rasmussen, 2003]. Cette approximation suppose a priori que la sortie issue du modèle d'intérêt est une réalisation d'un processus Gaussien. Comme évoqué par Marrel et al. [2008], cette formulation par processus Gaussien présente l'avantage d'obtenir les indices de sensibilité de manière analytique en y associant les incertitudes liées à l'approximation du modèle numérique par le méta-modèle. Néanmoins, ces formules sont souvent difficilement exploitables [Marrel, 2008]. Dans ce contexte, il est généralement préférable d'utiliser des méthodes d'estimation de Monte Carlo pour calculer ces indices de sensibilité en appelant directement le méta-modèle par processus Gaussien [Iooss et al., 2006]. Des méthodes, reposant sur des procédures bootstrap, permettent d'estimer l'erreur liée à l'utilisation du méta-modèle à la place du modèle d'intérêt et celle liée à l'échantillonnage de Monte Carlo impactant chaque indice de Sobol' estimé [Le Gratiet et al., 2013]. Par ailleurs, afin de modéliser un grand nombre de comportements physiques d'intérêt, nous devons nous reposer sur des codes stochastiques. A l'opposé des codes déterministes, ces modèles numériques stochastiques peuvent renvoyer des résultats différents lorsqu'ils sont appelés plusieurs fois avec le même jeu de variables d'entrée. Dans ce contexte, la modélisation par processus Gaussien bruité hétéroscédastiquement est une solution pour réaliser une analyse de sensibilité globale.

La Section 4.1 revient sur la formulation théorique de la méta-modélisation par processus Gaussien pour approcher un modèle numérique coûteux. La Section 4.2 est consacrée à l'utilisation du méta-modèle processus Gaussien pour mener l'analyse de sensibilité globale du modèle numérique d'intérêt par l'intermédiaire de l'estimation des indices de Sobol'. Pour terminer, la Section 4.3 propose une approche basée sur une régression par processus Gaussien bruité hétéroscédastiquement pour estimer les indices de Sobol' dans le cadre d'un code stochastique.

Introduction

In uncertainty quantification (UQ), Monte Carlo techniques for the estimation of Sobol' indices require a large number of numerical model evaluations, see Section 3.3, which can quickly exceed the limits of numerical resources. A widely used method to circumvent this computational issue consists in replacing the time-consuming numerical model by a mathematical approximation, called a metamodel or a surrogate model or also a surface response [Box and Draper, 1987], that simulates the behavior of the computer code and requires a lower computational cost. Constructing such a metamodel relies on a limited number of forward model responses. The limited number of points where the model of interest is evaluated is called the design of experiments (DoE), e.g., Latin Hypercube design see Section 3.5.

Several metamodels can be found in the literature, among whom Polynomial Chaos (PC) expansions [Soize and Ghanem, 2004], support vector machine [Smola and Schölkopf, 2004], artificial neural networks [Dreyfus, 2005] or Gaussian Process (GP) models [Rasmussen, 2003]. Hereafter, we will focus on the GP model which is often used in the UQ scientific field for its flexibility and prediction error quantification, [see Marrel et al., 2015b, Le Gratiet et al., 2017]. GP model, sometimes called GP regression, is equivalent to the kriging principle developed in geostatistics [Krige, 1951]. The kriging method has been firstly developed for spatial interpolation problems of a random field at unobserved locations. Then, Sacks et al. [1989] extended the concept to numerical models by considering the correlation between two outputs of a code depending on the distance between inputs. The principle of a GP model is to treat the deterministic response of the numerical code as a realization of a random Gaussian process described by its mean function and its covariance function, also called the kernel. Such kernel is a positive-definite function of two distinct input parameters allowing to define the prior covariance between any two values of the function of interest. Many kernels can be used, each one corresponding to a different set of prior assumptions made about the function of interest [Rasmussen, 2003, Stein, 2012, Duvenaud, 2014]. A kernel can incorporate a number of parameters which specify the shape of the covariance function. These parameters, also known as hyper-parameters, can be either estimated by minimizing a loss function with a leave-one-out cross-validation procedure or maximizing a likelihood function [Bachoc, 2013].

In this chapter, we will present how to approximate a numerical model based on a DoE thanks to a GP regression. In many industrial applications, we do not have direct access to the function to be approximated but only to noisy versions of it. It is the case when we are dealing with a stochastic simulation. In this chapter, we highlight the fact that kriging model can be adapted to such noisy observations.

In some cases, as mentioned by Chen et al. [2005], by using GP regression, analytical formulae can be available for Sobol' index computation, avoiding the necessity to use a Monte Carlo scheme [Homma and Saltelli, 1996, Jansen, 1999, Monod et al., 2006]. In this latter context, a method giving confidence intervals for the Sobol' index estimates and taking into account both the metamodel uncertainty and the numerical errors on the Sobol' index estimations is suggested in [Le Gratiet et al., 2013]. They consider a sampling strategy to estimate the Sobol' indices and they infer the sampling errors thanks to a bootstrap procedure as proposed by Archer et al. [1997]. Then by considering the numerical model as a realization of a GP model, it is possible to take into account the metamodel error in the estimation of Sobol' indices.

The first section explains theoretically the metamodeling procedure using Gaussian process regression. Then, in Section 4.2 we present a kriging-based sensitivity analysis relying on the estimation of Sobol' indices. Finally in Section 4.3, Gaussian process metamodel framework with heteroscedastic noise is proposed in order to estimate Sobol' index in the context of a stochastic numerical model.

4.1 Theoretical formulation

Let us assume a deterministic real-valued function of the d-dimensional input parameter vector $\mathbf{x} = (x_1, \ldots, x_p) \in \mathcal{P} \to \mathbb{R}$ defined as:

$$\begin{array}{ll} \mathcal{P} \subset \mathbb{R}^p & \to \mathbb{R} \\ \mathbf{x} = (x_1, \ldots, x_p) & \longmapsto y(\mathbf{x}) \end{array} . \tag{4.1}$$

By considering the design of experiments (DoE), denoted by $\mathbf{X}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where the numerical model in Equation (4.1) has been evaluated, and by denoting by $\mathbf{y}^n = y(\mathbf{X}^n)$ the values of $y(\mathbf{x})$ at points in $\mathbf{X}^n$. Gaussian process (GP) regression treats the deterministic response $y(\mathbf{x})$ as a realization of a Gaussian stochastic process $Y(\mathbf{x})$, including a regression part and a centered square-integrable process. This random process can be written as:

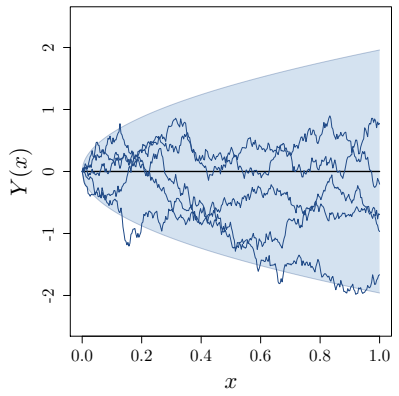
$$Y(\mathbf{x}) = \mu(\mathbf{x}) + Z(\mathbf{x}) \,,$$

where $Z(\mathbf{x})$ is a centered stationary Gaussian process of known covariance kernel $C : (\mathbf{u}, \mathbf{v}) \in \mathcal{P}^2 \to C(\mathbf{u}, \mathbf{v}) \in \mathbb{R}$. The deterministic function $\mu(\mathbf{x})$ approximates the trend of the observations with respect to the inputs and the covariance structure defines the prior dependence between the different values of the computer code responses. Examples of Gaussian process regression model are given in Figure 4.1. Various samples path of Gaussian processes are represented by considering different means and covariance functions in a 1-D or 2-D input space. For the multidimensional input space representation (last panel), only one sample obtained from the Gaussian process with a zero mean function and a Matérn 5/2 covariance function is displayed.

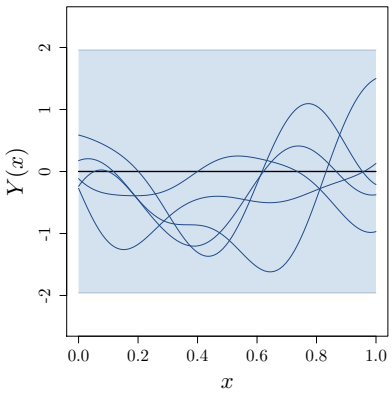
In the literature, it is a common practice to consider the deterministic regression part of the Gaussian stochastic process as a linear combination of elementary functions. In this context, $\mu(\mathbf{x})$ can be written as:

$$\mu(\mathbf{x}) = \sum_{j=0}^{k} \beta_j \mathrm{f}_j(\mathbf{x}) = \mathbf{F}(\mathbf{x})\boldsymbol{\beta} \,, \tag{4.2}$$

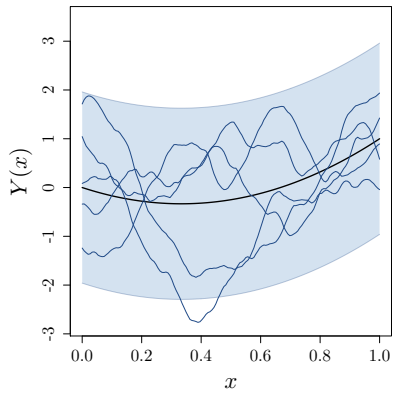
where $\mathbf{F}(\mathbf{x}) = [\mathrm{f}_0(\mathbf{x}), \ldots, \mathrm{f}_k(\mathbf{x})]$ is a vector of fixed basis functions, and $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_k]^T$ is the regression coefficient vector. In our study, we have decided to use first order polynomial to model the trend. As mentioned by Martin and Simpson [2005] and later by Marrel et al. [2008], such one-degree polynomial function is sufficient, and sometimes mandatory, in order to capture the global trend of the numerical model. Therefore, the deterministic function can be written as $\mu(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$.



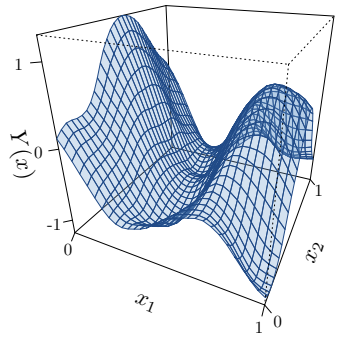
(a) $\mu(x) = 0$ and Brownian covariance function



(b) $\mu(x) = 0$ and Gaussian covariance function



(c) $\mu(x) = -2x + 3x^2$ and Matérn 3/2 covariance function



(d) $\mu(x) = 0$ and Matérn 5/2 covariance function

Figure 4.1 – Different sample paths of Gaussian processes (blue lines) considering various means and covariance functions. For the first three panels, the black line shows the deterministic mean function μ and the shaded area corresponds to 95% confidence intervals. For the last panel, only one sample is considered.

The covariance function of the stochastic part $Z(\mathbf{x})$ is hereafter assumed stationary such as $C(\mathbf{u},\mathbf{v}) = \sigma^2 R(\mathbf{u}-\mathbf{v};\boldsymbol{\theta})$. It is parameterized by the vector $\boldsymbol{\theta}$ and the process variance σ^2. Our study is focused on a family of correlation functions that can be written as a product of one-dimensional correlation kernel:

$$C(\mathbf{u},\mathbf{v}) = \sigma^2 R(\mathbf{u}-\mathbf{v};\boldsymbol{\theta}) = \sigma^2 \prod_{l=1}^{p} g(u_l - v_l;\theta_l)\,\cdot \tag{4.3}$$

A non-exhaustive list of one-dimensional stationary covariance kernels g is presented in Table 4.1. For further details on the available covariance kernels, the reader is referred to the work of Sacks et al. [1989] or more recently of Rasmussen [2003], where authors give a complete review of covariance functions with their drawbacks and advantages. The choice of the covariance kernel is a crucial aspect of a GP regression. Indeed, the covariance function will allow to control the level of smoothness for the GP.

Name	Formula
squared exponential[1]	$g(u-v;\theta) = \sigma^2 \exp\left(-\dfrac{(u-v)^2}{2\theta^2}\right)$
Matérn 5/2	$g(u-v;\theta) = \sigma^2 \left(1 + \dfrac{\sqrt{5}\lvert u-v\rvert}{\theta} + \dfrac{5\lvert x-y\rvert^2}{3\theta^2}\right) \exp\left(-\dfrac{\sqrt{5}\lvert u-v\rvert}{\theta}\right)$
Matérn 3/2	$g(u-v;\theta) = \sigma^2 \left(1 + \dfrac{\sqrt{3}\lvert u-v\rvert}{\theta}\right) \exp\left(-\dfrac{\sqrt{3}\lvert u-v\rvert}{\theta}\right)$
exponential	$g(u-v;\theta) = \sigma^2 \exp\left(-\dfrac{\lvert u-v\rvert}{\theta}\right)$
cosine	$g(u-v;\theta) = \sigma^2 \cos\left(\dfrac{u-v}{\theta}\right)$

[1] Also known as Gaussian kernel, exponentiated quadratic or radial basis function.

Table 4.1 – Examples of one-dimensional stationary kernels.

Depending on the observations obtained from the DoE, two frameworks of GP regressions can be derived. The first one consists in considering that the observations are noise-free. At the opposite, in the second one, we consider that the data are tainted by a white noise.

The noise-free case y is modeled by the conditional Gaussian process $\{Y_n(\mathbf{x})\,,\ \mathbf{x}\in\mathcal{P}\} := \{[Y(\mathbf{x})|Y(\mathbf{X}^n)=\mathbf{y}^n]\,,\ \mathbf{x}\in\mathcal{P}\}$, where $\mathbf{y}^n = (y(\mathbf{x}_1),\ldots,y(\mathbf{x}_n))$. We then get, for any $\mathbf{x}\in\mathcal{P}$, the ordinary kriging equations:

$$Y_n(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), s^2(\mathbf{x})) \tag{4.4}$$

with

$$m(\mathbf{x}) = \mu(\mathbf{x}) + \mathbf{c}(\mathbf{x})^T \mathbf{C}^{-1}(\mathbf{y}^n - \boldsymbol{\mu})\,, \tag{4.5}$$

$$s^2(\mathbf{x}) = C(\mathbf{x},\mathbf{x}) - \mathbf{c}(\mathbf{x})^T \mathbf{C}^{-1}\mathbf{c}(\mathbf{x})\,. \tag{4.6}$$

We denote by $\boldsymbol{\mu} = \mu(\mathbf{X}^n)$ the vector of trend values on $\mathbf{X}^n$, by $\mathbf{C} = (C(\mathbf{x}_i, \mathbf{x}_j))_{1 \leq i,j \leq n}$ the covariance matrix of $Y(\mathbf{X}^n)$, and by $\mathbf{c}(\mathbf{x}) = (C(\mathbf{x}, \mathbf{x}_i))_{1 \leq i \leq n}$ the vector of covariances between $Y(\mathbf{x})$ and $Y(\mathbf{X}^n)$.

We present in Figure 4.2 an example of GP regression in a noise-free framework. We notice from this figure that the GP regression mean function m interpolates the data points from the DoE.

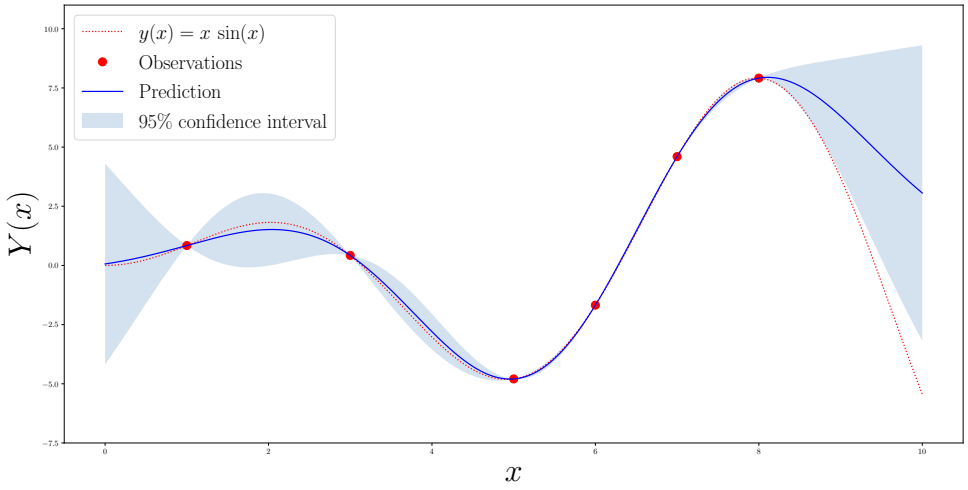


Figure 4.2 – Gaussian process regression with noise-free observations and a radial basis covariance function, see Table 4.1. The variance parameter equals $\sigma^2 = 1$, the length scale parameter $\theta = 10$ and the mean $\mu(x)$ is null. The dashed pink line represents the function of interest $f(x) = x \sin(x)$, the red circles represent the noise-free observations, the black line represents the GP regression mean $m(x)$, and the shaded area corresponds to 95% confidence intervals.

The noisy case In many industrial cases, exact evaluations of y cannot be obtained directly from the DoE. We have, for each $i = 1, \ldots, n$, a noisy evaluation $\tilde{y}_i = y(\mathbf{x}_i) + \varepsilon_i$. Where, ε_i is a centered noise with the corresponding noise variance τ_i^2, i.e., $\varepsilon_i \sim \mathcal{N}(0, \tau_i^2)\,(1 \leq i \leq n)$. We then consider, as a first approximation, that the vector $(\varepsilon_1, \ldots, \varepsilon_n)$ is a centered Gaussian random vector with diagonal covariance matrix $\mathrm{diag}(\tau_1^2, \ldots, \tau_n^2)$ denoted by $\boldsymbol{\Delta}$. Provided that the process Y and the Gaussian measurement errors ε_i are stochastically independent, the process Y is still Gaussian conditionally on the heteroscedastic noisy observations $\tilde{y}_i$ and its conditional mean and variance function are given by the following slightly modified kriging equations:

$$Y_n(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), s^2(\mathbf{x}))\,,$$

with,

$$m(\mathbf{x}) = \mu(\mathbf{x}) + \mathbf{c}(\mathbf{x})^T (\mathbf{C} + \boldsymbol{\Delta})^{-1} (\tilde{\mathbf{y}} - \boldsymbol{\mu})\,, \tag{4.7}$$

$$s^2(\mathbf{x}) = C(\mathbf{x}, \mathbf{x}) - \mathbf{c}(\mathbf{x})^T (\mathbf{C} + \boldsymbol{\Delta})^{-1} \mathbf{c}(\mathbf{x})\,\cdot \tag{4.8}$$

We present in Figure 4.3 an example of GP regression in a noisy case. We can notice that in the noisy framework, the GP regression mean function does not interpolate the observations and the variance is no longer equal to zero at the observations.

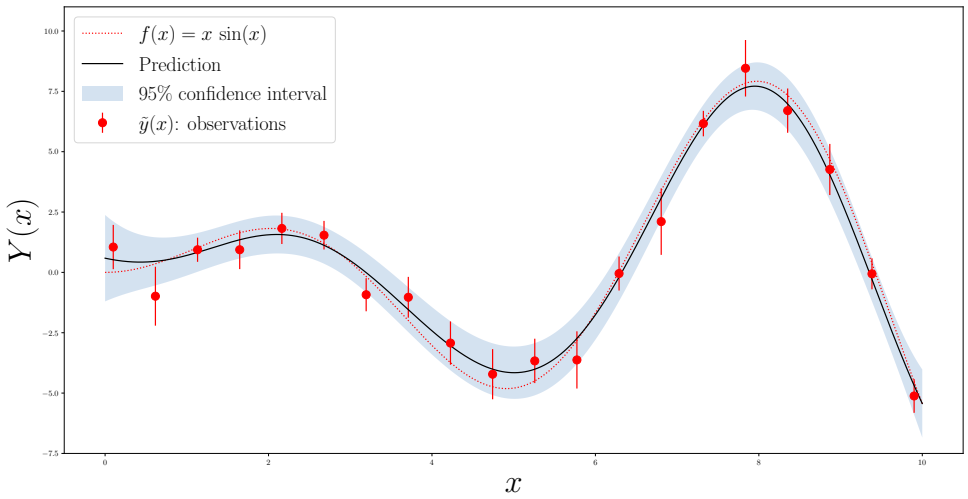


Figure 4.3 – Gaussian process regression with heteroscedastic noisy observations and a radial basis covariance function, see Table 4.1. The variance parameter equals $\sigma^2 = 2.43^2$, the length scale parameter $\theta = 1.65$ and the mean is null. The dashed pink line represents the function of interest $f(x) = x\sin(x)$, the red circles represent the noisy observations, the black line represents the Gaussian process regression mean, and the shaded area corresponds to 95% confidence intervals.

To compute the mean and variance of a GP regression model, see Equation (4.8), the estimation of several parameters is needed. Firstly, the kernels rely on some intrinsic parameters which are usually referred to as hyper-parameters. These hyper-parameters specify the precise shape of the covariance function. Kriging model, with a regression part defined as in Equation (4.2), is characterized by the regression parameter vector $\boldsymbol{\beta}$. In the literature two approaches are commonly used [Bachoc, 2013]. Firstly, a solution for hyper-parameters estimation consists in minimizing a loss function with a leave-one-out cross-validation scheme. On the other hand, we can estimate these parameters by maximum likelihood or maximum a posteriori not described hereafter.

Cross-validation estimate Let us assume hereafter the noise-free case previously described. In order to choose the hyper-parameters, a natural approach is to compare the error from the prediction of various kriging models and finally to select the one with the lowest error. Cross-validation procedure consists in splitting the DoE into two disjoint sets, one dedicated to training and the other one to estimate the performance of the surrogate model.

The principle of cross-validation is usually based on the k-fold setting where the DoE is split into k disjoint and equally sized subsets. The validation of the kriging model is done on a single subset and the training is performed based on the union of the remaining $k-1$ subsets [Rasmussen, 2003]. This procedure is then repeated k times, each time with a different subset for validation, in order to compute an averaged error from a loss function. A particular and popular k-fold cross-validation in GP regression is when $k = n$ (n is the length of the DoE). This procedure is known as leave-one-out cross-validation. Any loss function can be used with such approach but a usual one is the mean squared

error loss function. This error in leave-one-out cross-validation is defined as:

$$MSE_{LOO} = \frac{1}{n}\sum_{i=1}^{n}(m_i(\mathbf{x}_i) - y(\mathbf{x}_i))^2\,,$$

where m_i is the mean predictor of the kriging model trained on all points from the DoE except the i^{th} one and $y(\mathbf{x}_i)$ is the observation for the i^{th} point.

As it can be seen in Equation (4.5), the variance parameter σ^2 has no influence on the mean predictor of the kriging model. Consequently, minimizing the MSE_{LOO} cannot provide an estimate for this parameter. Nevertheless, as proposed by Bachoc [2013], a leave-one-out cross-validation criterion can also be derived in order to estimate the variance hyper-parameter.

Maximum likelihood estimate For estimating the parameters of a GP regression model, a commonly used numerical procedure is based on maximum likelihood estimation [Fang et al., 2005]. The idea is to quantify the adequacy between model realizations and a distribution. Let us assume a noise-free GP regression model parameterized by the vector of regression coefficients $\boldsymbol{\beta}$, the kriging variance σ^2 and the autocorrelation parameters $\boldsymbol{\theta}$, such as the mean function and the vector of trend values in Equation (4.5) are respectively defined as :

$$\mu(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\,,$$

and

$$\boldsymbol{\mu} = \mathbf{F}\boldsymbol{\beta}\,,$$

where $\mathbf{f}(\mathbf{x})$ is the vector of trend values at $\mathbf{x}$ and $\mathbf{F} = (\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_n))^T$ is the experimental matrix.

Considering a covariance kernel as in Equation (4.3) we have $\mathbf{C} = \sigma^2 R(\mathbf{u} - \mathbf{v}; \boldsymbol{\theta}) = \sigma^2 \mathbf{R}_{\boldsymbol{\theta}}$, where $\mathbf{R}_{\boldsymbol{\theta}}$ depends only on $\boldsymbol{\theta}$. Due to the Gaussian assumption, the likelihood for parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and σ can be written as:

$$\mathcal{L}(\mathbf{y}^n|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}\sqrt{\det\mathbf{R}_{\boldsymbol{\theta}}}}\exp\left(-\frac{1}{2}\frac{(\mathbf{y}^n - \mathbf{F}\boldsymbol{\beta})^T\mathbf{R}_{\boldsymbol{\theta}}^{-1}(\mathbf{y}^n - \mathbf{F}\boldsymbol{\beta})}{\sigma^2}\right),$$

Given the correlation parameters $\boldsymbol{\theta}$, the maximum likelihood estimator of $\boldsymbol{\beta}$ and of σ^2 are respectively:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\mathbf{F}^T\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{F})^{-1}\ \ \mathbf{F}^T(\mathbf{R}_{\boldsymbol{\theta}})^{-1}\mathbf{y}^n\,,$$

and,

$$\widehat{\sigma}^2(\boldsymbol{\theta}) = \frac{1}{n}(\mathbf{y}^n - \mathbf{F}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))^T\mathbf{R}_{\boldsymbol{\theta}}^{-1}(\mathbf{y}^n - \mathbf{F}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))\,\cdot$$

By replacing the vector of regression coefficients and the process variance by their optimal values respectively $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\widehat{\sigma}^2(\boldsymbol{\theta})$, the likelihood can be written as:

$$\mathcal{L}(\boldsymbol{\theta}) = (2\pi\widehat{\sigma}^2(\boldsymbol{\theta}))^{-\frac{n}{2}}(\det\mathbf{R}_{\boldsymbol{\theta}})^{-\frac{1}{2}}\exp(-\frac{n}{2})\,\cdot$$

Due to the fact that the likelihood can take extremely small values, it is often helpful to consider log-likelihood to avoid numerical issues defined as:

$$\log\mathcal{L}(\boldsymbol{\theta}) = -\frac{n}{2}\log\Big(\widehat{\sigma}^2(\boldsymbol{\theta})(\det\mathbf{R}_{\boldsymbol{\theta}})^{\frac{1}{n}}\Big) - \frac{n}{2}\Big(\log(2\pi) + 1\Big)\,\cdot \tag{4.9}$$

Then, an optimization can be performed by minimizing the opposite of the log-likelihood function defined in Equation (4.9) regarding the hyperparameter $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\mathcal{D}_{\boldsymbol{\theta}}}{\operatorname{argmin}} \Big\{ -\log\mathcal{L}(\boldsymbol{\theta}) \Big\},$$

where $\mathcal{D}_{\boldsymbol{\theta}}$ is the domain of definition for the admissible values of $\boldsymbol{\theta}$.

Thus, the estimation of $\boldsymbol{\theta}$ can be reduced to the following numerical optimization:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}\in\mathcal{D}_{\boldsymbol{\theta}}}{\operatorname{argmin}} \Big\{ \log\Big(\widehat{\sigma}^2(\boldsymbol{\theta}) (\det \mathbf{R}_{\boldsymbol{\theta}})^{\frac{1}{n}} \Big) \Big\}.$$

4.2 Kriging-based Sobol' indices

In this section, we present a methodology to estimate sensitivity indices using a Gaussian process (GP) regression model. The Sobol' index produced by Monte Carlo estimation using a kriging surrogate model is tainted with a twofold error. Firstly, a sampling error from the use of a Monte Carlo sampling procedure and then a metamodel error due to the fact that the numerical model of interest is substituted by a surrogate model. Nevertheless, in order to make a rigorous global sensitivity analysis, it is important to assess the impact of these two combined errors on the estimated Sobol' indices. We focus on the kriging-based sensitivity analysis, developed by Le Gratiet et al. [2013], allowing to take into account both the sampling error and the metamodel one, assumed to have no interaction.

Hereafter, we are interested in the closed Sobol' index presented in Section 3.2 but the methodology can be extended to other indices, e.g., total effect. As suggested by Marrel et al. [2009] and later by Le Gratiet et al. [2013], the idea is to substitute the numerical model in Equation (3.5) with the Gaussian process Y_n defined in Equation (4.4), such as:

$$\tilde{S}_{\mathbf{u}} = \frac{Var_{\mathbf{X}_{\mathbf{u}}}(\mathbb{E}_{\mathbf{X}_{-\mathbf{u}}}[\, Y_n(\mathbf{X}) \mid \mathbf{X}_{\mathbf{u}}])}{Var_{\mathbf{X}}(Y_n(\mathbf{X}))}.$$

As Y_n is a random process, the resulting indices are also random. In a similar fashion, we can substitute in the estimator of the closed Sobol' index the numerical model responses $y(\mathbf{x})$ by the Gaussian process Y_n, such as:

$$\widehat{\tilde{S}}_{\mathbf{u}} = \frac{\frac{1}{s}\sum_{i=1}^{s} Y_n(\mathbf{x}_i)\, Y_n(\mathbf{x}_i^{\mathbf{u}}) - (\frac{1}{s}\sum_{i=1}^{s} Y_n(\mathbf{x}_i))\,(\frac{1}{s}\sum_{i=1}^{s} Y_n(\mathbf{x}_i^{\mathbf{u}}))}{\frac{1}{s}\sum_{i=1}^{s} Y_n(\mathbf{x}_i)^2 - (\frac{1}{s}\sum_{i=1}^{s} Y_n(\mathbf{x}_i))^2}, \tag{4.10}$$

where the designs $\mathbf{P} = \{\mathbf{x}_i\}_{i=1}^{s}$ and $\mathbf{P}^{\mathbf{u}} = \{\mathbf{x}_i^{\mathbf{u}}\}_{i=1}^{s}$ are the ones introduced in Section 3.3.

In order to estimate the sampling error due to the Monte Carlo estimation and the one due to the substitution of $y(\boldsymbol{x})$ by a kriging model, we have to investigate the distribution of the estimator described in Equation (4.10). The method to compute the distribution of the estimator is described in [Le Gratiet et al., 2013] and presented in Algorithm 2.

The sample $\left(\widehat{\tilde{s}}_{\mathbf{u}}^{k,l}\right)_{\substack{k=1,\dots,N_Y \\ l=1,\dots,B}}$ of $\widehat{\tilde{S}}_{\mathbf{u}}$ obtained from Algorithm 2 is of size $N_Y \times B$. From this output, we can obtain an estimate, denoted by $\overline{\tilde{S}}_{\mathbf{u}}$, for $\mathbb{E}[\widehat{\tilde{S}}_{\mathbf{u}}]$:

$$\overline{\tilde{S}}_{\mathbf{u}} = \frac{1}{N_Y B} \sum_{\substack{k=1,\dots,N_Y \\ l=1,\dots,B}} \widehat{\tilde{s}}_{\mathbf{u}}^{k,l}.$$

Algorithm 2: Evaluation of the distribution of $\widehat{\widetilde{S}}_{\mathbf{u}}$ in Equation (4.10) adapted from [Le Gratiet et al., 2013]

Build $Y_n(\mathbf{x})$ from the n observations $\mathbf{y}^n$ of $y(\mathbf{x})$ at points in $\mathbf{X}^n$ (see Equation (4.4)).
Generate two designs $\mathbf{P}$ and $\mathbf{P}^{\mathbf{u}}$ from independent random vectors distributed according to the input vector (see Section 3.3).
Set N_Y the number of samples for $Y_n(\mathbf{x})$ and B the number of bootstrap samples for evaluating the Monte Carlo integrations.
for $k = 1, \dots, N_Y$ **do**
  Sample a realization $y_n(\mathbf{x})$ of the Gaussian process $Y_n(\mathbf{x})$ for each $\mathbf{x} \in \left\{ \{\mathbf{x}_i\}_{i=1}^s, \{\mathbf{x}_i^{\mathbf{u}}\}_{i=1}^s \right\}$.
  Compute $\widehat{\widetilde{s}}_{\mathbf{u}}$ the k^{th} realization of $\widehat{\widetilde{S}}_{\mathbf{u}}$ using Equation (4.10) with the realization $y_n(\mathbf{x})$.
  for $l = 2, \dots, B$ **do**
    Sample with replacements two set of samples of size s, $\mathbf{v}$ and $\tilde{\mathbf{v}}$ respectively from $\{\mathbf{x}_i\}_{i=1}^s$ and $\{\mathbf{x}_i^{\mathbf{u}}\}_{i=1}^s$.
    Compute $\widehat{\widetilde{s}}_{\mathbf{u}}^{k,l}$ from $y_n(\mathbf{x})$ with $\mathbf{x} \in \{\mathbf{v}, \tilde{\mathbf{v}}\}$.
  return The sample $\left(\widehat{\widetilde{s}}_{\mathbf{u}}^{k,l}\right)_{\substack{k=1,\dots,N_Y \\ l=1,\dots,B}}$

In a similar manner, the variance of $\widehat{\widetilde{S}}_{\mathbf{u}}$ can be estimated by:

$$\widehat{\sigma}^2(\widehat{\widetilde{S}}_{\mathbf{u}}) = \frac{1}{N_Y B - 1} \sum_{\substack{k=1,\dots,N_Y \\ l=1,\dots,B}} \left(\widehat{\widetilde{s}}_{\mathbf{u}}^{k,l} - \overline{\widetilde{S}}_{\mathbf{u}} \right).$$

This variance takes both into account the uncertainty of the Monte Carlo integrations and the one of the kriging model approximation. We can firstly estimate the part of variance related to the kriging model such as:

$$\widehat{\sigma}^2_{Y_n}(\widehat{\widetilde{S}}_{\mathbf{u}}) = \frac{1}{B} \sum_{l=1}^{B} \frac{1}{N_Y - 1} \sum_{k=1}^{N_Y} \left(\widehat{\widetilde{s}}_{\mathbf{u}}^{k,l} - \frac{\sum_{i=1}^{N_Y} \widehat{\widetilde{s}}_{\mathbf{u}}^{i,l}}{N_Y} \right).$$

Moreover, the part of the variance due to the Monte Carlo integrations can be estimated with:

$$\widehat{\sigma}^2_{MC}(\widehat{\widetilde{S}}_{\mathbf{u}}) = \frac{1}{N_Y} \sum_{k=1}^{N_Y} \frac{1}{B - 1} \sum_{l=1}^{B} \left(\widehat{\widetilde{s}}_{\mathbf{u}}^{k,l} - \frac{\sum_{i=1}^{B} \widehat{\widetilde{s}}_{\mathbf{u}}^{k,i}}{B} \right).$$

We neglect the part of variance due to a potential interaction between Monte Carlo integrations and metamodeling.

4.3 GP-based Sobol' indices for stochastic numerical model.

In this section, we do not deal with a deterministic numerical model, but with a stochastic one, i.e., when the same set of inputs leads to different output values. Stochas-

tic numerical models are often needed to properly model some physical phenomena, e.g., acoustic wave propagation in turbulent fluids [Iooss et al., 2002] or atmospheric pollution [Reich et al., 2012]. Such models are governed by some intrinsic alea, which is described as an uncontrollable random input variable denoted by V. Hereafter, V is considered as a random field whose each realization is governed by a random seed value. Let us consider the following stochastic numerical model:

$$f : \begin{array}{ll} \mathcal{P} \subset \mathbb{R}^p & \to \mathbb{R} \\ \mathbf{x} & \mapsto Y = f(\mathbf{x}, V) \end{array} ,$$

where $\mathbf{x}$ is the controllable input parameter vector of size p that belongs to the input space $\mathcal{P}$, Y is the quantity of interest obtained from the stochastic numerical model f.

In the context of global sensitivity analysis, each input parameter is now considered as a random variable X_j with its uncertainty modeled by a probability distribution, such as $\mathbf{X} = (X_1, \ldots, X_p)$. These one-dimensional probability distributions reflect the practitioner's belief in the uncertainty on the parameter values and the X_j's are assumed to be mutually independent. Moreover, in all this chapter, we assume that $\mathbf{X}$ and V are independent. As highlighted by Hart et al. [2017], global sensitivity analysis, especially Monte Carlo estimation of Sobol' indices [Homma and Saltelli, 1996, Jansen, 1999, Monod et al., 2006], is not trivial when dealing with a stochastic numerical code. In the literature, the different treatments of the inherent randomness have led to the formulation of various Sobol' index extensions to stochastic models. We focus on the extension which relies on the elimination of the internal randomness by representing the probability distribution of the random output Y with quantitative measures, such as variance [Iooss and Ribatet, 2009] or quantiles [Browne et al., 2016]. In the study, we consider the mean value of Y relative to the intrinsic randomness of the code. As a result, the stochastic simulator is reduced to a deterministic function and the closed Sobol' index defined in Equation (3.5) can be reformulated:

$$\tilde{S}_{\mathbf{u}} = \frac{Var_{\mathbf{X_u}}(\mathbb{E}_{\mathbf{X_{-u}}}[\mathbb{E}_V[f(\mathbf{X}, V)]|\mathbf{X_u}])}{Var_{\mathbf{X}}(Y)} .$$

We usually replace the mean quantity by its empirical mean. Estimation of the mean is based on Monte Carlo sampling and consequently consists in repeating calculations with the same sets of controllable inputs $\mathbf{x}$. Let us consider that for each exploration in the input space, the simulator is run K-times to properly discard the intrinsic randomness by estimating the expectation from the output samples.

By adapting the formalism of Sobol' in [Sobol', 1993] presented in Section 3.3 to our stochastic numerical model context, we get an estimate of the closed Sobol' index:

$$\widehat{\tilde{S}}_{\mathbf{u}} = \frac{\frac{1}{s}\sum_{i=1}^{s}\frac{1}{K}\sum_{k=1}^{K} y_{i,k}\, y^{\mathbf{u}}_{i,k} - \left(\frac{1}{s}\sum_{i=1}^{s}\frac{1}{K}\sum_{k=1}^{K} y_{i,k}\right)^2}{\frac{1}{s}\sum_{i=1}^{s}\left(\frac{1}{K}\sum_{k=1}^{K} y_{i,k}\right)^2 - \left(\frac{1}{s}\sum_{i=1}^{s}\frac{1}{K}\sum_{k=1}^{K} y_{i,k}\right)^2} . \tag{4.11}$$

The major drawback of this procedure is that it may be time consuming due to the combination of sampling and replication. The total number of calls to the code to compute the estimator in Equation (4.11) is equal to $Ks(p+1)$, with p the number of uncertain input parameters. As mentioned previously when dealing with deterministic code, a popular solution to avoid this computational issue consists in replacing the numerical model by a metamodel. Nevertheless, for stochastic numerical models, classical metamodel procedures are not pertinent anymore if the inherent randomness of the stochastic code is not

taken into account. To overcome this limitation, we propose the use of an heteroscedastic GP regression model to provide an efficient metamodel of $\mathbb{E}_V[f(\mathbf{X}, V)|\mathbf{X}]$ [Ginsbourger et al., 2008].

In our context, exact evaluations of the expectation $\mathbb{E}_V[f(\mathbf{X}, V)|\mathbf{X}]$ of the stochastic numerical model cannot be obtained directly. We rather compute:

$$\begin{aligned}\forall i = 1, \ldots, n,\ \tilde{y}_i &= \frac{1}{K}\sum_{k=1}^{K} f(\mathbf{x}_i, V = v_k)\,, \\ &= \mathbb{E}_V[f(\mathbf{X}, V)|\mathbf{X} = \mathbf{x}_i] + \varepsilon_i\,.\end{aligned}$$

We model $(\varepsilon_i)_{1\leq i\leq n}$ by a Gaussian vector with independent components with mean zero and variance τ_i^2 defined as:

$$\tau_i^2 = \frac{1}{K}\left(\frac{1}{K-1}(\sum_{k=1}^{K}(f(\mathbf{x}_i, V = v_k) - \frac{1}{K}\sum_{k=1}^{K} f(\mathbf{x}_i, V = v_k))^2\right)\,.$$

We then approximate y by a conditional Gaussian process $\{[Y(\mathbf{x})|Y(\mathbf{X}^n) = \mathbf{y}^n]\,,\ \mathbf{x} \in \mathcal{P}\}$ and we assume that the process Y is independent from the observation noise. We subsequently use formulae in Equations (4.7) and (4.8) with $\mathbf{\Delta}$ the diagonal matrix $\mathrm{diag}(\tau_1^2, \ldots, \tau_n^2)$.

Conclusion

In this chapter, we present a metamodeling approach, based on Gaussian process, to alleviate the computational cost of Sobol' index estimation. Indeed, high-fidelity numerical models are often very greedy in terms of computing time to be directly used to conduct global sensitivity analysis based on Monte Carlo methods. Gaussian process metamodeling, also known as kriging, has been widely used to perform sensitivity analysis. The proposed methodology provides an estimate and a confidence interval for each Sobol' index. We adapt the estimator defined by Equation (4.10) to deal with our stochastic simulator, metamodeling the expectation with respect to the inherent randomness by a heteroscedastic Gaussian process.

5 Model calibration

La calibration de modèle consiste à estimer des paramètres de code numérique. La pertinence des paramètres d'entrée lors de l'utilisation de modèles numériques s'avère être un aspect déterminant. En effet, le modèle peut prendre en compte de nombreuses physiques et être le plus complexe possible, si les paramètres qui le constituent sont faux alors la simulation n'aura aucune valeur. Ces méthodes d'estimation de paramètres peuvent être classées en deux groupes : les méthodes fréquentistes et Bayésiennes [Kantas et al., 2009]. Les méthodes développées dans ce chapitre sont basées sur des algorithmes de filtrage Bayésien issues du domaine de l'assimilation de données [Moireau and Chapelle, 2011, Olivier and Smyth, 2017, Nemeth et al., 2013]. Ces méthodes d'inférence présentent l'intérêt de traiter des observations obtenues en continu. Par ailleurs, ces algorithmes de filtrage peuvent être facilement exécutés en parallèle et ne nécessitent généralement pas le gradient de la fonction à minimiser, ce qui est d'un grand intérêt pour les codes numériques "boites noires" coûteux en temps de calcul. En effet, du fait de la complexité de développement, afin que les codes soient les plus représentatifs de la réalité, leur exécution peut prendre un temps non négligeable. Ce temps d'exécution est grandement diminué par rapport à l'expérimentation réelle, néanmoins, il reste un critère décisif dans le choix des méthodes d'inférence utilisées dans le domaine de la calibration de modèle numérique. Historiquement, les méthodes d'assimilation de données ont été conçues afin d'améliorer les prévisions en météorologie. L'objectif initial de ces approches d'inférence était de mettre à jour l'estimation de l'état d'un modèle afin de produire des prévisions plus précises. Cependant, les travaux présentés montreront que l'assimilation de données peut être utilisée pour la calibration de modèles. La Section 5.1 conceptualise le principe de l'inférence Bayésienne et présente la vision récursive qui en découle. Dans la Section 5.2, deux méthodes d'assimilation de données reposant sur le filtre de Kalman sont présentées, sa forme classique proposée par Kalman [1960] et sa variante dite de Monte Carlo proposée par Evensen [1994]. La Section 5.3 généralise ces méthodes d'assimilation de données, initialement développées pour l'amélioration de l'estimation de l'état d'un modèle, à l'inférence de paramètres d'entrée considérés comme statiques.

Introduction

In many fields of engineering, estimating model parameters from measurements is a crucial problem and several approaches have been developed. We focus on nonlinear model calibration problems, also known as inverse problems. Over the last few years, parameter estimation has been a subject of studies [Kanso et al., 2006, Carmassi et al., 2018, Rubio et al., 2018]. In this domain, two approaches are competing [Kantas et al., 2009]. On the one hand, there are the frequentist approaches in which parameters are assumed to be deterministic but unknown. An estimation of these unknown parameters can be done through a statistical minimization of the error between the measurements and model outputs. On the other hand, Bayesian approaches assume that the unknown parameters are modeled using probability distributions. These Bayesian model calibration techniques use measurements to update some prior probability distribution [Kennedy and O'Hagan, 2001, Tarantola, 2005].

One can categorize these methods as off-line if the data are processed in batches of observations or online if the data are sequentially processed when new observations become available. In the literature, Bayesian model calibration has been mainly applied in a batch manner by typically using Markov Chain Monte Carlo methods (MCMC). The best-known MCMC method is the Metropolis–Hastings algorithm [Hastings, 1970]. Such inference methods can require thousands of sampling points before achieving the convergence of the posterior. This requirement can quickly be expensive due to the fact that each sampling point corresponds to an evaluation of the numerical model. To overcome this computational burden, a possible solution is to use a surrogate model to replace the costly model, [see Marzouk et al., 2007, Yan and Zhou, 2019]. Recent decades have been marked by a simultaneous development of sensor technologies and high-fidelity numerical modeling capabilities. At the intersection of these two advances lies an interesting evolution of real-time monitoring of a system. Consequently, our research efforts have been mostly directed toward online techniques. In this context, the model calibration can be carried out using an inference procedure based on sequential Bayesian techniques.

In geosciences, these sequential Bayesian techniques are called data assimilation (DA) methods. DA techniques allow to combine all sources of information available from observations and numerical models [Blayo et al., 2014]. These approaches have been extensively used to approximate the state of systems from noisy observations in many applications such as oceanography, weather forecasting, seismology or finance [Ghil and Malanotte-Rizzoli, 1991, Emerick and Reynolds, 2012]. Two points of view are facing in the DA field: variational and sequential methods (also known as filtering methods) [Kalnay, 2003]. In variational approaches, the measured data are incorporated over an entire time-interval. While filtering methods update the state of a system sequentially as soon as a new set of noisy observations becomes available. In this thesis, we focus our attention on filtering DA techniques which allow to combine computational models with noisy observations in order to improve the system state. In particular, one can focus on the Kalman filtering (KF) method. It consists of a forecast step, which evolves the underlying dynamical systems, and an analysis step, which adjusts the distribution of states. Under the hypotheses of Gaussianity and linearity, the KF is the exact solution to estimate recursively the probability distribution function of interest. Nevertheless, these hypotheses cannot hold in many situations, in particular the linear assumption. In this context, the ensem-

ble Kalman filter formulated by Evensen [2009] has been proposed and can be seen as a Monte Carlo approximation of the KF. The method relies on an ensemble of finite size of realizations to represent the error statistics but it is still based on the Gaussian hypothesis. The ensemble Kalman filter has been efficiently used in several applied studies, such as weather prediction, oceanography, reservoir engineering [Evensen and Van Leeuwen, 1996, Anderson, 2001, Aanonsen et al., 2009].

Recently research efforts have been devoted to the use of ensemble Kalman filtering approaches in order to study inverse problems for parameter estimation of numerical models. This method can be considered as particle-based derivative-free Bayesian algorithm and is sometimes referred to as ensemble Kalman inversion (EKI) [Iglesias et al., 2013, Schillings and Stuart, 2017]. This inference technique is a sequential Monte Carlo method, working on an ensemble of parameter particles, and transforming them from a prior distribution into a posterior one. As highlighted by Kovachki and Stuart [2019], the EKI is a procedure which behaves like the well-known gradient descent method, but without computing gradients. Instead, this filtering approach relies on an ensemble and is thus inherently parallelizable. Indeed, the forward calls of the numerical model can be parallelized across multiple processing units [Houtekamer et al., 2014, Ruiz et al., 2015]. This is a crucial asset when you are dealing with time-consuming black-box numerical models.

In Section 5.1, we give an overview of the Bayesian inference approach. In Section 5.2 we present two data assimilation strategies: the Kalman filter and its Monte Carlo approximation named ensemble Kalman filter. Section 5.3 is dedicated to the extension of ensemble Kalman filter for model calibration.

5.1 Bayesian inference

In Bayesian inference approach, we are concerned at characterizing the distribution of the input parameters conditioned on the measured data [Aster et al., 2018]. By employing such probabilistic paradigm, we can coherently quantify and reduce uncertainties in the input parameters with regard to all available information. The conditional probability distribution of the unknown parameters $\mathbf{X} \in \mathcal{P} \subset \mathbb{R}^p$ given the random data vector $\mathbf{Y} = (Y_1, \ldots, Y_m)^T$ is denoted by $p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$. The unknown parameters and the data are represented as jointly distributed random vectors:

$$(\mathbf{Y}, \mathbf{X}) \sim p_{\mathbf{Y},\mathbf{X}}(\mathbf{y}, \mathbf{x}) = p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}) \,,$$

where, $p_{\mathbf{X}}(\mathbf{x})$ is the prior distribution and $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ is the likelihood function.

On the one hand, the prior distribution represents our belief about the epistemic uncertainty affecting the input parameters before incorporating the data. This distribution can be based on heuristics such as expert knowledge. On the other hand, the likelihood function establishes the probabilities of obtaining the observations for a given parameter values. By conditioning on the realized observations, we can obtain the posterior distribution thanks to the Bayes' law:

$$p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{X}}(\mathbf{x})}{p_{\mathbf{Y}}(\mathbf{y})} \,, \tag{5.1}$$

where $p_{\mathbf{Y}}(\mathbf{y})$ is referred to the normalizing constant of $p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$, also known as the marginal likelihood or model evidence. This quantity, which represents the distribution

of $\mathbf{y}$ whichever the value of the parameter vector, is given by:

$$p_{\mathbf{Y}}(\mathbf{y}) = \int_{\mathcal{P}} p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}\,. \tag{5.2}$$

By combining Equation (5.1) and Equation (5.2), the Bayes' law can be formulated as:

$$p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{X}}(\mathbf{x})}{\int_{\mathcal{P}} p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}}\,.$$

Due to the fact that the model evidence does not depend on $\mathbf{x}$, one can rather consider the following relation of proportionality:

$$p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) \propto p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{X}}(\mathbf{x})\,. \tag{5.3}$$

In many applications, one is interested in estimating the unknown input parameters in a recursive manner. For performing such recursive inverse problems, a popular and flexible framework consists in relying on state-space formulations. The state-space models, also known as Hidden Markov Models (HMM), are considered as a powerful modeling framework for a variety of problems [Murphy, 2012, Bishop, 2006]. A state-space model consists of a pair of discrete-time stochastic processes denoted by $\{\mathbf{X}_{(k)}\}_{k\geq 0}$ and $\{\mathbf{Y}_{(k)}\}_{k\geq 1}$, whose realizations are respectively $\mathbf{x} = \{\mathbf{x}_{(k)}\}_{k\geq 0}$ and $\mathbf{y} = \{\mathbf{y}_{(k)}\}_{k\geq 1}$.

The law of the process is defined through the laws $p_{\mathbf{X}_{(0:k)}}(\mathbf{x}_{(0:k)})$ for all $k \geq 0$. Consequently, we have the following general result:

$$\begin{aligned} p_{\mathbf{X}_{(0:k)}}(\mathbf{x}_{(0:k)}) =& p_{\mathbf{X}_{(0)}}(\mathbf{x}_{(0)}) \times p_{\mathbf{X}_{(1)}|\mathbf{X}_{(0)}}(\mathbf{x}_{(1)}|\mathbf{x}_{(0)}) \times \\ & p_{\mathbf{X}_{(2)}|\mathbf{X}_{(0:1)}}(\mathbf{x}_{(2)}|\mathbf{x}_{(0:1)}) \ldots \times p_{\mathbf{X}_{(k)}|\mathbf{X}_{(0:k-1)}}(\mathbf{x}_{(k)}|\mathbf{x}_{(0:k-1)})\,. \end{aligned} \tag{5.4}$$

The discrete-time stochastic process $\{\mathbf{X}_{(k)}\}_{k=0}^{K}$, taking its values in a general state space $\mathcal{P}$, is considered as a Markov process.

Definition 3. *A sequence $\{\mathbf{X}_{(k)}\}_{k=0}^{K}$ is a Markov chain, if for any positive integer k and any possible realizations $(\mathbf{x}_{(0)}, \ldots, \mathbf{x}_{(K)})$ of the random variables, the conditional distribution of $\mathbf{X}_{(k)}$ given $\mathbf{X}_{(k-1)} = \mathbf{x}_{(k-1)}, \ldots, \mathbf{X}_{(0)} = \mathbf{x}_{(0)}$ is:*

$$p_{\mathbf{X}_{(k)}|\mathbf{X}_{(0:k-1)}}(\mathbf{x}_{(k)}|\mathbf{x}_{(0:k-1)}) = p_{\mathbf{X}_{(k)}|\mathbf{X}_{(k-1)}}(\mathbf{x}_{(k)}|\mathbf{x}_{(k-1)})\,.$$

By using the characteristic of a Markov chain, we can reformulate Equation (5.4) as:

$$p_{\mathbf{X}_{(0:k)}}(\mathbf{x}_{(0:k)}) = p_{\mathbf{X}_{(0)}}(\mathbf{x}_{(0)}) \prod_{l=1}^{k} p_{\mathbf{X}_{(l)}|\mathbf{X}_{(l-1)}}(\mathbf{x}_{(l)}|\mathbf{x}_{(l-1)})\,. \tag{5.5}$$

A Markov process is then entirely defined with its initial distribution $p_{\mathbf{X}_{(0)}}(\mathbf{x}_{(0)})$ and the transition distribution $p_{\mathbf{X}_{(l)}|\mathbf{X}_{(l-1)}}(\mathbf{x}_{(l)}|\mathbf{x}_{(l-1)})$. Graphical models are often used to describe stochastic models such as the class of Markov chains. A graphical representation of a Markov Chain Model is depicted in Figure 5.1.

Figure 5.1 – Graphical representation of a Markov chain process.

Lastly, the couple $(\mathbf{X}, \mathbf{Y})$ is said to be a hidden Markov chain in statistics if knowing the state, the observations are independent, i.e., property of conditional independence, such as:

$$p_{\mathbf{Y}_{(1:k)}|\mathbf{X}_{(0:k)}}(\mathbf{y}_{(1:k)}|\mathbf{x}_{(0:k)}) = \prod_{l=1}^{k} p_{\mathbf{Y}_{(l)}|\mathbf{X}_{(l)}}(\mathbf{y}_{(l)}|\mathbf{x}_{(l)}) \cdot \tag{5.6}$$

By combining Equation (5.5) and Equation (5.6) with the Bayes' rule, described in Equation (5.3), we get the posterior distribution as a product :

$$p_{\mathbf{X}_{(0:k)}|\mathbf{Y}_{(1:k)}}(\mathbf{x}_{(0:k)}|\mathbf{y}_{(1:k)}) \propto p_{\mathbf{X}_{(0)}}(\mathbf{x}_{(0)}) \prod_{l=1}^{k} p_{\mathbf{Y}_{(l)}|\mathbf{X}_{(l)}}(\mathbf{y}_{(l)}|\mathbf{x}_{(l)}) p_{\mathbf{X}_{(l)}|\mathbf{X}_{(l-1)}}(\mathbf{x}_{(l)}|\mathbf{x}_{(l-1)}) \cdot \tag{5.7}$$

Equation (5.7) states that a new update of the probability distribution can be obtained as soon as new observations are obtained. A graphical representation of a hidden Markov Chain Model is depicted in Figure 5.2.

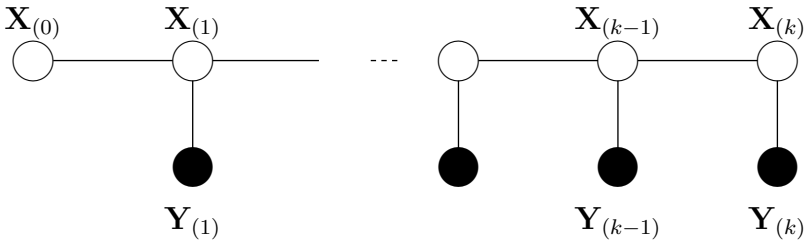


Figure 5.2 – Hidden Markov chain representation where $\mathbf{X}$ and $\mathbf{Y}$ represent respectively the hidden-states and the observations.

In this context, a recursive Bayesian procedure, called optimal filter, consists in finding the probability distribution function (pdf) of $\mathbf{X}_{(k)}$ given past and current observations, $\mathbf{Y}_{(1:k)}$. The conditional pdf of interest is hereafter denoted by $p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k)})$. The recursive Bayesian procedure relies on two distinct steps respectively named prediction and analysis.

Here we are assuming $p_{\mathbf{X}_{(k-1)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k-1)}|\mathbf{y}_{(1:k-1)})$ known, the prediction step, also known as forecast step, consists in making an approximation of the next state vector given all available information:

$$p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k-1)}) = \int_{\mathcal{P}} p_{\mathbf{X}_{(k)}|\mathbf{X}_{(k-1)}}(\mathbf{x}_{(k)}|\mathbf{x}_{(k-1)})\, p_{\mathbf{X}_{(k-1)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k-1)}|\mathbf{y}_{(1:k-1)})\, d\mathbf{x}_{(k-1)} \tag{5.8}$$

Then, the observation obtained at the current iteration is introduced thanks to the analysis step, and allows to correct the previous approximation as:

$$p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k)}) = \frac{p_{\mathbf{Y}_{(k)}|\mathbf{X}_{(k)}}(\mathbf{y}_{(k)}|\mathbf{x}_{(k)})\; p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k-1)})}{\int_{\mathcal{P}} p_{\mathbf{Y}_{(k)}|\mathbf{X}_{(k)}}(\mathbf{y}_{(k)}|\mathbf{x}_{(k)})\; p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k-1)})\; d\mathbf{x}_{(k)}} \cdot \tag{5.9}$$

In many physical problems the integrals in Equation (5.8) and Equation (5.9) cannot be computed due to the curse of dimensionality. The filtering methods consist in numerical approximations of these two integrals. In this context, different data assimilation methods have been developed. In the next section, we will describe a sequential data assimilation procedure called Kalman filter and its Monte Carlo variant named as ensemble Kalman filter.

5.2 Data assimilation

In the data assimilation framework, state-space models are often based on two key components:

A dynamics model - Let us assume that a model of the natural processes of interest is available as a discrete stochastic-dynamical system:

$$\mathbf{x}_{(k)} = \mathcal{M}_{(k-1,k)}(\mathbf{x}_{(k-1)}, \boldsymbol{\epsilon}^m_{(k)}) \cdot$$

Here $\mathbf{x}_{(k)} \in \mathcal{P} \subset \mathbb{R}^p$ is the model state vector, $\mathcal{M}_{(k-1,k)} : \mathbb{R}^p \to \mathbb{R}^p$ is usually a nonlinear function from iteration $k-1$ to k, and $\boldsymbol{\epsilon}^m_{(k)} \in \mathbb{R}^p$ is the model error. In many applications, we can represent this model error as a stochastic additive term, such as:

$$\mathbf{x}_{(k)} = \mathcal{M}_{(k-1,k)}(\mathbf{x}_{(k-1)}) + \boldsymbol{\epsilon}^m_{(k)}\,, \tag{5.10}$$

where $\boldsymbol{\epsilon}^m_{(k)}$ is the model error, represented here as a stochastic additive term following a Gaussian distribution such that $\boldsymbol{\epsilon}^m_{(k)} \overset{iid}{\sim} \mathcal{N}\left(0, \mathbf{Q}_{(k)}\right)$.

An observation model - Noisy measurements are available at discrete iterations and are considered as components of the observation vector denoted by $\mathbf{y}_{(k)} \in \mathbf{Y} \subset \mathbb{R}^m$. These collected data are related to the model state vector $\mathbf{x}_{(k)}$ at iteration k such as:

$$\mathbf{y}_{(k)} = \mathcal{H}_{(k)}(\mathbf{x}_{(k)}, \boldsymbol{\epsilon}^o_{(k)}) \text{ with } \boldsymbol{\epsilon}^o_{(k)} \overset{iid}{\sim} \mathcal{N}\left(0, \mathbf{R}_{(k)}\right)\,,$$

where $\mathcal{H}_{(k)} : \mathbb{R}^p \to \mathbb{R}^m$ is the observation function (generally nonlinear), and $\boldsymbol{\epsilon}^o_{(k)}$ is the observation error which accounts for the deficiencies in the formulation of the observation function, and instrumental error of the sensing devices. By assuming the observation error is represented as a stochastic additive term following a Gaussian distribution, the observation model can be reformulated:

$$\mathbf{y}_{(k)} = \mathcal{H}_{(k)}(\mathbf{x}_{(k)}) + \boldsymbol{\epsilon}^o_{(k)} \cdot \tag{5.11}$$

In order to alleviate the mathematical formulation and the computational burden, standard assumptions about model and observation errors are: additivity, Gaussiannity, unbiasedness and mutual independence. Unfortunately, these assumptions cannot always hold in some industrial applications.

5.2.1 Kalman filter

The Kalman filter (KF), introduced by Kalman [1960], allows to estimate the posterior distribution, defined in Equation (5.9), recursively as new observed data are obtained. This filtering technique provides the exact distribution when the state-space model is

linear-Gaussian. In the literature, discrete-time linear-Gaussian state-space models have the following form:

$$\forall k \in \mathbb{N}^*, \left\{ \begin{array}{l} \mathbf{x}_{(k)} = \mathbf{M}_{(k-1,k)}\mathbf{x}_{(k-1)} + \boldsymbol{\epsilon}^m_{(k)} \\ \mathbf{y}_{(k)} = \mathbf{H}_{(k)}\mathbf{x}_{(k)} + \boldsymbol{\epsilon}^o_{(k)} \end{array} \right. , \tag{5.12}$$

where, $\mathbf{M}_{(k-1,k)}$ and $\mathbf{H}_{(k)}$ are respectively the matrices representing the linear model and the observation operator, model errors $\boldsymbol{\epsilon}^m_{(k)}$, and observation errors $\boldsymbol{\epsilon}^o_{(k)}$ are white Gaussian noises, of zero mean and of respective covariance $\mathbf{Q}_{(k)}$ et $\mathbf{R}_{(k)}$. Moreover, one can assume that there is no correlation between model and observation errors, such as:

$$\mathbb{E}[\boldsymbol{\epsilon}^o_{(k)}\ \boldsymbol{\epsilon}^{m\ T}_{(k)}] = 0\,,$$

and that they are supposed to be independent of the initial condition.

KF approach has been widely used in many applications [Brown, 1986, Harvey, 1990, Wells, 2013]. Let us consider that the initial state, denoted by $\mathbf{X}_{(0)}$, is Gaussian, of expectation $\mathbf{x}_b$, and covariance $\mathbf{P}_b$. Then the state-space system described in Equation (5.12) can be written as:

$$\left\{ \begin{array}{l} \mathbf{X}_{(0)} \sim p_{\mathbf{X}_{(0)}}(\mathbf{x}_{(0)}) = \mathcal{N}(\mathbf{X}_{(0)}; \mathbf{x}_b, \mathbf{P}_b) \\ \mathbf{X}_{(k)}|\mathbf{X}_{(k-1)} \sim p_{\mathbf{X}_{(k)}|\mathbf{X}_{(k-1)}}(\mathbf{x}_{(k)}|\mathbf{x}_{(k-1)}) = \mathcal{N}(\mathbf{X}_{(k)}; \mathbf{M}_{(k-1,k)}\ \mathbf{X}_{(k-1)}, \mathbf{Q}_{(k)}) \\ \mathbf{Y}_{(k)}|\mathbf{X}_{(k)} \sim p_{\mathbf{Y}_{(k)}|\mathbf{X}_{(k)}}(\mathbf{y}_{(k)}|\mathbf{x}_{(k)}) = \mathcal{N}(\mathbf{Y}_{(k)}; \mathbf{H}_{(k)}\ \mathbf{X}_{(k)}, \mathbf{R}_{(k)}) \end{array} \right. .$$

Let us assume that the Gaussian pdf $p_{\mathbf{X}_{(k-1)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k-1)}|\mathbf{y}_{(1:k-1)})$ is known through the mean $\mathbf{x}^a_{(k-1)}$ and the covariance matrix $\mathbf{P}^a_{(k-1)}$. One can forecast the state vector from iteration $k-1$ to time k. By substituting the pdf in Equation (5.8) by the Gaussian distributions, we obtain the following formula:

$$p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k-1)}) = \int_{\mathcal{P}} \mathcal{N}(\mathbf{x}_{(k)}\ ;\ \mathbf{M}_{(k-1,k)}\ \mathbf{x}_{(k-1)}, \mathbf{Q}_{(k)})\, \mathcal{N}(\mathbf{x}_{(k-1)}\ ;\ \mathbf{x}^a_{(k-1)}, \mathbf{P}^a_{(k-1)})\, d\mathbf{x}_{(k-1)}$$

We can show that:

$$\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k-1)} \sim p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k-1)}) = \mathcal{N}(\mathbf{X}_{(k)}\ ;\ \mathbf{x}^f_{(k)}\ ,\ \mathbf{P}^f_{(k)})\,,$$

where,

$$\begin{aligned} \mathbf{x}^f_{(k)} &= \mathbb{E}[\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k-1)} = \mathbf{y}_{(1:k-1)}] \\ &= \mathbb{E}[\mathbf{M}_{(k-1,k)}\ \mathbf{X}_{(k-1)} + \boldsymbol{\epsilon}^m_{(k)}|\mathbf{Y}_{(1:k-1)} = \mathbf{y}_{(1:k-1)}] \\ &= \mathbf{M}_{(k-1,k)}\ \mathbb{E}[\mathbf{X}_{(k-1)}|\mathbf{Y}_{(1:k-1)} = \mathbf{y}_{(1:k-1)}] + \mathbb{E}[\boldsymbol{\epsilon}^m_{(k-1)}|\mathbf{Y}_{(1:k-1)} = \mathbf{y}_{(1:k-1)}] \\ &= \mathbf{M}_{(k-1,k)}\ \mathbf{x}^a_{(k-1)}\,. \end{aligned}$$

and

$$\begin{aligned} \mathbf{P}^f_{(k)} &= \mathbb{E}[(\mathbf{X}_{(k)} - \mathbf{x}^f_{(k)})(\mathbf{X}_{(k)} - \mathbf{x}^f_{(k)})^T] \\ &= \mathbb{E}[(\mathbf{M}_{(k-1,k)}(\mathbf{X}_{(k-1)} - \mathbf{x}^a_{(k-1)}) - \boldsymbol{\epsilon}^m_{(k)})\ (\mathbf{M}_{(k-1,k)}(\mathbf{X}_{(k-1)} - \mathbf{x}^a_{(k-1)}) - \boldsymbol{\epsilon}^m_{(k-1)})^T] \\ &= \mathbf{M}_{(k-1,k)}\mathbf{P}^a_{(k-1)}\mathbf{M}^T_{(k-1,k)} + \mathbb{E}[\boldsymbol{\epsilon}^m_{(k)}\boldsymbol{\epsilon}^{m\ T}_{(k)}] \\ &+ \mathbf{M}_{(k-1,k)}\mathbb{E}[(\mathbf{X}_{(k-1)} - \mathbf{x}^a_{(k-1)})\boldsymbol{\epsilon}^{m\ T}_{(k)}] + \mathbb{E}[\boldsymbol{\epsilon}^m_{(k)}(\mathbf{X}_{(k-1)} - \mathbf{x}^a_{(k-1)})^T]\mathbf{M}^T_{(k-1,k)} \\ &= \mathbf{M}_{(k-1,k)}\mathbf{P}^a_{(k-1)}\mathbf{M}^T_{(k-1,k)} + \mathbf{Q}_{(k)}\,. \end{aligned}$$

At iteration k, the forecast Gaussian distribution $p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k-1)})$ is known thanks to the forecast step through the mean $\mathbf{x}^f_{(k)}$, and the covariance matrix $\mathbf{P}^f_{(k)}$. The analysis step consists in updating this pdf using the new set of observations stacked into the vector $\mathbf{y}_{(k)}$, and in finding the filtering distribution $p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k)})$, described in Equation (5.9). By knowing that:

$$p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k-1)}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}^f_{(k)}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_{(k)} - \mathbf{x}^f_{(k)})^T \ {\mathbf{P}^f_{(k)}}^{-1} \ (\mathbf{x}_{(k)} - \mathbf{x}^f_{(k)})\right] ,$$

and

$$p_{\mathbf{Y}_{(k)}|\mathbf{X}_{(k)}}(\mathbf{y}_{(k)}|\mathbf{x}_{(k)}) = \frac{1}{(2\pi)^{n/2}|\mathbf{R}_{(k)}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}_{(k)} - \mathbf{H}_{(k)}\mathbf{x}_{(k)})^T \ {\mathbf{R}_{(k)}}^{-1} \ (\mathbf{y}_{(k)} - \mathbf{H}_{(k)}\mathbf{x}_{(k)})\right] .$$

Under linear-Gaussian assumption, KF provides the filtering distribution by using Bayes' rule. Then, Equation (5.9) can be reduced to:

$$\begin{aligned} p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k)}) &= \frac{\mathcal{N}(\mathbf{y}_{(k)};\ \mathbf{H}_{(k)}\mathbf{x}_{(k)}, \mathbf{R}_{(k)})\ \mathcal{N}(\mathbf{x}_{(k)};\ \mathbf{x}^b_{(k)}, \mathbf{P}^b_{(k)})}{\mathcal{N}(\mathbf{y}_{(k)};\ \mathbf{H}_{(k)}\mathbf{x}^b_{(k)}, \mathbf{H}_{(k)}\mathbf{P}^b_{(k)}\mathbf{H}^T_{(k)} + \mathbf{R}_{(k)})} , \\ &= \mathcal{N}(\mathbf{x}_{(k)}; \mathbf{x}^a_{(k)}, \mathbf{P}^a_{(k)}) , \end{aligned}$$

where

$$\begin{aligned} \mathbf{x}^a_{(k)} &= \mathbf{x}^f_{(k)} + \mathbf{K}_{(k)}(\mathbf{y}_{(k)} - \mathbf{H}_{(k)}\mathbf{x}^f_{(k)}) , \\ \mathbf{P}^a_{(k)} &= \mathbf{P}^f_{(k)} - \mathbf{K}_{(k)}\mathbf{H}_{(k)}\mathbf{P}^f_{(k)} , \end{aligned}$$

with

$$\mathbf{K}_{(k)} = \mathbf{P}^f_{(k)} \ \mathbf{H}^T_{(k)}(\mathbf{R}_{(k)} + \mathbf{H}_{(k)}\mathbf{P}^f_{(k)}\mathbf{H}^T_{(k)})^{-1} . \tag{5.13}$$

Within this recursive context, the gain, described in Equation (5.13), is often called the Kalman gain. It is a ratio based on how much we trust the prediction versus the measurement. If we are confident in our measurements and unconfident in our predictions the Kalmain gain will favor the measured data, and vice versa. Moreover, due to the fact that $p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k)})$ is a Gaussian distribution, the updated mean $\mathbf{x}^a_{(k)}$ and the covariance matrix $\mathbf{P}^a_{(k)}$ are sufficient information for characterizing the full filtering distribution.

By considering the background state vector $\mathbf{x}_b$ and its associated error covariance matrix denoted by $\mathbf{P}_b$, the recursive steps of the Kalman filter can be summarized as follows:

Initialization -

$$\mathbf{x}_b \text{ and } \mathbf{P}_b$$

Forecast step -

$$p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k-1)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k-1)}) = \mathcal{N}(\mathbf{x}_{(k)}; \mathbf{x}^f_{(k)}, \mathbf{P}^f_{(k)})$$

$$\begin{aligned} \mathbf{x}^f_{(k)} &= \mathbf{M}_{(k-1,k)} \ \mathbf{x}^a_{(k-1)} \\ \mathbf{P}^f_{(k)} &= \mathbf{M}_{(k-1,k)} \ \mathbf{P}^a_{(k-1)} \ \mathbf{M}^T_{(k-1,k)} + \mathbf{Q}_{(k)} \end{aligned}$$

Analysis step -

$$p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k)}) = \mathcal{N}(\mathbf{x}_{(k)}; \mathbf{x}^a_{(k)}, \mathbf{P}^a_{(k)})$$

$$\begin{aligned}
\mathbf{K}_{(k)} &= \mathbf{P}^f_{(k)}\ \mathbf{H}^T_{(k)}(\mathbf{R}_{(k)} + \mathbf{H}_{(k)}\mathbf{P}^f_{(k)}\mathbf{H}^T_{(k)})^{-1} \\
\mathbf{P}^a_{(k)} &= (\mathbf{I} - \mathbf{K}_{(k)}\mathbf{H}_{(k)})\mathbf{P}^f_{(k)} \\
\mathbf{x}^a_{(k)} &= \mathbf{x}^f_{(k)} + \mathbf{K}_{(k)}(\mathbf{y}_{(k)} - \mathbf{H}_{(k)}\mathbf{x}^f_{(k)})
\end{aligned}$$

Unfortunately, Kalman filter technique has to deal with some limitations. One limitation of the implementation of the KF is the filter divergence. If the input statistical information is misspecified, the KF approach cannot infer a solution close to the true target [Fitzgerald, 1971]. Another limitation is due to the computational restriction involved by the necessary storage of a $p \times p$ state covariance matrix which becomes intractable for high p. A final limitation is due to the inconsistency of the statistical and dynamical hypotheses of the KF approach, i.e., Gaussianity and linearity assumptions. Indeed, the observation or the background errors can often be physically non-Gaussian, e.g., variables constrained by thresholds. Moreover, nonlinearity of the dynamics or the observation function affects the KF in two ways. Firstly, the transposed of these functions is not defined. Secondly, nonlinearity of the functions destroys the Gaussianity of statistics. In this context, the extended Kalman filter (EKF), which is a first-order expansion of the Kalman filter, has been developed. The EKF has been successful used in different applications, such as meteorology and oceanography [Ghil et al., 1981, Ghil and Malanotte-Rizzoli, 1991]. Nevertheless in high dimensional applications, the EKF cannot be implemented due to the high cost associated with the iterative construction of the covariance matrix. Besides, the implementation of the EKF relies on the local linear tangent which leads to neglect the nonlinear effects. Consequently, this method is efficient only for weakly nonlinear models. Otherwise, one may rely to more complex strategy such as the ensemble Kalman filter.

5.2.2 Ensemble Kalman filter

As seen in the previous section, the Kalman filter solves the optimal filter by given explicit recursive expressions of the two first moments of the probability distribution functions. This is optimal only in the linear Gaussian case. In the case of nonlinear models and/or non Gaussian pdfs, the Kalman filter is no more optimal and can also easily fail the estimation process. In this context, Evensen introduces in [Evensen, 1994] a Monte Carlo approximation of the Kalman filter called the Ensemble Kalman filter (EnKF). This Monte Carlo filtering method has been used in many application studies due to its robustness, its ease of implementation, and its efficient accuracy, e.g., oceanography, reservoir modeling, and weather forecasting [Evensen and Van Leeuwen, 1996, Houtekamer and Mitchell, 2001]. This method relies on the use of an ensemble to represent the error statistics. Nevertheless, the EnKF is still based on the Gaussian hypothesis, i.e., its analysis and update steps only rely on the mean and covariance. Let us consider the following state-space formulation:

$$\forall k \in \mathbb{N}^*, \left\{ \begin{array}{l} \mathbf{x}_{(k)} = \mathcal{M}_{(k-1,k)}(\mathbf{x}_{(k-1)}) + \boldsymbol{\epsilon}^m_{(k)} \\ \mathbf{y}_{(k)} = \mathcal{H}_{(k)}(\mathbf{x}_{(k)}) + \boldsymbol{\epsilon}^o_{(k)} \end{array} \right. , \tag{5.14}$$

where, $\mathcal{M}_{(k-1,k)}$ and $\mathcal{H}_{(k)}$ are nonlinear functions, model errors $\boldsymbol{\epsilon}^m_{(k)}$, and observation errors $\boldsymbol{\epsilon}^o_{(k)}$ are white Gaussian noises, of zero mean and of respective covariance $\mathbf{Q}_{(k)}$ et $\mathbf{R}_{(k)}$.

At iteration k, we have an ensemble of size N_{ens}, of forecast parameter estimates $\mathbf{x}^f_{(k)} = [\mathbf{x}^{f(1)}_{(k)}, \ldots, \mathbf{x}^{f(N_{ens})}_{(k)}]^T \in \mathbb{R}^{p \times N_{ens}}$ where the superscript $\cdot^{f(i)}$ denotes the i-th forecast member of the ensemble. The mean of the forecast members of the ensemble is given by:

$$\bar{\mathbf{x}}^f_{(k)} = \frac{1}{N_{ens}} \sum_{i=1}^{N_{ens}} \mathbf{x}^{f(i)}_{(k)} \,.$$

Then, the forecast covariance matrix can be defined as:

$$\boldsymbol{P}^f_{(k)} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} (\mathbf{x}^{f(i)}_{(k)} - \bar{\mathbf{x}}^f)(\mathbf{x}^{f(i)}_{(k)} - \bar{\mathbf{x}}^f_{(k)})^T \,.$$

As said previously the structure of the EnKF is the same as the one of the Kalman filter. Thus, we need to compute the Kalman Gain $\boldsymbol{K}_{(k)}$ defined by:

$$\mathbf{K}_{(k)} = \mathbf{P}^f_{(k)}\ \mathbf{H}^T_{(k)} (\mathbf{R}_{(k)} + \mathbf{H}_{(k)} \mathbf{P}^f_{(k)} \mathbf{H}^T_{(k)})^{-1} \,. \tag{5.15}$$

In Equation (5.15), the observation operator, denoted by $\mathbf{H}_{(k)} \in \mathbb{R}^{m \times p}$, is linear or has been linearized. Nevertheless, for most applications this condition of linear (or linearized) observation operator cannot be applied. In that context, we can replace the terms $\mathbf{P}^f_{(k)}\ \mathbf{H}^T_{(k)}$ and $\mathbf{H}_{(k)} \mathbf{P}^f_{(k)} \mathbf{H}^T_{(k)}$ of the Kalman Gain equation as [Houtekamer and Mitchell, 2001]:

$$\mathbf{P}^f_{(k)}\ \mathbf{H}^T_{(k)} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} (\mathbf{x}^{f(i)}_{(k)} - \bar{\mathbf{x}}^f_{(k)})(\mathcal{H}_{(k)}(\mathbf{x}^{f(i)}_{(k)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}^f_{(k)}))^T \,, \tag{5.16}$$

and,

$$\mathbf{H}_{(k)} \mathbf{P}^f_{(k)} \mathbf{H}^T_{(k)} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} (\mathcal{H}_{(k)}(\mathbf{x}^{f(i)}_{(k)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}^f_{(k)}))(\mathcal{H}_{(k)}(\mathbf{x}^{f(i)}_{(k)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}^f_{(k)}))^T \,. \tag{5.17}$$

Equation (5.16) and Equation (5.17) linearize the nonlinear measurement function $\mathcal{H}_{(k)}$ to the observation operator $\mathbf{H}_{(k)}$ by a linearization process using ensemble members. The difference between the original EnKF and the stochastic version presented in this section is that the observations are now treated as random variables. Indeed, it has been proven by Burgers et al. [1998] that, in order for the EnKF analysis error covariance to be consistent with the one of the KF, one have to treat the observations as random variables following a Gaussian distribution with mean equal to the observed value and covariance equal to $\mathbf{R}_{(k)}$. Most of the time, this observation error covariance matrix is diagonal according to the assumption of independent observations. Consequently, an ensemble of observations of the same size N_{ens} is generated by adding noise terms to the observation set $\mathbf{y}_{(k)}$ such that:

$$\mathbf{y}^{(i)}_{(k)} = \mathbf{y}_{(k)} + \boldsymbol{\epsilon}^{o(i)}_{(k)} \text{ with } \boldsymbol{\epsilon}^{o(i)}_{(k)} \sim \mathcal{N}(0, \mathbf{R}_{(k)}),\ i = 1 \ldots N_{ens} \,.$$

Then the computation of the Kalman gain $\boldsymbol{K}_{(k)}$ can be done. We can independently update the ensemble members using:

$$\mathbf{x}^{a(i)}_{(k)} = \mathbf{x}^{f(i)}_{(k)} + \boldsymbol{K}_{(k)} \left(\mathbf{y}^{(i)}_{(k)} - \mathcal{H}_{(k)}(\mathbf{x}^{f(i)}_{(k)}) \right) \,,$$

where the superscript $\cdot^{a(i)}$ denotes the i-th updated member of the ensemble. The last step of the EnKF method is the forecast step of the ensemble parameters at $k+1$ and involves an ensemble of N_{ens} updated parameters for iteration k.

$$\mathbf{x}^{f(i)}_{(k+1)} = \mathcal{M}_{(k,k+1)}(\mathbf{x}^{a(i)}_{(k)}) + \boldsymbol{\epsilon}^{m(i)}_{(k+1)}, \; i = 1 \dots N_{ens} \,,$$

with $\boldsymbol{\epsilon}^{m(i)}_{(k+1)} \sim \mathcal{N}(0, \mathbf{Q}_{(k+1)})$.

As studied by Le Gland et al. [2009], the solution of the stochastic EnKF converges for $N_{ens} \to \infty$ with a rate of $\circ(N_{ens}^{-\frac{1}{2}})$. In the EnKF method, two main sources of sampling errors can be considered. Firstly, the one due to the use of a finite ensemble of model realizations, secondly the one due to the introduction of stochastic observation perturbations. To reduce the sampling error, one can rely on the use of Latin hypercube sampling strategy, see Section 3.5, instead of a conventional Monte Carlo method. A graphical representation of the ensemble Kalman filter is illustrated in Figure 5.3 combining two sources of information: model forecasts (in green) and observations (in blue). The presented filtering approach uses several members in order to track the hidden-state of the system at each iteration. The uncertainties affecting both the forecast, the analysis, and the observations can be obtained as displayed by the ellipsoids. The algorithm of this method is summarized in Algorithm 3.

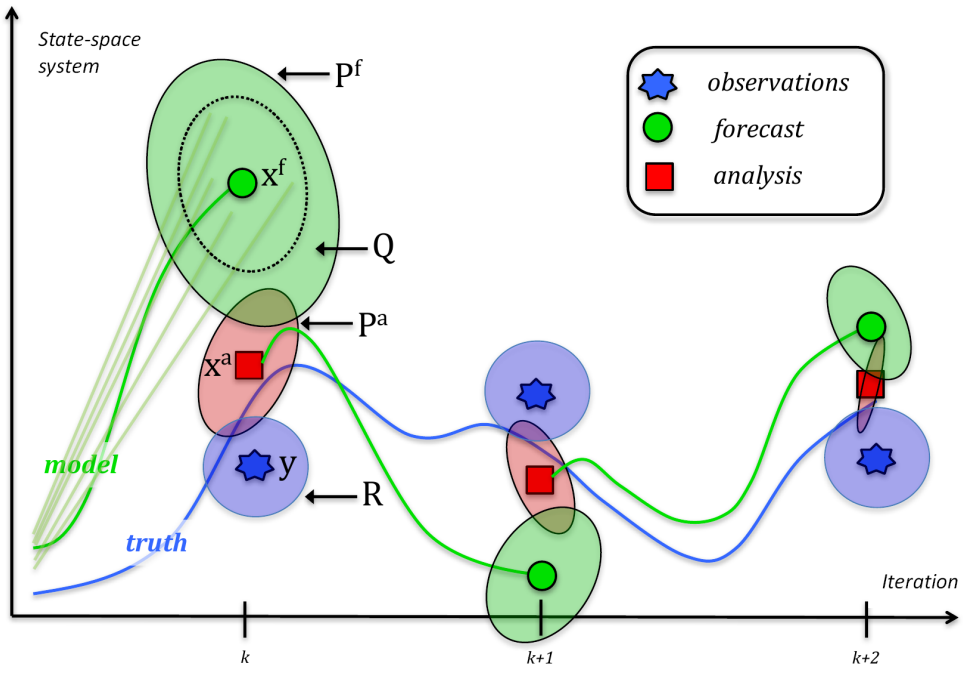


Figure 5.3 – Sketch of the ensemble Kalman filter adapted from [Tandeo et al., 2020]. The ellipses represent the forecast $\mathbf{P}^f$ and analysis $\mathbf{P}^a$ error covariances, the model $\mathbf{Q}$ and observation $\mathbf{R}$ error covariances of the state-space model defined in Equation (5.14)

.

Algorithm 3: Ensemble Kalman Filter.

Data:
number of members in the ensemble N_{ens};
prior guess of the parameter vector $\mathbf{x}_b$ and prior parameter covariance matrix $\mathbf{P}_b$;
some measurements $\{\mathbf{y}_{(k)}\}_{k=1,\dots,K}$;
error covariance matrix $\{\mathbf{R}_{(k)}\}_{k=1,\dots,K}$ and artificial error covariance matrix $\{\mathbf{Q}_{(k)}\}_{k=0,\dots,K}$.

Initialisation step:

for $i = 1$ *to* N_{ens} **do**

$$\mathbf{x}_{(0)}^{a(i)} = \mathbf{x}_b + \boldsymbol{\epsilon}^b \text{ with, } \boldsymbol{\epsilon}^b \sim \mathcal{N}(0, \mathbf{P}_b)$$

for $k = 1$ *to* K **do**

Forecast step:

for $i = 1$ *to* N_{ens} **do**

$$\mathbf{x}_{(k)}^{f(i)} = \mathcal{M}_{(k-1,k)}(\mathbf{x}_{(k-1)}^{a(i)}) + \boldsymbol{\epsilon}_{(k)}^{m(i)} \text{ with, } \mathbf{e}_{(k)}^{m(i)} \sim \mathcal{N}(0, \mathbf{Q}_{(k)})$$

$$\bar{\mathbf{x}}_{(k)}^{f} = \frac{1}{N_{ens}} \sum_{i=1}^{N_{ens}} \mathbf{x}_{(k)}^{f(i)} \text{ and } \mathbf{P}_{(k)}^{f} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} (\mathbf{x}_{(k)}^{f(i)} - \bar{\mathbf{x}}_{(k)}^{f})(\mathbf{x}_{(k)}^{f(i)} - \bar{\mathbf{x}}_{(k)}^{f})^T$$

Update step:

$$\mathbf{P}_{(k)}^{f}\ \mathbf{H}_{(k)}^{T} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathbf{x}_{(k)}^{f(i)} - \bar{\mathbf{x}}_{(k)}^{f}\right) \left(\mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}_{(k)}^{f})\right)^T$$

$$\mathbf{H}_{(k)}\mathbf{P}_{(k)}^{f}\mathbf{H}_{(k)}^{T} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}_{(k)}^{f})\right) \left(\mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}_{(k)}^{f})\right)^T$$

$$\mathbf{K}_{(k)} = \mathbf{P}_{(k)}^{f}\ \mathbf{H}_{(k)}^{T} \left(\mathbf{R}_{(k)} + \mathbf{H}_{(k)}\mathbf{P}_{(k)}^{f}\mathbf{H}_{(k)}^{T}\right)^{-1}$$

for $i = 1$ *to* N_{ens} **do**

$$\mathbf{y}_{(k)}^{(i)} = \mathbf{y}_{(k)} + \boldsymbol{\epsilon}_{(k)}^{o(i)} \text{ with } \boldsymbol{\epsilon}_{(k)}^{o(i)} \sim \mathcal{N}(0, \mathbf{R}_{(k)})$$

$$\mathbf{x}_{(k)}^{a(i)} = \mathbf{x}_{(k)}^{f(i)} + \mathbf{K}_{(k)} \left(\mathbf{y}_{(k)}^{(i)} - \mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)})\right)$$

The forecast covariance $\mathbf{P}^f$ can be badly estimated due to its evaluation based on an ensemble of limited size. The misestimation of the forecast covariance produces an accumulation of the sampling error which leads to the filter divergence [Pham et al., 1998, Anderson and Anderson, 1999, Anderson, 2007]. One way around proposed in [Anderson and Anderson, 1999] is to inflate the error covariance matrix by a scalar factor, denoted by λ, greater than 1:

$$\mathbf{P}_{(k)}^{f} = \lambda \mathbf{P}_{(k)}^{f}\,.$$

5.3 Data assimilation technique for parameter estimation

Data assimilation methods, such as the ensemble Kalman filter, have been developed for estimating the state of a dynamical system from noisy observations. Nevertheless, ensemble Kalman methods in the context of inverse problems such as parameter estimation have been developed in oil industry applications where this inference problem is known as history matching [Li et al., 2007, Oliver et al., 2008]. Recently, there has been a growing interest in applying EnKF to inverse problems. This recursive inference method can be seen as a derivative-free procedure [Stuart and Zygalakis, 2015, Reich, 2018], which allows to have high parallelism capabilities. The algorithm of this model calibration, based on the ensemble Kalman filter, is summarized in Algorithm 4. The idea is to consider an artificial dynamic for the vector of parameters such as:

$$\forall i \in \mathbb{N}^*,\ \mathbf{x}_{(k)}^{f(i)} = \mathbf{x}_{(k-1)}^{a(i)} + \boldsymbol{\epsilon}_{(k)}^{m(i)} \cdot \tag{5.18}$$

Conclusion

This chapter highlights the concept of recursive Bayesian inference. In this context, data assimilation techniques have been presented with a particular focus on the Kalman filter and its Monte Carlo variant called ensemble Kalman filter. Such data assimilation techniques combine forecasts from a numerical model with noisy observations, based on the model and observation equations of a state-space formulation. These sequential inference procedures have been developed to sequentially update the probability distribution on states of partially observed systems. This chapter introduces how the ensemble Kalman filtering approach can be extended to numerical model calibration problems. Moreover, the use of ensemble Kalman filtering methods to perform inverse problems for parameter estimation is really convenient in the context of black-box time-consuming numerical models because it is a derivative-free procedure with high parallelism capabilities.

Algorithm 4: Ensemble Kalman Filter for model calibration.

Data:
number of members in the ensemble N_{ens};
prior guess of the parameter vector $\mathbf{x}_b$ and prior parameter covariance matrix $\mathbf{P}_b$;
some measurements $\{\mathbf{y}_{(k)}\}_{k=1,\dots,K}$;
error covariance matrix $\{\mathbf{R}_{(k)}\}_{k=1,\dots,K}$ and artificial error covariance matrix $\{\mathbf{Q}_{(k)}\}_{k=0,\dots,K}$.

Initialisation step:

for $i = 1$ *to* N_{ens} **do**

$$\mathbf{x}_{(0)}^{a(i)} = \mathbf{x}_b + \boldsymbol{\epsilon}^b \text{ with, } \boldsymbol{\epsilon}^b \sim \mathcal{N}(0, \mathbf{P}_b)$$

for $k = 1$ *to* K **do**

Forecast step:

for $i = 1$ *to* N_{ens} **do**

$$\mathbf{x}_{(k)}^{f(i)} = \mathbf{x}_{(k-1)}^{a(i)} + \boldsymbol{\epsilon}_{(k)}^{m(i)} \text{ with, } \mathbf{e}_{(k)}^{m(i)} \sim \mathcal{N}(0, \mathbf{Q}_{(k)})$$

$$\bar{\mathbf{x}}_{(k)}^{f} = \frac{1}{N_{ens}} \sum_{i=1}^{N_{ens}} \mathbf{x}_{(k)}^{f(i)} \text{ and } \mathbf{P}_{(k)}^{f} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} (\mathbf{x}_{(k)}^{f(i)} - \bar{\mathbf{x}}_{(k)}^{f})(\mathbf{x}_{(k)}^{f(i)} - \bar{\mathbf{x}}_{(k)}^{f})^T$$

Update step:

$$\mathbf{P}_{(k)}^{f}\ \mathbf{H}_{(k)}^{T} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathbf{x}_{(k)}^{f(i)} - \bar{\mathbf{x}}_{(k)}^{f}\right) \left(\mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}_{(k)}^{f})\right)^T$$

$$\mathbf{H}_{(k)}\mathbf{P}_{(k)}^{f}\mathbf{H}_{(k)}^{T} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}_{(k)}^{f})\right) \left(\mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}_{(k)}^{f})\right)^T$$

$$\mathbf{K}_{(k)} = \mathbf{P}_{(k)}^{f}\ \mathbf{H}_{(k)}^{T} \left(\mathbf{R}_{(k)} + \mathbf{H}_{(k)}\mathbf{P}_{(k)}^{f}\mathbf{H}_{(k)}^{T}\right)^{-1}$$

for $i = 1$ *to* N_{ens} **do**

$$\mathbf{y}_{(k)}^{(i)} = \mathbf{y}_{(k)} + \boldsymbol{\epsilon}_{(k)}^{o(i)} \text{ with } \boldsymbol{\epsilon}_{(k)}^{o(i)} \sim \mathcal{N}(0, \mathbf{R}_{(k)})$$

$$\mathbf{x}_{(k)}^{a(i)} = \mathbf{x}_{(k)}^{f(i)} + \mathbf{K}_{(k)} \left(\mathbf{y}_{(k)}^{(i)} - \mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)})\right)$$

Part III

Contributions to the industrial application

I have not failed. I have just found 10,000 things that do not work.

Thomas Edison

6

Quantification and reduction of uncertainties in a wind turbine numerical model based on a global sensitivity analysis and a recursive Bayesian inference approach

---

Les résultats présentés dans ce chapitre ont donné lieu à un article soumis pour publication au journal scientifique "International Journal for Numerical Methods in Engineering" [Hirvoas et al.]. L'objectif de ce chapitre est d'apporter une réponse à la quantification et la réduction des incertitudes dans le contexte de la modélisation numérique d'éolienne. L'intérêt de cette étude est dans un premier temps de déterminer les paramètres liés aux propiétés du modèle de l'éolienne ayant une influence sur la fatigue de cette dernière. Puis dans un second temps, nous nous intéressons à la réduction des incertitudes entâchant ces paramètres influents. La méthode d'inférence développée au cours de ces travaux est basée sur un algorithme de filtrage Bayésien d'ensemble, issu du domaine de l'assimilation de données [Evensen, 2009, Iglesias et al., 2013, Schillings and Stuart, 2017, Kovachki and Stuart, 2019]. Cette approche de filtrage, appelée filtre de Kalman d'ensemble, peut être facilement exécutée en parallèle, ce qui est d'un grand intérêt pour les codes numériques multi-physiques utilisés dans le domaine éolien [Jonkman and Buhl Jr., 2005, Le Cunff et al., 2013, DNV GL, 2013], où le coût de calcul de la simulation directe est déjà un défi en soi. Cependant, afin d'analyser la possibilité d'estimation des paramètres d'entrée du modèle numérique, une étude d'identifiabilité a été réalisée. Pour celà, nous utilisons le lien entre l'analyse de sensibilité globale et d'identifiabilité abordé par Dobre et al. [2012]. Cette méthode d'inférence récursive vise à tirer pleinement profit des quantités importantes de données fournies par les capteurs placés sur les éoliennes modernes en production, en combinant de manière optimale les données de production et les modèles numériques afin d'obtenir des modèles hautement fidèles des éoliennes. Ces travaux de recherche se placent dans le contexte industriel du développement d'un "jumeau numérique d'une éolienne terrestre". La méthodologie basée sur le filtre de Kalman d'ensemble laisse envisager la mise à jour en continu des modèles numériques aéro-servo-élastiques, utilisés pour l'estimation de la durée de vie de l'éolienne, afin de tenir compte des éventuelles modifications des propriétés de la structure au cours de sa durée de vie. Le chapitre est organisé comme suit. La méthodologie d'analyse de sensibilité globale dans le contexte de modèle numérique stochastique et coûteux en temps de calcul est présentée dans la Section 6.1. Dans la Section 6.2, nous explorons comment le filtre de Kalman d'ensemble

peut être utilisé dans les problèmes de calibration de modèles numériques. La section 6.3 est consacrée à la présentation du modèle numérique de l'éolienne, ses paramètres d'entrée considérés comme incertains et les grandeurs de sortie retenues pour quantifier et réduire les incertitudes. Dans la section 6.4, une étude de cas numérique d'éoliennes est utilisée pour illustrer la procédure proposée et ses performances dans la calibration de paramètres avec des données pseudo-expérimentales bruitées.

Introduction

In the current profound worldwide energy transition, wind power generation is developing rapidly. As a consequence, wind turbines monitoring, performance optimization and lifetime assessment are becoming major issues. In the context of digitalization of the industry, the exploitation of collected data can be optimized by combination with wind turbine numerical models. Such numerical models can be complex and costly as they involve nonlinear dynamic equations with different physics as well as stochastic loading from the wind. Moreover, some input parameters of the models can be poorly or badly known as the structure ages over time and defaults can appear. Consequently, model predictions are affected by these uncertainties. Characterization and reduction of these uncertainties is important for decision making [De Rocquigny et al., 2008]. It is the case in wind energy applications where uncertainties are ubiquitous both in external conditions and in the models used during design process. In this context, uncertainty quantification and reduction methods have been developed [Smith, 2013]. As mentioned by Hart et al. [2017], even the concept of sensitivity is delicate when dealing with stochastic models, as the one in our industrial application whose stochasticity is due to the stochastic nature of the wind external solicitation. Note also that the models used in wind energy applications are often time consuming [Jonkman and Buhl Jr., 2005, Perdrizet et al., 2013]. Therefore most of the commonly used methodologies for uncertainty quantification are inappropriate in our setting. An historical strategy for uncertainty quantification in wind energy fields was to take into account uncertainties by employing Monte Carlo methods [Kwon, 2010, Jin and Tian, 2010]. Nevertheless, for high-fidelity numerical models, such uncertainty quantification approaches based on Monte Carlo methods become cumbersome due to the computational cost. Advanced methods such as polynomial chaos expansion, stochastic collocation or Gaussian process have been developed to alleviate this computational issue, [see Petrone et al., 2011, Wang et al., 2016, Murcia et al., 2018].

In the present work, we aim at investigating a complete framework to quantify and reduce the input parameter uncertainties involved in a finite element wind turbine model. Such methods to reduce uncertainties involved in models used in wind energy are currently investigated [Sørensen and Toft, 2010]. In the literature, Van Buren et al. [2013] has developed a framework to quantify and reduce such uncertainties based on ANOVA decomposition and Bayesian inference. However, our approach is innovative due to the fact that we are dealing with a high-fidelity wind turbine numerical model and by the recursive aspect of our inference procedure. Recently, similar research work on recursive inference with a low-fidelity wind turbine numerical model has been led by Branlard et al. [2020]. They develop a digital twin concept in order to estimate turbine states; e.g., wind speed, torque; based on the Kalman filter. Our main contribution is twofold.

Firstly, the framework allows quantifying the sources of uncertainties affecting the fatigue behavior of the structural components of the wind turbine. To perform such analysis, we usually use an aero-servo-elastic software fed by model parameters and wind solicitation. Each wind field is computed using a stochastic turbulent wind simulator. The obtained outputs of the simulations are different time-series describing the behavior of the wind turbine, which are reduced to some quantities of interest (QoIs), see Figure 6.1. The function $\mathcal{D}$ representing the time-consuming numerical model is defined as:

$$\mathbf{y} = \mathcal{D}(\mathbf{x}, V)\,, \tag{6.1}$$

where, $\mathbf{x} = (x_1, ..., x_p) \in \mathcal{P} \subset \mathbb{R}^p$ are the model input parameters, V is a stochastic process modeling the external wind solicitation, $\mathbf{y} = (y_1, ..., y_m) \in \mathbb{R}^m$ is the vector of discretized functional output; e.g., generated power, structural accelerations or loads. Let g be the function mapping the functional loads of the structure in $\mathbf{y}$ to the damage quantity of interest (QoI), such as the damage-equivalent load (DEL) [Freebury and Musial, 2000]:

$$g(\mathbf{y}) := g \circ \mathcal{D}(\mathbf{x}, V)\,. \tag{6.2}$$

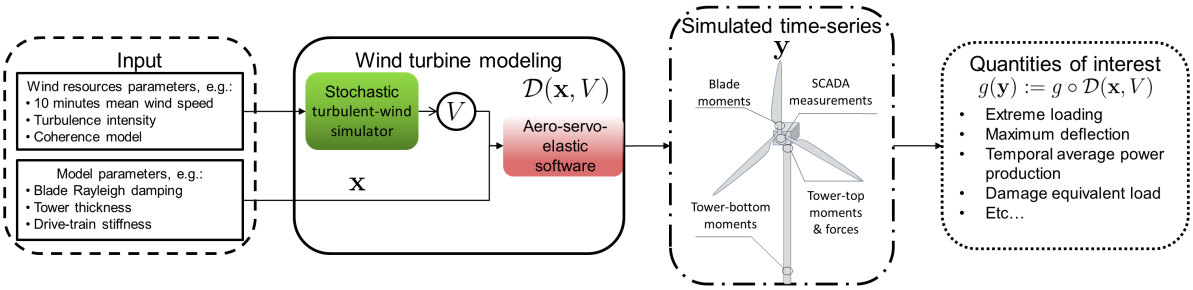


Figure 6.1 – General sketch for wind turbine modeling.

Global Sensitivity Analysis (GSA) methods have been developed to quantify the uncertainty in QoI with regard to the input parameters, their individual contributions, or the contribution of their interactions. We propose a variance-based GSA methodology, relied on the so-called Sobol' indices [Sobol', 1993], for stochastic computer simulations. Such techniques, which often refer to the probabilistic framework and Monte Carlo (MC) methods, require a lot of calls to the numerical model. The uncertain input parameters are modeled by independent random variables gathered into a random vector and characterized by their probability distribution. Variance-based SA for time consuming deterministic computer models has been mainly performed by approximating the model by a mathematical function, also known as a surrogate model. Among the different surrogate models, we focus on Gaussian process (GP), also known as kriging, which is characterized by its mean and covariance functions. One advantage of the GP model is to provide both a prediction of the numerical model and the associated uncertainty. Such surrogate modeling assumes that prior beliefs about the numerical code can be modeled by a Gaussian process. Nowadays, in industrial applications, numerical models are often run at different levels of complexity and then a hierarchy of simulations is available. In this context, if several resolutions of simulation are obtained, multi-fidelity GP regression has been proposed to predict the output of a costly numerical model, [see Forrester et al.,

2007, Forrester and Keane, 2009]. In particular, Parussini and Perdikaris have proposed a recursive formulation of the approach combined with other approximations to enhance the computational efficiency, [see Perdikaris et al., 2015, 2016, Parussini et al., 2017]. These multi-fidelity formulations of the GP regression are computing time efficient in terms of number of model evaluations in comparison to the simple GP regression. However in the present industrial application, we did not implement multi-fidelity formulations of GP regression as we could only run a high-resolution version of the stochastic simulator $\mathcal{D}$. In order to take into account the inherent randomness from wind turbine simulation, our approach consists in focusing on the mean behavior of the high-resolution runs of the stochastic simulator described by the deterministic model $\mathbf{x} \mapsto \mathbb{E}_V[g \circ \mathcal{D}(\mathbf{x}, V)]$. GP regression is then used to reduce the numerical costs. More precisely, noisy evaluations of the conditional expectation are computed via Monte Carlo and then filtered using heteroscedastic GP modeling. Lastly, variance-based sensitivity indices are computed by running the GP based surrogate in a so-called pick-freeze estimation procedure [Le Gratiet et al., 2017].

Secondly, after identification of the less influential input parameters on the fatigue behavior of the wind turbine, we propose a Bayesian inference framework to carry out a model calibration procedure based on in situ measurements. It uses measurements $\mathbf{y}^{mes}$ to update some prior probability distributions about the unknown input parameters $\mathbf{X} \sim p(\mathbf{x})$ and yields some posterior probability distributions, through the Bayes' theorem $p(\mathbf{x}|\mathbf{y}^{mes}) \propto p(\mathbf{y}^{mes}|\mathbf{x})p(\mathbf{x})$ [1]. Numerous batch techniques have been developed to solve such Bayesian problems. Nevertheless, recent decades have been marked by a simultaneous development of sensor technologies and Internet of Things capabilities. Thus, our research efforts have been directed towards inference techniques where the data are sequentially processed when new observations become available. In this context, the model parameter inference can be carried out using sequential Bayesian techniques. In geosciences, these techniques are called data assimilation methods. We perform the calibration using a recursive Bayesian inference approach based on an Ensemble Kalman Filter (EnKF) [Evensen, 2009].

However, such recursive inverse problems can be solved assuming that several conditions of well-posedness and identifiability are achieved. These conditions have been summarized by Hadamard, Jacques [1902]. As highlighted by Dobre et al. [2012], a relationship between the non-identifiability of input parameters and the GSA can be established. Indeed, input parameters with null total sensitivity indices on the measured outputs imply their non-identifiability. Therefore, for the purpose of identifiability a second GSA is conducted on the calibration parameters. However, due to the functional behavior of the measurements, we propose to first reduce their dimensionality through principal component analysis. Then, a GP is fitted to the different principal components and used to compute an aggregated Sobol' index for each model parameter [Lamboni et al., 2011].

Last but not least, the proposed framework has been applied to an industrial wind turbine numerical model. The developed recursive inference procedure has shown promising results in the industrial inversion problem.

The chapter is organized as follows. The GSA methodology in the context of stochastic time-consuming numerical model is introduced in Section 6.1. In Section 6.2, we explore

1. Random variables are written in upper case roman letters and particular realizations of a random variable are written in corresponding lower case letters.

how the EnKF can be employed in model calibration problems. Section 6.3 is devoted to present the wind turbine numerical model, its uncertain input parameters, and the selected output quantities used for quantifying and reducing the uncertainties. In Section 6.4, a wind turbine numerical case study is used to illustrate the proposed framework and its performance in calibrating parameters with noisy pseudo-experimental output data.

6.1 Kriging based global sensitivity analysis

6.1.1 Introduction

The aim of sensitivity analysis is to quantify the relative influence of input parameters on some QoI produced from the model outputs of the numerical model. In the context of model calibration, conducting such an analysis can help to identify which input parameters should be properly estimated. One may distinguish two categories of methods: local and global. While local sensitivity analysis considers small perturbations of the inputs around some nominal values, global sensitivity analysis (GSA) varies the inputs on their whole variation range [Saltelli et al., 2000]. Among the large number of available approaches, variance-based sensitivity analysis introduced by Sobol' [1990] proposes to measure the sensitivity by computing the so-called Sobol' indices. When no analytical formulae of these indices are available, one way to perform their estimation is to rely on Monte Carlo (MC) techniques, which require a huge number of model evaluations. In the context of costly numerical codes as, e.g., the wind turbine numerical model under interest, the use of a cheap metamodel in place of the true costly model is thus crucial. In addition to being computationally expensive, the numerical model we are dealing with is stochastic. This means that from a same set of input parameters, the output can have different values depending on the wind conditions. This specificity has to be carefully taken into account when estimating sensitivity measures under interest. More precisely, let us use the formalism introduced in Equations (6.1) and (6.2) to the model in hand. We are interested in measuring the sensitivity of a QoI $g(\mathbf{y}) \in \mathbb{R}$ with respect to the input $\mathbf{x}$. In the context of GSA, each input is now considered as a random variable X_j with its uncertainty modeled by a probability distribution, such as $\mathbf{X} = (X_1, \ldots, X_p)$. These one-dimensional probability distributions reflect the practitioner's belief in the uncertainty on the parameter values and the X_j are assumed to be independent from each other. Then, the QoI $g(\mathbf{Y}) := g \circ \mathcal{D}(\mathbf{X}, V)$ is a random variable itself.

The randomness of the QoI has two sources: the randomness from the parameters $\mathbf{X}$, and the one due to the stochasticity propagated from the model itself through V, which is assumed to be independent of $\mathbf{X}$. There exists at least two approaches to deal with this stochasticity in a GSA framework. The first one considers the full probability distribution of the QoI while the other one is only concerned with quantitative measures of the probability distribution, e.g., quantile or expectation [Etoré et al., 2018]. The latter is the one considered in this work by investigating the QoI averaged over the inherent randomness of the physical system. We are therefore interested in the sensitivity of the deterministic function f defined as:

$$f(\mathbf{X}) = \mathbb{E}_V[g \circ \mathcal{D}(\mathbf{X}, V)]\,.$$

The total variance of $f(\mathbf{X})$ can be split into different parts of variance under the assumption that the input parameters are independent (this is the so-called Hoeffding

decomposition, [see Hoeffding, 1948]). Each part of variance corresponds to the contribution of each set of parameters on the variance of the output $f(\mathbf{X})$. By considering the ratio of each part of variance to the total variance, we obtain a measure of importance for each set of input parameters that is called the Sobol' index [Sobol', 1990]. In the literature, functional analysis of variance (FANOVA) has been widely used to quantify the sensitivity of a model output to input variables (see Appendix A in [Owen, 2013] and [Saltelli, 2002]). Let us denote $\mathbf{u}$ a subset of $\{1, ..., p\}$, $-\mathbf{u}$ its complement and $|\mathbf{u}|$ its cardinality. Assuming $Var_{\mathbf{X}}[f(\mathbf{X})] < +\infty$, $Var_{\mathbf{X}}[f(\mathbf{X})] \neq 0$, we define for any $\mathbf{u}$ the closed Sobol' index of order $r = |\mathbf{u}|$ associated to the set of inputs $\mathbf{X_u} = \{X_j\}_{j \in \mathbf{u}}$ as:

$$S_{\mathbf{u}} = \frac{Var_{\mathbf{X_u}}[\mathbb{E}_{\mathbf{X_{-u}}}[f(\mathbf{X})|\mathbf{X_u}]]}{Var_{\mathbf{X}}[f(\mathbf{X})]} \,. \tag{6.3}$$

This index quantifies the main effect of all the variables within $\mathbf{X_u}$, including interactions, on $f(\mathbf{X})$. The most influential sets of input parameters can then be identified as the sets of input parameters with the largest Sobol' indices. Total Sobol' indices can also been defined as:

$$S^T_{\mathbf{u}} = 1 - S_{\mathbf{-u}} = 1 - \frac{Var_{\mathbf{X_{-u}}}[\mathbb{E}_{\mathbf{X_u}}[f(\mathbf{X})|\mathbf{X_{-u}}]]}{Var_{\mathbf{X}}[f(\mathbf{X})]} \,. \tag{6.4}$$

This index quantifies the effect of $\mathbf{X_u}$ plus the effect of all interactions between variables in $\mathbf{X_u}$ and variables in $\mathbf{X_{-u}}$ on Z.

A general approach to estimate Sobol' indices is based on Monte Carlo and requires an important number of evaluations of f. The high computational cost of the wind turbine model prevents from performing such estimation in reasonable time. It is the reason why we would like to rely on a metamodel to compute cheap evaluations of the initial costly computer code. In this study, we chose to approximate the true numerical code by a Gaussian process, also known as kriging metamodel, in order to apply a kriging based sensitivity analysis, e.g., in [Le Gratiet et al., 2013]. We firstly present the noiseless framework and then we detail the case where we are facing to heterogeneously noisy evaluations of the function f.

6.1.2 Ordinary kriging

First, a Gaussian process regression model is built to surrogate the function f. The principle of kriging based metamodeling [Krige et al., 1989] is to consider that our deterministic model f can be considered as a realization of a Gaussian process $\{Z(\mathbf{x}) \,,\ \mathbf{x} \in \mathcal{P}\}$ with mean function μ and covariance kernel C. Such covariance kernel (a.k.a. covariance function, kernel function, or kernel), is a positive-definite function of two distinct inputs $\mathbf{x}, \mathbf{x}'$ allowing to define the prior covariance between any two values of the function of interest. Many kernels can be used, each one corresponding to a different set of prior assumptions made about the function of interest [Stein, 2012, Rasmussen, 2003, Duvenaud, 2014]. Each kernel is defined by a number of parameters which specify the shape of the covariance function. These parameters, also known as hyper-parameters, can be either estimated by minimizing a loss function with a Leave-One-Out Cross-Validation procedure or maximizing a likelihood function [Bachoc, 2013]. In our study, we use the last mentioned approach to estimate these hyper-parameters of a 5/2 Matérn kernel.

For any $\mathbf{x} \in \mathcal{P}$, $f(\mathbf{x})$ is approximated by the conditional Gaussian process $\{Z_n(\mathbf{x}) \,,\ \mathbf{x} \in \mathcal{P}\} := \{[Z(\mathbf{x})|Z(\mathbf{X}^n) = \mathbf{z}] \,,\ \mathbf{x} \in \mathcal{P}\}$, where $\mathbf{z} = \{z_1, \dots z_n\}$ are evaluations of f on n points

$\mathbf{X}^n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathcal{P}$. The design $\{(\mathbf{x}_1, z_1), \dots, (\mathbf{x}_n, z_n)\}$ is called the learning sample. In the following, $\mathbf{X}^n$ is chosen as a Latin Hypercube sample [McKay et al., 1979] to guarantee a good exploration of our numerical model. We then get the ordinary kriging equations:

$$Z_n(\mathbf{x}) \sim \mathcal{N}(m_{OK}(\mathbf{x}), s^2_{OK}(\mathbf{x}))\,, \tag{6.5}$$

with

$$\begin{aligned} m_{OK}(\mathbf{x}) &= \mu(\mathbf{x}) + \mathbf{c}(\mathbf{x})^T \mathbf{C}^{-1}(\mathbf{z} - \boldsymbol{\mu})\,, \\ s^2_{OK}(\mathbf{x}) &= C(\mathbf{x}, \mathbf{x}) - \mathbf{c}(\mathbf{x})^T \mathbf{C}^{-1} \mathbf{c}(\mathbf{x})\,. \end{aligned}$$

We denote by $\boldsymbol{\mu} = \mu(\mathbf{X}^n)$ the vector of trend values on $\mathbf{X}^n$, by $\mathbf{C} = (C(\mathbf{x}_i, \mathbf{x}_j))_{1 \le i,j \le n}$ the covariance matrix of $Z(\mathbf{X}^n)$, and by $\mathbf{c}(\mathbf{x}) = (C(\mathbf{x}, \mathbf{x}_i))_{1 \le i \le n}$ the vector of covariances between $Z(\mathbf{x})$ and $Z(\mathbf{X^n})$.

6.1.3 Noisy kriging

In our context, exact evaluations of f can not be obtained directly. We have, for each $i = 1, \dots, n$, a noisy evaluation $\tilde{z}_i = f(\mathbf{x}_i) + \varepsilon_i$, where $\tilde{z}_i$ is defined as an empirical mean computed from a K-sample of $g \circ \mathcal{D}(\mathbf{x}_i, V)$, and ε_i is a centered noise whose variance is defined as

$$\tau_i^2 = \frac{1}{K} \left(\frac{1}{K-1} (\sum_{j=1}^{K} (\mathcal{D}(\mathbf{x}_i, V = v_j) - \frac{1}{K} \sum_{j=1}^{K} \mathcal{D}(\mathbf{x}_i, V = v_j))^2 \right) .$$

We then consider, as a first approximation, that the vector $(\varepsilon_1, \dots, \varepsilon_n)$ is a centered Gaussian random vector with diagonal covariance matrix $\mathrm{diag}(\tau_1^2, \dots, \tau_n^2)$ denoted by $\boldsymbol{\Delta}$. Then, provided that the process Z and the Gaussian measurement errors $\{\varepsilon_i\}_{1 \le i \le n}$ are stochastically independent, the process conditionally on the noisy observations $\{\tilde{Z}_n(\mathbf{x})\,,\ \mathbf{x} \in \mathcal{P}\} := \{[Z(\mathbf{x}) | \tilde{Z}(\mathbf{X}^n) = \{\tilde{z}_i\}_{1 \le i \le n}]\,,\ \mathbf{x} \in \mathcal{P}\}$ is still Gaussian, and its conditional mean and variance functions are given by the following slightly modified kriging equations:

$$\tilde{Z}_n(\mathbf{x}) \sim \mathcal{N}(m_{NK}(\mathbf{x}), s^2_{NK}(\mathbf{x}))\,, \tag{6.6}$$

with

$$\begin{aligned} m_{NK}(\mathbf{x}) &= \mu(\mathbf{x}) + \mathbf{c}(\mathbf{x})^T (\mathbf{C} + \boldsymbol{\Delta})^{-1} (\tilde{\mathbf{z}} - \boldsymbol{\mu})\,, \\ s^2_{NK}(\mathbf{x}) &= C(\mathbf{x}, \mathbf{x}) - \mathbf{c}(\mathbf{x})^T (\mathbf{C} + \boldsymbol{\Delta})^{-1} \mathbf{c}(\mathbf{x})\,. \end{aligned}$$

6.1.4 Kriging based Sobol' indices

Following Le Gratiet et al. [2013] and Marrel et al. [2009] the idea is to substitute f with $\tilde{Z}_n$ in Equation (6.3):

$$\tilde{S}_{\mathbf{u}} = \frac{Var_{\mathbf{X_u}}[\mathbb{E}_{\mathbf{X_{-u}}}[\,\tilde{Z}_n(\mathbf{X}) \mid \mathbf{X_u}]]}{Var_{\mathbf{X}}[\tilde{Z}_n(\mathbf{X})]}\,. \tag{6.7}$$

As $\tilde{Z}_n$ is a random process, the resulting indices are also random. These indices are estimated via Monte Carlo samples from two designs of experiments, using a pick-freeze

procedure. A design is a point set $\mathbf{P} = \{\mathbf{x}_i\}_{i=1}^s$ in which each point is obtained by sampling s times each input variable $X_j \in \mathcal{P}$, $j = 1, \ldots, p$. Each row of the design is a point $\mathbf{x}_i$ in $\mathcal{P}$, the j-th column of the design refers to a sample of X_j and for $\mathbf{u} \subseteq \{1, \ldots, p\}$, $\mathbf{x}_{i,\mathbf{u}} = \{x_{i,j}\}$, $j \in \mathbf{u}$. Given two points $\mathbf{x}$ and $\mathbf{x}'$, the hybrid point $(\mathbf{x}_\mathbf{u} : \mathbf{x}'_{-\mathbf{u}}) \in \mathcal{P}$ is defined as x_j if $j \in \mathbf{u}$ and x'_j if $j \notin \mathbf{u}$. Consider $\mathbf{P} = \{\mathbf{x}_i\}_{i=1}^s$ and $\mathbf{P}' = \{\mathbf{x}'_i\}_{i=1}^s$ two designs sampled from the distribution of the input random vector $\mathbf{X}$. One way to estimate the quantity in Equation (6.7) has been proposed by Homma and Saltelli [1996] and further studied in Janon *et al.*, see Lemma 1 in [Janon, 2012]. They propose the following estimator:

$$\widehat{\tilde{S}}_\mathbf{u} = \frac{\frac{1}{s}\sum_{i=1}^s \tilde{Z}_n(\mathbf{x}_i)\ \tilde{Z}_n(\mathbf{x}_{i,\mathbf{u}} : \mathbf{x}'_{i,-\mathbf{u}}) - \frac{1}{s}\sum_{i=1}^s \tilde{Z}_n(\mathbf{x}_i)\ \frac{1}{s}\sum_{i=1}^s \tilde{Z}_n(\mathbf{x}_{i,\mathbf{u}} : \mathbf{x}'_{i,-\mathbf{u}})}{\frac{1}{s}\sum_{i=1}^s \tilde{Z}_n(\mathbf{x}_i)^2 - (\frac{1}{s}\sum_{i=1}^s \tilde{Z}_n(\mathbf{x}_i))^2}. \tag{6.8}$$

Confidence intervals can be obtained via a bootstrap method, as described in Algorithm 1 of [Le Gratiet et al., 2013]. These intervals integrate two sources of uncertainty, the first one is related to the metamodel approximation, and the second one is related to the Monte Carlo integration.

Each step of the procedure was implemented in **R** using the *km* and *sobolGP* functions from respectively the *DiceKriging* and *Sensitivity* packages, [see Roustant et al., 2012, Iooss et al., 2019].

6.2 Bayesian inference for online parameter identification

In our context, we suppose that data are collected sequentially and we seek to refine our choice of parameters in the numerical model at each iteration. This problem can be seen as a supervised learning problem that we aim to solve sequentially as each pair of data points $\{v_{(k)}, \mathbf{y}_{(k)}\}$ is obtained at the iteration k [Kovachki and Stuart, 2019].

In the recursive Bayesian parameter estimation framework, developed in this paper, the unknown time-invariant input parameter vector $\mathbf{x}$ is modeled as a discrete Markov chain, the evolution of which is governed by a random walk process. In our context, the dynamic evolution of the input parameter and the measurement modelisation with the simulator responses can be formulated at iteration $\{k\}_{k=1\ldots T}$ as:

$$\begin{cases} \mathbf{x}_{(k)} &= \mathbf{x}_{(k-1)} + \delta_{(k)} \\ \mathbf{y}_{(k)} &= \mathcal{D}(\mathbf{x}_{(k)}, V = v_{(k)}) + \boldsymbol{\epsilon}_{(k)} \end{cases}, \tag{6.9}$$

where $\mathbf{x}_{(k)}$ is the input parameter vector, $v_{(k)}$ is a known realization of the stochastic external excitation at k, the Gaussian noises $\delta_{(k)} \sim \mathcal{N}(0, \mathbf{Q}_{(k)})$ and $\boldsymbol{\epsilon}_{(k)} \sim \mathcal{N}(0, \mathbf{R}_{(k)})$ are respectively an artificial dynamic noise and a combination of the model and observation errors. For the sake of readability, $\mathbf{y} \in \mathbb{R}^m$ will represent the vector gathering the measured responses obtained on the structure of interest.

Filtering techniques, a type of data assimilation, can be used to sequentially estimate the parameter vector in Equation (6.9) using the known input solicitation and the available measurements. Among all available filtering methods, the Kalman Filter (KF) [Kalman et al., 1960] has been widely applied when dealing with a linear system with Gaussian error sources. In this paper due to the nonlinearity in our numerical model, the Ensemble

Kalman filter (EnKF) [Evensen, 2009] is used to perform parameters estimation. The EnKF is a sequential Monte Carlo method that provides an alternative to the traditional KF. The method works on an ensemble of parameter estimates transforming them from the prior distribution into the posterior one. We propose to use a Latin Hypercube sampling technique coupled with a geometrical criteria maximizing the minimum distance between the design points instead of a conventional Monte Carlo method to generate the initial ensemble of parameter estimates, [see Johnson et al., 1990].

In the field of inverse problems, this inference method is referred to as Ensemble Kalman Inversion (EKI). The EnKF formulation used in [Snyder and Zhang, 2003] is adopted in this paper. At iteration k, we have an ensemble of size N_{ens}, of forecast parameter estimates $\mathbf{x}^f_{(k)} = [\mathbf{x}^{f(1)}_{(k)}, \dots, \mathbf{x}^{f(N_{ens})}_{(k)}] \in \mathbb{R}^{p \times N_{ens}}$ where the superscript $\cdot^{f(i)}$ denotes the i-th forecast member of the ensemble. The mean of the forecast members of the ensemble is given by:

$$\bar{\mathbf{x}}^f_{(k)} = \frac{1}{N_{ens}} \sum_{i=1}^{N_{ens}} \mathbf{x}^{f(i)}_{(k)} \,.$$

The error covariance matrix for the forecast estimate in the KF can be empirically estimated as:

$$\mathbf{P}^f_{(k)} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} (\mathbf{x}^{f(i)}_{(k)} - \bar{\mathbf{x}}^f_{(k)})(\mathbf{x}^{f(i)}_{(k)} - \bar{\mathbf{x}}^f_{(k)})^T \,.$$

As said previously the structure of the EnKF is the same as the one of the Kalman filter [Welch and Bishop, 1995]. Thus, we need to compute the Kalman Gain, referred to as $\mathbf{K}_{(k)}$ and defined by:

$$\mathbf{K}_{(k)} = \mathbf{P}^f_{(k)} \mathbf{M}^T \left(\mathbf{M} \mathbf{P}^f_{(k)} \mathbf{M}^T + \mathbf{R}_{(k)} \right)^{-1} ,$$

where the observation operator, denoted by $\mathbf{M} \in \mathbb{R}^{m \times N_{ens}}$, is linear or has been linearized from the function $\mathcal{D}$, [see Kopp and Orford, 1963].

Nevertheless, for most applications this condition of linear (or linearized) observation operator cannot be applied. In that context, as proposed in [Houtekamer and Mitchell, 2001], we can replace the terms $\mathbf{P}^f_{(k)}\mathbf{M}^T$ and $\mathbf{M}\mathbf{P}^f_{(k)}\mathbf{M}^T$ of the Kalman Gain equation by the following ones:

$$\frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathbf{x}^{f(i)}_{(k)} - \bar{\mathbf{x}}^f_{(k)} \right) \left(\mathcal{D}(\mathbf{x}^{f(i)}_{(k)}, V = v_{(k)}) - \mathcal{D}(\bar{\mathbf{x}}^f_{(k)}, V = v_{(k)}) \right)^T \tag{6.10}$$

and,

$$\frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathcal{D}(\mathbf{x}^{f(i)}_{(k)}, V = v_{(k)}) - \mathcal{D}(\bar{\mathbf{x}}^f_{(k)}, V = v_{(k)}) \right) \left(\mathcal{D}(\mathbf{x}^{f(i)}_{(k)}, V = v_{(k)}) - \mathcal{D}(\bar{\mathbf{x}}^f_{(k)}, V = v_{(k)}) \right)^T \cdot \tag{6.11}$$

It has been argued, in [Ambadan and Tang, 2009], that Equation (6.10) and Equation (6.11) are good approximations if the following hypothesis are verified:

$$\mathcal{D}(\bar{\mathbf{x}}^f_{(k)}, V = v_{(k)}) = \overline{\mathcal{D}(\mathbf{x}^f_{(k)}, V = v_{(k)})} = \frac{1}{N_{ens}} \sum_{i=1}^{N_{ens}} \mathcal{D}(\mathbf{x}^{f(i)}_{(k)}, V = v_{(k)}) \,,$$

$$\mathbf{x}^{f(i)}_{(k)} - \bar{\mathbf{x}}^f_{(k)} = \xi_i \text{ and } \|\xi_i\| \text{ is small for } i = 1 \dots N_{ens} \,.$$

This version of EnKF treats the observations as random variables [Evensen, 2009]. Indeed, an ensemble of observations of the same size N_{ens} is generated by adding noise terms to the observation set $\mathbf{y}_{(k)}$ such that:

$$\mathbf{y}_{(k)}^{(i)} = \mathbf{y}_{(k)} + \mathbf{e}_{(k)}^{o(i)}\,, \text{ with } \mathbf{e}_{(k)}^{o(i)} \sim \mathcal{N}(0, \mathbf{R}_{(k)}), i = 1 \dots N_{ens}\,.$$

The noise terms are sampled from the distribution of the error covariance matrix $\mathbf{R}_{(k)}$. The stochastic EnKF has been showed to have the advantage to "re-Gaussianize" the ensemble distribution thanks to the observation perturbations [Lawson and Hansen, 2004]. Maintaining Gaussianity has a positive impact on analysis quality of the ensemble filter by maintaining the correct forecast error covariance. Most of the time, the measurement observational error covariance matrix is diagonal according to the assumption of independent observations. Using the presented approximation, the computation of the Kalman gain $\mathbf{K}_{(k)}$ can be done. We can independently update the ensemble members using:

$$\mathbf{x}_{(k)}^{a(i)} = \mathbf{x}_{(k)}^{f(i)} + \mathbf{K}_{(k)} \left(\mathbf{y}_{(k)}^{(i)} - \mathcal{D}(\mathbf{x}_{(k)}^{f(i)}, V = v_{(k)}) \right)\,.$$

where the superscript $\cdot^{a(i)}$ denotes the i-th updated member of the ensemble. The last step of the EnKF method is the forecast step of the ensemble parameters at $k+1$ and involves an ensemble of N_{ens} updated parameters for iteration k, such as:

$$\mathbf{x}_{(k+1)}^{f} = \mathbf{x}_{(k)}^{a} + \delta_{(k)}\,, \text{ with } \delta_{(k)} \sim \mathcal{N}(0, \mathbf{Q}_{(k)})\,.$$

The presented method is fully described in Algorithm 3.

Algorithm 5: Ensemble Kalman Filter for parameter inference, a.k.a. Ensemble Kalman Inversion.

Data:
number of members in the ensemble N_{ens};
prior guess of the parameter vector $\mathbf{x}_b$ and prior parameter covariance matrix $\mathbf{P}_b$;
some measurements $\{\mathbf{y}_{(k)}\}_{k=1,...,T}$ and known realization of the external solicitation $\{v_{(k)}\}_{k=1,...,T}$;
error covariance matrix $\{\mathbf{R}_{(k)}\}_{k=1,...,T}$ and artificial error covariance matrix $\{\mathbf{Q}_{(k)}\}_{k=0,...,T}$.

Initialisation step:

for $i = 1$ *to* N_{ens} **do**

$$\mathbf{x}_{(0)}^{a(i)} = \mathbf{x}_b + \boldsymbol{\epsilon}^b \text{ with, } \boldsymbol{\epsilon}^b \sim \mathcal{N}(0, \mathbf{P}_b)$$

for $k = 1$ *to* T **do**

Forecast step:

$$\mathbf{x}_{(k)}^{f} = \mathbf{x}_{(k-1)}^{a} + \delta_{(k)}\,, \text{ with } \delta_{(k)} \sim \mathcal{N}(0, \mathbf{Q}_{(k)})$$

$$\bar{\mathbf{x}}_{(k)}^{f} = \frac{1}{N_{ens}} \sum_{i=1}^{N_{ens}} \mathbf{x}_{(k)}^{f(i)} \text{ and } \mathbf{P}_{(k)}^{f} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} (\mathbf{x}_{(k)}^{f(i)} - \bar{\mathbf{x}}_{(k)}^{f})(\mathbf{x}_{(k)}^{f(i)} - \bar{\mathbf{x}}_{(k)}^{f})^T$$

Update step:

$$\mathbf{P}_{(k)}^{f}\mathbf{M}^T = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathbf{x}_{(k)}^{f(i)} - \bar{\mathbf{x}}_{(k)}^{f}\right) \left(\mathcal{D}(\mathbf{x}_{(k)}^{f(i)}, V = v_{(k)}) - \mathcal{D}(\bar{\mathbf{x}}_{(k)}^{f}, V = v_{(k)})\right)^T$$

$$\mathbf{M}\mathbf{P}_{(k)}^{f}\mathbf{M}^T = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathcal{D}(\mathbf{x}_{(k)}^{f(i)}, V = v_{(k)}) - \mathcal{D}(\bar{\mathbf{x}}_{(k)}^{f}, V = v_{(k)})\right) \left(\mathcal{D}(\mathbf{x}_{(k)}^{f(i)}, V = v_{(k)}) - \mathcal{D}(\bar{\mathbf{x}}_{(k)}^{f}, V = v_{(k)})\right)^T$$

$$\mathbf{K}_{(k)} = \mathbf{P}_{(k)}^{f}\mathbf{M}^T(\mathbf{M}\mathbf{P}_{(k)}^{f}\mathbf{M}^T + \mathbf{R}_{(k)})^{-1}$$

for $i = 1$ *to* N_{ens} **do**

$$\mathbf{y}_{(k)}^{(i)} = \mathbf{y}_{(k)} + \mathbf{e}_{(k)}^{o(i)} \text{ with } \mathbf{e}_{(k)}^{o(i)} \sim \mathcal{N}(0, \mathbf{R}_{(k)})$$

$$\mathbf{x}_{(k)}^{a(i)} = \mathbf{x}_{(k)}^{f(i)} + \mathbf{K}_{(k)} \left(\mathbf{y}_{(k)}^{(i)} - \mathcal{D}(\mathbf{x}_{(k)}^{f(i)}, V = v_{(k)})\right)$$

6.3 Description of the wind-turbine numerical model

Dynamic analysis of wind turbines involves strong interactions between the turbines' aerodynamics, the control system, and the structural mechanics. The main solicitations are the environmental conditions and the rotating machinery during operating term. In order to model and simulate the nonlinear response of wind turbine structures under such solicitations, various servo-aero-elastic software have been developed, such as OpenFAST

[NREL, 2018], Bladed [DNV GL, 2013], HAWCK2 [Larsen and Hansen, 2007] or Deeplines Wind™ [Principia].

In our study, a simulator of a Senvion MM82 wind turbine has been developed using Deeplines Wind™ software from technical specifications. This software is a fully coupled simulation tool taking into account the aerodynamics of the aero-generator, the elasticity of the structural wind turbine components (mast, blades and drive-train systems), and the control system [Le Cunff et al., 2013]. The software architecture developed by IFPEN [2] and Principia [3] is fully modular with different dynamic libraries (DLL) called by the solver. The integration in time is performed with an implicit Newmark integration scheme. The developed simulator includes a nonlinear beam finite element formulation to model the structural components. The aerodynamic loads acting on the turbine rotor are dynamically computed by employing the Blade-Element Momentum (BEM) theory for Horizontal Axis Wind Turbine (HAWT). A Deeplines Wind™ software validation, based on code comparisons [Perdrizet et al., 2013], has shown accurate results in various conditions.

Wind turbine simulation consists of two stages: first the generation of the input turbulent wind field and then the fully coupled servo-aero-elastic simulation. The generation of the input stochastic process is done by using a simulator called Turbsim [Jonkman, 2009]. This simulator has some deterministic inputs such as the turbulence intensity, the mean wind speed, the mean flow angles, the spectrum and the spatial correlation model. In our model, we have used an IEC [4] Kaimal turbulence spectrum with an exponential spatial coherence. Nevertheless, these deterministic values cannot uniquely determine a stochastic wind field and a pseudo-random number generator has to be used in order to create random phases for the velocity time. Then structural responses are time computed with a multi-physics numerical code such as Deeplines Wind™, following the formalism introduced in Equation (6.1). A wind turbine structure can encounter a variety of operating conditions. Each of the operating conditions, modeled by the stochastic process V, is parameterized by measurable deterministic quantities mentioned in Figure 6.1. In this paper we will perform the study at an under-rated average wind speed of 8 m/s corresponding to the most common operating regime of the considered turbine. All computed responses are based on 10-minute effective simulations of the MM82 Senvion wind turbine. By effective, we mean that the transient start-up behavior is removed from the analysis. The transient start-up behavior can be decomposed in a ramp time wind and an oversight periods. The oversight period has been set according to auto-correlation studies of the outputs. This period permits to remove the effect of the ramp time period on the numerical model responses. A numerical simulation lasts 15 minutes on an Intel Xeon Scalable Gold 6140 processor operating at 2.3 GHz.

From the structural time responses computed by the Deeplines Wind™ model, we obtain some QoIs describing the fatigue behavior of the wind turbine. They are obtained by post-treating the resulting time series of internal loads at different locations of the analysed design, see Figure 6.2 and Table 6.1. In our study, we denote by g the function mapping the functional loads of the structure to the damage QoIs, see Equation (6.2). Each fatigue QoI has been estimated by using the damage-equivalent load (DEL). The DEL is computed for a set of parameter values and different realizations of the stochastic

2. see https://www.ifpenergiesnouvelles.fr/
3. see http://www.principia-group.com/
4. International Electrotechnical Commission

process V. It is defined as the regular load amplitude that would create in N_{ref} cycles the same fatigue as the considered irregular load history. The DEL is computed based on the Palgrem Miner's rule [Sutherland, 1999]:

$$DEL = \left(\frac{\sum_{i=1}^{N_c} S_i^m \cdot n_i}{N_{ref}} \right)^{\frac{1}{m}} . \tag{6.12}$$

where $i = 1, \ldots, N_c$ corresponds to each range bin, S_i is the cycle range value and n_i is the number of cycles for the i-th bin. The exponent m is the negative inverse slope of the cyclic stress against the cycles to failure curve (S-N curve) and N_{ref} is the reference number of cycles usually set to an arbitrary value. The cycles in an irregular load history are computed using the Rainflow counting method [Cosack, 2011].

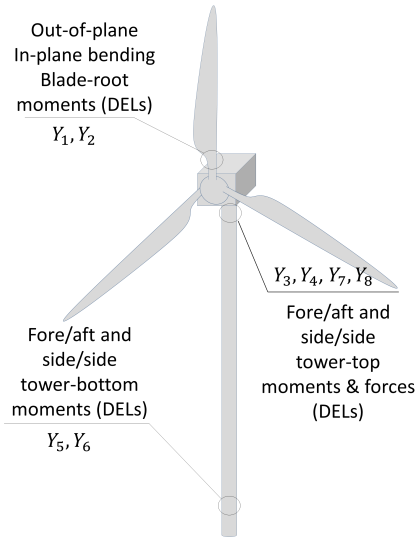


Figure 6.2 – Recorded time series of loads at different locations of the wind turbine

Load's position	m	Type of load	
Blade root moments	10	Out-of-plane bending	In-plane bending
Tower top moment	3	Fore-aft bending	Side-to-side bending
Tower bottom moment	3	Fore-aft bending	Side-to-side bending
Tower top force	3	Fore-aft force	Side-to-side force

Table 6.1 – Fatigue damage equivalent loads used for the global sensitivity analyses with their corresponding Wöhler's exponent, i.e., the negative inverse slope of the S-N curve.

In order to ensure that the variation induced by the input parameters is distinguishable from the one induced by the realization of the stochastic process V, multiple wind realizations have to be generated. A convergence study has been led in order to determine the number of realizations needed to encompass the variation of the selected realization of the stochastic process. The QoIs analyzed in this study are the DEL at different locations of the wind turbine, see Equation (6.12). The number of stochastic process realizations used during the convergence study varies from 1 to 30.

Figure 6.3 shows the convergence of the tower bottom fore-aft bending moment DEL on the number of realizations used for their averaging estimation. We assume a 10-minute mean hub-height wind speed of 8 m/s and a wind fluctuation following the Kaimal IEC. The IEC 61400-1 standard has three turbulence categories: A, B and C, [see IEC, 2005]. We have decided to use the A class corresponding to the highest turbulence intensity, i.e., the ratio of standard deviation of fluctuating wind velocity to the mean wind speed is around 24 %. A time series respecting the 10-minute statistics, which will be different at each generation, is created thanks to the simulator [Jonkman, 2009]. According to

certification guidelines, see design load case 1.2 in [DNV GL, 2016], the fatigue analysis has to be led with 6 wind realizations of 10 min time period. Nevertheless, as we can underline with the last mentioned figure, the empirical mean computed from these 6 realizations does not seem to be a reliable estimator. In other words, 6 simulations are not sufficient for QoIs' statistics to converge. With our industrial numerical model, a compromise has been made to balance the quality of the empirical estimator and the computing time goal by fixing the number of realizations to 10 for the GSA of the damage equivalent loads.

The aim of this paper is to identify and reduce the structural sources of uncertainty on the fatigue QoIs. We will focus our study on some wind turbine properties represented by 13 parameters gathered in the vector $\mathbf{x}$. A literature review has been done to specify the uncertainty in the parameter values. Based on expert knowledge, all these parameters were considered independent of one another with Gaussian distributions due to their physical properties.

Here follows a description of the 13 considered parameters, see Table 6.2. For the support structural properties, six parameters have been considered, including nacelle mass, nacelle center of mass, tower Rayleigh damping, inertial nacelle and drive-train torsion stiffness. The tower thickness has been changed by uniformly scaling the distributed tower thickness. The boundary values have been set by changing the first fore-aft tower frequency mode by $\pm 10\%$ of its reference value.

The uncertainties in blade structural properties have been represented using five parameters. The blade structural responses have led to the definition of the uncertainty range. Indeed, the frequency of the edge-wise (ES) and flap-wise (FS) mode were changed about 10% each from their nominal value. These modifications of the mode were done by uniformly scaling the associated stiffness and the distributed blade mass of all blades. Blade imbalance effects have been also included by applying a different change value to each blade. One blade is modified to be a value that is higher than the nominal value, and another one modified to a lower value. The third blade remains unchanged at the nominal value.

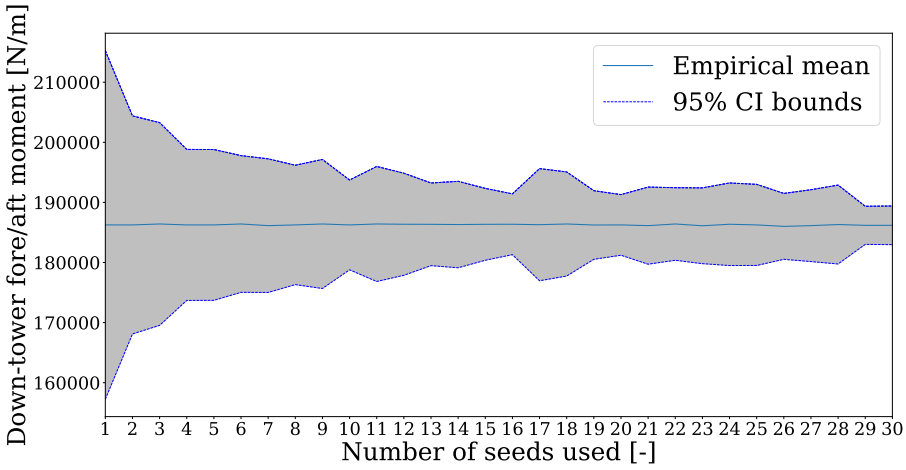


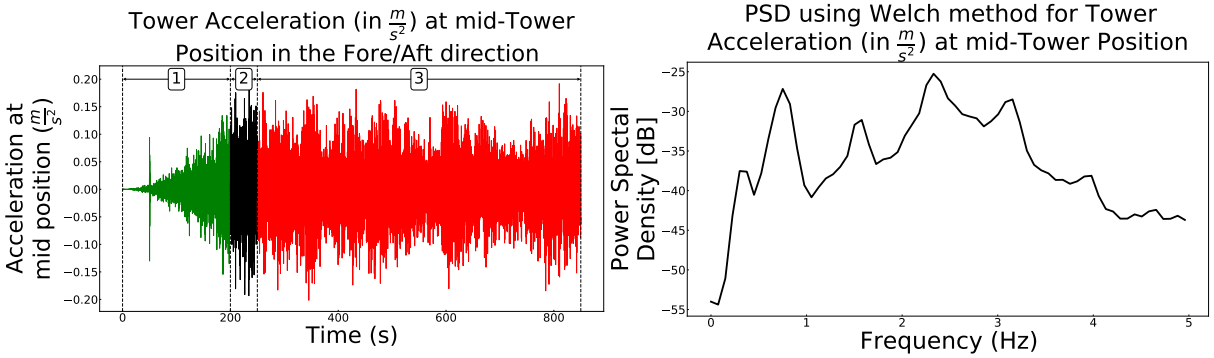
Figure 6.3 – Convergence of the tower bottom fore-aft bending moment DEL as a function of the number of turbulent seeds used for its evaluation. 95% confidence interval around the estimated empirical mean is also represented (grey area). The wind turbulence intensity is set around 24 %, [see IEC, 2005].

For the aerodynamic properties, we have considered the wind turbine yaw misalignment by changing the yaw angle of the turbine. For the individual blade pitch error, a constant offset angle is applied to two of the blades, respectively above and below nominal value.

Table 6.2 – Structural properties - uncertainties affecting the input parameters of the wind turbine model.

Input parameter	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	REF
Nacelle Mass - N_{mass} [kg]	6.90e+04	2.30e+03	[Witcher, 2017]
Nacelle center of mass - N_{CMx} [m]	1.00	3.33e-02	[Robertson et al., 2019b]
Tower thickness - e [%]	0	7	IFPEN $\pm$10% 1 FA
Tower Rayleigh Damping - β_{TR} [-]	3.10e-02	9.93e-03	[Koukoura, 2014]
Inertial Nacelle - I_{zz} $[kg \cdot m^2]$	7.00e+05	2.33e+04	IFPEN $\pm$10% μ
Drive-train Torsional stiffness - K_D $[\frac{N \cdot m^2}{rad}]$	4.45e+09	1.48e+08	[Holierhoek et al., 2010]
Blade Flap wise Stiffness - α_{BF} $[N \cdot m^2]$	1.00	3.33e-02	IFPEN $\sim \pm$10% 1 FS
Blade Edge wise Stiffness - α_{BE} $[N \cdot m^2]$	1.00	3.33e-02	IFPEN $\sim \pm$10% 1 ES
Blade mass coefficient - α_{mass} [%]	1.00	1.67e-02	[Witcher, 2017]
Blade Rayleigh Damping - β_{BR} [-]	5.39e-03	1.45e-03	[Robertson et al., 2019b]
Blade mass imbalance - η_B [%]	2.50	8.33e-01	[Robertson et al., 2019b]
Yaw misalignment - ω [°]	0	6.67	[Quick et al., 2017]
Individual pitch error - Ω [°]	0.10	3.33e-02	[Simms et al., 2001]

After an appropriate sensitivity analysis leading to the identification of the less influential input parameters on the fatigue QoIs, we can fix their value to a nominal one without affecting the fatigue behavior of the structure. Then, the uncertainties of the other parameters has to be reduced by employing an appropriate inference method based on in situ measurements. In this context, let us consider a wind turbine instrumented with accelerometers. We assume that bi-axial measuring devices are located at mid and top tower height position. From these sensors, we can record four functional acceleration time series. Then, the power spectral density (PSD) of each measured acceleration time series is computed using Welch's method [Welch, 1967], see Figure 6.4.



(a) **(b)**

Figure 6.4 – On the left side (a): simulated acceleration in $\frac{m}{s^2}$ of the wind turbine tower in the fore-aft direction obtained at the accelerometer device located at mid-tower decomposed in a ramp time wind $\boxed{1}$, an oversight period $\boxed{2}$ and a dynamical period of interest $\boxed{3}$. On the right side (b): estimated PSD of the period of interest using Welch's method [Welch, 1967].

6.4 Illustration of the proposed framework on the wind-turbine numerical model

6.4.1 GSA of the fatigue QoIs

The Deeplines Wind™ numerical model presented in Section 6.3 is used with 13 uncertain input parameters, each one having its associated probability distribution, see Table 6.2. The total Sobol' indices associated to these parameters for each DEL have been estimated using the heteroscedastic noisy GP model-based Sobol' index procedure as described in Section 6.1. A Latin Hypercube Sampling (LHS) of size 500 has been used to emulate the numerical model. Then, an augmented LHS of size 150 has been generated to determine the accuracy of the surrogate models. To apply this approach 6,500 forward wind turbine numerical simulations were submitted on the 206 TFlops supercomputer of IFPEN.

The function *sobolGP* performs a kriging-based GSA by taking into account the complete conditional predictive distribution of the surrogate model. The function estimates total Sobol' indices thanks to the Jansen estimators, [see Jansen, 1999]. Jansen Sobol' index estimators are accurate for large and small total indices. Moreover, by taking into account the complete conditional predictive distribution of the trained surrogate model in the estimation procedure of the total Sobol' indices, we can evaluate the uncertainty due to Monte Carlo estimation, but also due to metamodeling [Le Gratiet et al., 2013]. The results for the total Sobol' indices with their corresponding 95% confidence intervals are summarized in Figure 6.5. A threshold of 2.5e-2 was chosen to display a separation between input parameters with high and low total Sobol' indices. Figure 6.6 represents the different sources of uncertainty for the estimation of total Sobol' indices, obtained thanks to the function *sobolGP* for the DEL of the out-of-plane bending blade-root moment.

By considering all the total Sobol' indices obtained for the different fatigue QoIs presented in Figure 6.5, we can notice that only 6 parameters have indices values greater than the threshold. Consequently, we can fix other parameters to any specific value in the range of variability without affecting the fatigue QoIs. During the recursive Bayesian estimation, these non-influential parameters will be arbitrarily set at their mean value described in Table 6.2. By employing this method, we will reduce the parameter dimension in the inference procedure without affecting the fatigue QoIs which are crucial for assessing the lifespan of wind turbines.

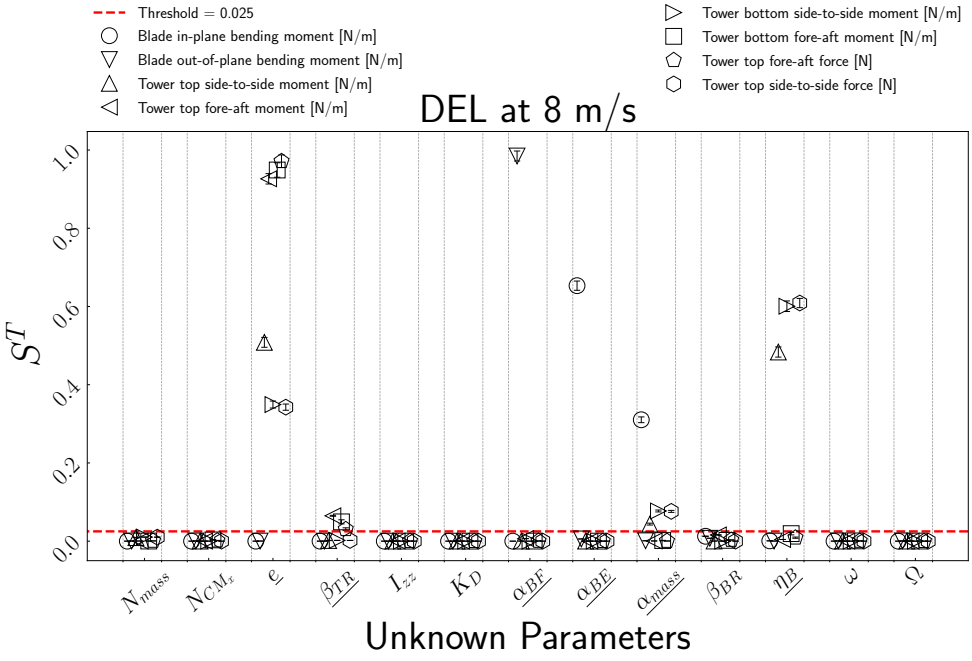


Figure 6.5 – Total Sobol' index estimates (y-axis) with their 95% confidence interval (CI) corresponding to each of the 13 inputs (x-axis) for the different fatigue outputs. The dashed line is a threshold arbitrarily chosen to 2.5e-2. CIs are obtained by taking into account the uncertainties due to both the surrogate and the Monte Carlo (MC) estimation. The number of samples for the conditional Gaussian process, to quantify the uncertainty of the kriging, was set to 100. The one due to MC integration was computed by bootstrapping with 100 samples.

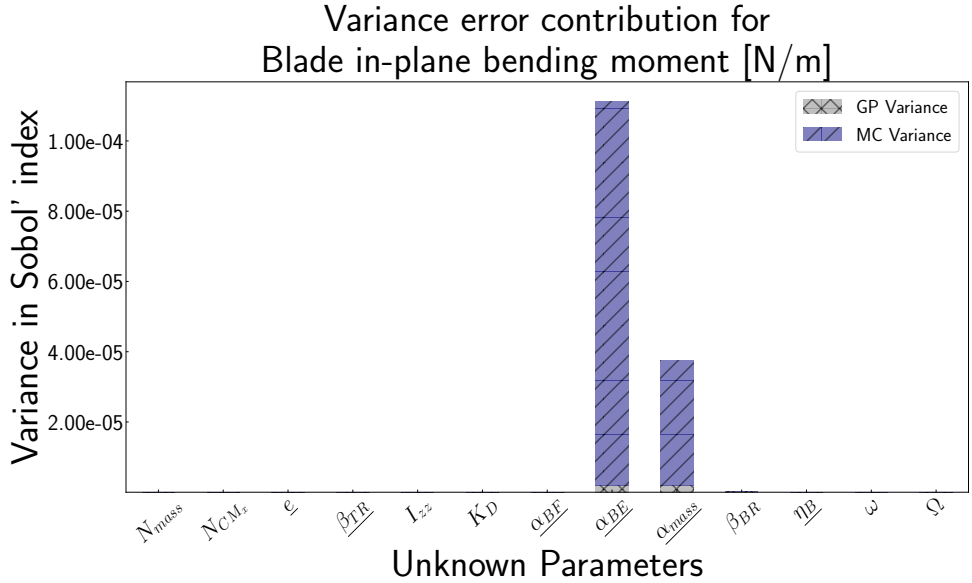


Figure 6.6 – Splitting of the variance of total Sobol' index estimators (y-axis) corresponding to each of the parameters (x-axis) for the out-of-plane bending blade-root moment DEL. The number of samples for the conditional Gaussian process, to quantify the uncertainty of the kriging approximation, was set to 100. The one due to Monte Carlo integration was computed by bootstrapping with 100 samples.

6.4.2 Identifiability study

It is possible that the considered experimental measurement settings do not offer enough information for the identification of some input parameters. In this context, another interesting property of the GSA underlined in Proposition 4.2 in [Dobre et al., 2012] is that nullity of the total sensitivity index for a specific input parameter implies its non-identifiability from the output under consideration. Thus, a GSA led on the measured outputs will determine which parameters cannot be inferred, although it is not a sufficient condition for identifiability.

In our industrial application study, the measured outputs, obtained from the accelerometers, are expressed as discretized time series. We are interested in their response in the frequency-domain by using the power spectral density (PSD). Discretized PSD series involve a substantial dimensionality and a high degree of redundancy. To bypass this issue, we have firstly focused our study on an orthogonal decomposition, of the different discretized PSDs, in order to reduce their dimensionality. This orthogonal decomposition will be performed by a data-driven method called Principal Component Analysis (PCA) [Wold et al., 1987]. PCA allows the functional output expansion in a new reduced space spanned by the most significant directions in term of variance of the discretized functional output.

In our study, a method based on PCA and GSA with GP model is used to compute an aggregated Sobol' index for each input parameter of the model. As described in [Lamboni et al., 2011], the proposed index synthesizes the influence of the parameter on the whole time series output.

In our study to ensure that the variation of the input parameters is distinguishable from the realization of the stochastic process V, 10 wind realizations have been used in this GSA. A new LHS of size 300 with a geometrical criteria maximizing the minimum distance between the design points has been used to emulate the numerical model. In Table 6.3, we summarize the total aggregated Sobol' indices obtained with the GP model built on the trained set from the lastly mentioned LHS. In this analysis, parameters with total Sobol' index values under a threshold set at 1e-02 will be considered as non-identifiable from the measured output. We can conclude that none of the significant input parameters can be considered a-priori as non-identifiable.

Table 6.3 – Total Sobol' indices for each output used during the recursive inference procedure described in details in Section 6.2. Estimated total Sobol' indices higher than the arbitrary threshold are underlined.

Measured outputs	e [%]	β_{TR} [−]	α_{BF} [%]	α_{BE} [%]	α_{mass} [%]	η_B [%]
Tower middle fore-aft acceleration's PSD	2.44e-01	7.64e-01	6.07e-03	1.37e-04	3.59e-04	4.46e-03
Tower middle side-to-side acceleration's PSD	3.84e-01	3.95e-01	1.38e-01	6.73e-04	8.60e-02	7.17e-03
Tower top fore-aft acceleration's PSD	1.21e-01	6.70e-01	2.09e-03	5.91e-02	6.76e-02	1.05e-01
Tower top side to side acceleration's PSD	6.56e-02	6.24e-01	1.36e-03	9.67e-02	1.39e-01	9.52e-02

6.4.3 Recursive Bayesian inference study

The 6 input parameters having an influential effect on the fatigue behavior of the structure and potentially identifiable are considered during the inference procedure. These unknown input parameters define the model parameter vector to be estimated, i.e., $\mathbf{x} = [e,\ \beta_{TR},\ \alpha_{BF},\ \alpha_{BE},\ \alpha_{mass},\ \eta_B]^T$.

In this section we focus on pseudo-experimental numerical tests in order to validate the inference procedure. These tests consist in running direct numerical analyses considering known values of input parameters, and then adding a Gaussian noise of known variance to the observed outputs. The dynamic response of the wind turbine structure is simulated using the mean values of the unknown model parameters described in Table 6.2. The noisy pseudo-experimental outputs used to recursively update the wind turbine model are the PSD of the acceleration time series obtained for side to side and fore-aft at the two different tower positions.

Algorithm 5 is used to recursively estimate the expected values of the unknown input parameters at each iteration step. The output measurement noise covariance matrix $\boldsymbol{R}_k$ is assumed to be diagonal, i.e., cross-correlations between the noise components are disregarded. Usually, the amplitudes of the measurement noise can be estimated based on the characteristics of the used sensors. Nevertheless, in our pseudo-experimental numerical case, these amplitudes have been arbitrarily chosen. Indeed, to mimic real-life applications, noise is incorporated in the simulated data by considering a noise covariance matrix such as the obtained standard deviation is equivalent to a 1% signal-to-noise ratio.

For the initialization of the Bayesian estimation procedure, the initial prior of the value of the input parameters is assumed to be:

$$\mathbf{x}_b = [-0.10,\ 4.00e-02,\ 0.98,\ 1.05,\ 9.85e-01,\ 0.04]^T \cdot$$

The initial error covariance matrix of the input parameters, denoted by $\boldsymbol{P}_b$, is also assumed to be diagonal. In other words, the initial prior of the input parameters are assumed to be statistically independent. Diagonal elements of $\boldsymbol{P}_b$ represent the practitioner's belief on the input parameters uncertainties, such as $\boldsymbol{P}_b = \mathrm{diag}(7.00e-02,\ 9.93e-03,\ 3.33e-02,\ 3.33e-02,\ 1.67e-02,\ 8.33e-03)$.

The number of inference iterations T and of the number of members N were respectively fixed at $T = 20$ and $N = 100$. This choice was mainly guide by the maximal simulation budget allocated to our study and by the fact that modest ensemble size is a reasonable practice as observed in industrial setups [Evensen, 2009, Eknes and Evensen, 2002, Houtekamer et al., 2005]. Figure 6.7 shows the results of the identification. It can be observed that the considered input parameters are well recovered. Table 6.4 reports the final a posteriori estimate of the input parameters.

Table 6.4 – Target, initial prior and final a posteriori estimates of the input parameters of the wind turbine numerical model.

	e [%]	β_{TR} [–]	α_{BF} [%]	α_{BE} [%]	α_{mass} [%]	η_B [%]
Target	0	3.10e-02	1	1	1	2.50e-02
Prior estimates	-1.00e-01	4.00e-02	9.80e-01	1.05	9.85e-01	4.00e-02
A posteriori estimates (T=20)	3.26e-03	3.08e-02	9.98e-01	1.00e+00	1.00e+00	2.52e-02

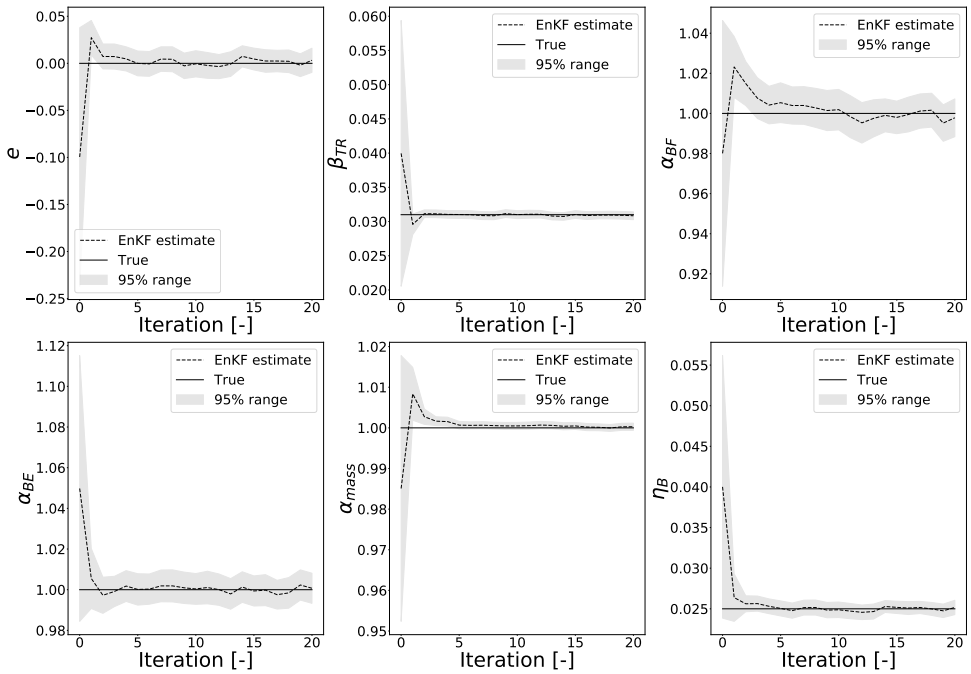


Figure 6.7 – Iteration evolution of the a posteriori estimates of the parameters. Results obtained by running EnKF presented in Section 6.2 with $N = 100$ members of the ensemble used for the estimation and considering pseudo-experimental measures.

6.4.4 Robustness analysis

To test the effectiveness of the proposed EnKF procedure, different noise levels affecting the synthetic data have been considered. We have chosen different structures of noise covariance matrices such as the obtained standard deviations affecting the measurements are respectively equivalent to 3% and 5% signal-to-noise ratios. The performed analysis have highlighted that the incorporation of higher noise leads to a harder identification of input parameters. The estimation of these parameters is less reliable because their identifiability property becomes weaker. The issue of identifiability is a crucial aspect due to the presence of noise in real-life applications. However, the proposed recursive Bayesian inference method has the ability to give confidence intervals on the inferred parameters due to its probabilistic framework.

Conclusion

This chapter presents a framework to quantify and reduce the uncertainties from the input parameters of a wind turbine numerical model.

The contributions of this paper are twofold. First, we have proposed a global sensitivity analysis based on Sobol' indices using a Gaussian process model with heteroscedastic nugget effect to quantify uncertainties of a stochastic and time-consuming wind turbine numerical model. The procedure we present is efficient to balance the inherent uncertainty of the stochastic numerical model and the one from the input parameters. More precisely the GSA has been performed on the fatigue quantities of interest and showed that only a restricted number of unknown parameters happens to influence these responses. Since fatigue quantities of interest are a crucial wind turbine design and life criteria, these influential input parameters have to be properly estimated in order to give confidence in fatigue analysis results.

Consequently, the second objective of this paper was to develop a recursive inference procedure to properly determine these parameters based on available measurements. But first was addressed the question of parameter non-identifiability by employing a global sensibility study on the measured outputs. As previously stated, performing such sensitivity analysis is not a sufficient condition for identifiability. Finally for the inference task, this paper demonstrates the applicability and computational efficiency of the ensemble Kalman filter (EnKF) for this type of challenging problem. The EnKF-based approach has been integrated into the commercial software Deeplines Wind™. The proposed methodology was verified using numerically simulated response data. For future work, the recursive Bayesian estimation procedure will be further tested by incorporating other measured output data.

7

Wind turbine quantification and reduction of uncertainties based on a data-driven data assimilation approach

Ce chapitre propose une procédure de quantification et de réduction des incertitudes impactant les simulations numériques utilisées pour estimer la fatigue d'une structure éolienne. L'étude présentée repose sur les travaux menés par Hirvoas et al., où une quantification et une réduction des incertitudes liées aux propriétés du modèle de l'éolienne sont réalisées à la fois par une analyse de sensibilité globale et une approche de filtrage Bayésien d'ensemble. Nous étendons la procédure aux paramètres incertains considérés lors de la modélisation d'un champ de vent synthétique nécessaire pour mener à bien une simulation aéro-servo-élastique en se basant sur une approche par vecteur d'état augmenté. Néanmoins, contrairement aux paramètres relevant des propriétés du modèle de l'éolienne qui évoluent lentement ou pas, ceux rattachés à la sollicitation extérieure ont un caractère dynamique qui doit être pris en compte lors de l'inférence récursive. Le travail proposé dans ce chapitre consiste à remplacer le modèle dynamique inconnu et utilisé dans la méthode d'assimilation de données par des simulations statistiques basées sur une base de données. Nous nous intéressons tout particulièrement aux méthodes d'assimilation de données par analogues, qui consistent à combiner les méthodes analogues et une méthode de filtrage stochastique tel que le filtre de Kalman d'ensemble [Tandeo et al., 2015, Lguensat et al., 2017]. Cette approche d'assimilation de données dite basée données (data-driven) est évaluée sur un cas industriel d'une éolienne en opération. Les données mesurées sont exploitées par la méthode pour récursivement réduire les incertitudes qui entachent les paramètres à la fois liés aux propriétés du modèle et à la modélisation d'un champ de vent synthétique. Le plan du chapitre est le suivant. Section 7.1 décrit les différentes incertitudes considérées dans cette étude. Dans la Section 7.2, le cadre théorique de l'assimilation de données dite basée données (data-driven) avec un intérêt spécifique pour la méthode de filtrage d'ensemble de Kalman couplée à la stratégie de prévision par analogue est détaillé. Enfin, les résultats d'une application de cette procédure de quantification et de réduction des incertitudes à une éolienne de référence sont présentés dans la section 7.3.

Introduction

A major challenge in wind energy industry is to propose robust designs withstanding on known environmental conditions. Design standards [IEC, 2005] are mainly based on dynamic load simulations describing the structural behavior of the wind turbine under different wind and operational conditions weighted by their probability of occurrence. Most of the time the number of wind scenarios considered during the conception phase is moderate and far from exploring the set of environmental conditions. Moreover, the dynamic response of the structure and its lifetime can be affected by some uncertainties or evolution in the wind turbine properties. Consequently, the prediction of the lifetime consumption of the operating wind turbine by taking into account all the inherent uncertainty is crucial. In that context, the quantification and reduction of uncertainties involved in the aero-servo-elastic numerical models play an important role to determine the effective fatigue loads of the turbine.

The approach intoduced in this paper generalizes the one in [Hirvoas et al.] by taking into account the uncertainties affecting the parameters related to the wind inflow on top of the parameters of the structure properties. It relies on a complete framework including a global sensitivity analysis, an identifiability analysis, and a recursive Bayesian inference approach. First, a global sensitivity analysis based on the estimation of Sobol' indices thanks to surrogate models allows to determine the most relevant input parameters in the variability of the fatigue loads of a wind turbine. After assessing the identifiability properties of these influential parameters, a second objective is to reduce their uncertainty by using an ensemble Kalman filter. Data assimilation allows to gather all the information obtained from real time measurements of the physical system and from the numerical model. The procedure is closely related to the industrial concept of digital twin which consists in combining measurements from the wind turbine with a numerical model to build a digital equivalent of the real-world structure. However, unlike the model properties having a static or slow time-variant behavior, the parameters related to the external conditions have a dynamic that has to be learnt from data.

Modern wind turbines in production are currently monitored thanks to a large amount of sensors. Then, data monitored by sensors can be used to learn the non-explicit dynamic behavior of the inflow related parameters. In the present work, we focus on non-parametric learning strategies. In the literature, several non-parametric methods have been developed such as regression machine learning [Brunton et al., 2016], echo state networks [Pathak et al., 2018] or more recently residual neural networks [Bocquet et al., 2020]. Our study investigates an analog forecasting method relying on nearest neighbors principle [Lorenz, 1969]. The aforementioned non-parametric procedure has been firstly coupled with data assimilation filtering schemes in [Tandeo et al., 2015] and further detailed by Lguensat et al. [2017]. In the present work, we propose an algorithm interfacing Python library AnDA [1] combining analog forecasting with ensemble data assimilation, with the algorithms developed in [Hirvoas et al.]. The algorithm we propose takes profit of the parallelization capabilities of the current high performance computing architectures which allows for example to evaluate the real-time damage of an operating wind turbine using a digital twin.

The outline of this paper is as follows. Firstly, Section 7.1 describes the different uncertainties involved in the framework of this study. In Section 7.2, the theoretical

1. see https://github.com/ptandeo/AnDA

framework of data-driven data assimilation with a specific focus on the ensemble Kalman filtering scheme coupled with the analog forecasting strategy is detailed. Finally, results of an application of this complete procedure of uncertainty quantification and reduction to a reference wind turbine are presented in Section 7.3.

7.1 Context

Before their exploitation, wind turbine rotors are designed thanks to a site classification strategy. It relies on design standard classes characterized by the reference turbulence intensity I_{ref} defined as the mean turbulence intensity expected at 15 m/s mean wind speed and the reference wind $\overline{u}_{ref}$ defined as the extreme 10-minute average wind speed with a recurrence period of 50 years. In the IEC-61400-1 standard [IEC, 2005], two safety classes are considered. The first one, named as normal safety class, allows to cover most applications by giving specific values for I_{ref} and $\overline{u}_{ref}$. In Table 7.1, the corresponding values for the nine categories of the normal safety are given. The proposed parameter values are supposed to represent many different sites and consequently do not give a precise representation of a specific site. The second category is mentioned as a special safety class S which allows to consider site-specific values for the wind speed and turbulence terms.

Table 7.1 – Safety class design classification of the wind turbines: the normal safety class containing nine categories from I-A to III-C and the special safety class S [IEC, 2005]

Class		I	II	III	S
$\overline{u}_{ref}$	[m/s]	50	42.5	37.5	
A	I_{ref} [-]	0.16			Site-specific values
B	I_{ref} [-]	0.14			
C	I_{ref} [-]	0.12			

For both classes, the design relies on numerical aero-servo-elastic simulations under different environmental and operational conditions, weighted by the probability of occurrence. They allow to estimate the ultimate and fatigue loads in order to testify the structural integrity. Nevertheless, operating wind turbines experience real wind and operational conditions that are different from the ones mentioned in the design standard classes. Consequently, there is a need for an estimation of the remaining fatigue life of the components based on the real wind solicitation seen by the structure. Moreover, the wind turbine itself can present some uncertainties or evolution in its mechanical properties (defaults appearance, degradation with time) that will affect the dynamic response of the structure and its lifetime.

As a consequence, these aero-servo-elastic numerical models involve many uncertain and potentially variable over time parameters. The ubiquitous uncertainty may be found in the parameters of the wind turbine numerical model as well as in the external conditions. To ensure the tracking of fatigue and defaults of an operating wind turbine structure, it is important to quantify the impact of these uncertainties on predictions and then to reduce them based on the combination of measurements and model predictions. For that purpose, the field of uncertainty quantification is well-adapted.

In that context, we propose to determine the sources of uncertainties affecting the wind field parameters and the wind turbine numerical model properties using an aug-

mented state vector approach. First, the uncertainty of wind field parameters has to be determined. In our context, these parameters are used to characterize a synthetic three-dimensional turbulent wind field based on the Kaimal spectrum with an exponential coherence model, see Section 2.1 for details. Eight input parameters related to the wind field have been identified to be tainted by uncertainties, see Table 7.2. We have considered the mean and the standard deviation of the wind speed at hub height, the vertical wind shear exponent, the mean wind inflow direction relative to the wind turbine in terms of vertical or horizontal inflow angles, and the longitudinal turbulence length scale parameter. Moreover, we have supposed as unknown the input coherence decrement and offset parameter, see Equation (2.3).

In an operational context, some information on the mean and standard deviation of the wind speed at hub height can be obtained from 10-minute data measured from a nacelle mounted cup-anemometer. Nevertheless, these measurements are known to be very perturbed and never fully describe the parameters of interest due mainly to the wake effect of the rotor and the non-perfect transfer function used to retrieve them. In this work, we assume that the 10-minute mean and standard deviation wind speed can be obtained from the 10-minute data obtained from the cup-anemometer modulo an additive error term. So that the mean wind speed at hub height based on the anemometer can be express as:

$$\overline{u} = \overline{u}_{scada} + \Delta\overline{u}\,,$$

where $\overline{u}_{scada}$ is the 10-minute mean wind speed obtained from the cup-anemometer mounted on the wind turbine nacelle and $\Delta\overline{u}$ is an additive error assumed to follow the distribution defined in Table 7.2.

In a similar manner, the wind speed standard deviation can be obtained from the measurement obtained by the cup-anemometer mounted on the nacelle of the wind turbine as:

$$\sigma_u = \sigma_{scada} + \Delta\sigma_u\,,$$

where σ_{scada} is the 10-minute standard deviation wind speed obtained from the nacelle cup-anemometer of the wind turbine nacelle and $\Delta\sigma_u$ is an additive error assumed to follow the distribution defined in Table 7.2.

Unless having high frequency Supervisory Control And Data Acquisition (SCADA) data, no information can be obtained on the other parameters. Consequently, an investigation of the distribution of the uncertainty affecting these remaining wind inflow parameters has to be properly made. Table 7.2 summarizes the wind-inflow parameters that we consider unknown and their respective uncertainty modeling. In particular, we adapt the Gaussian distribution proposed by Dimitrov et al. [2015] for the 10-minute vertical wind shear exponent, such as:

$$\begin{aligned} \mu_\alpha &= 0.088(\ln(\overline{u}_{scada}) - 1) \\ \sigma_\alpha &= 1/\overline{u}_{scada} \end{aligned}\,. \tag{7.1}$$

Table 7.2 summarizes the wind-inflow parameters that we consider unknown and their respective uncertainty modeling.

Table 7.2 – Wind field parameters - uncertainties affecting the inputs of the wind turbine model. $\mathcal{U}$: uniform distribution and $\mathcal{G}$: Gaussian distribution.

Input	Variable	Unit	Distribution	Parameters	REF
Error of hub mean wind speed SCADA vs undisturbed inflow	$\Delta\overline{u}$	[m/s]	$\mathcal{U}$	Min: $-0.1 \cdot \overline{u}_{scada}$ Max: $0.1 \cdot \overline{u}_{scada}$	IFPEN
Error of hub standard deviation SCADA vs undisturbed inflow	$\Delta\sigma_u$	[m/s]	$\mathcal{U}$	Min: $-0.2 \cdot \sigma_{scada}$ Max: $0.2 \cdot \sigma_{scada}$	IFPEN
Vertical wind inflow angle	ϕ_v	[°]	$\mathcal{U}$	Min: 0 Max: 10	IFPEN
Horizontal wind inflow angle	ϕ_h	[°]	$\mathcal{U}$	Min: -15 Max: 15	IFPEN
Longitudinal turbulence length scale	Λ_u	[m]	$\mathcal{U}$	Min: 20 Max: 170	[Dimitrov et al., 2017] [Solari and Piccardo, 2001]
Decrement parameter of coherence model	a	[-]	$\mathcal{U}$	Min: 1.5 Max: 26	[Robertson et al., 2019a]
Offset parameter of coherence model	b'	[-]	$\mathcal{U}$	Min: 0 Max: 0.17	[Robertson et al., 2019a] [Saranyasoontorn et al., 2004]
Vertical wind shear exponent	α	[-]	$\mathcal{G}$	$\mu = \mu_\alpha$ $\sigma = \sigma_\alpha$, see Equation (7.1)	[Dimitrov et al., 2015]

Moreover, as suggested in [Hirvoas et al.], a total of twelve parameters can be considered as uncertain in the aero-servo-elastic wind turbine numerical model properties. All these input parameters are assumed to be independent of one another with Gaussian or truncated Gaussian distributions obtained from expert knowledge or literature. Considering the support structural properties of the turbine model, we have selected six parameters such as nacelle mass and center of mass, tower Rayleigh damping, inertial nacelle and drive-train torsion stiffness. Lastly, the geometry of the tower, resulting from fabrication tolerances, has been also included in these uncertainties by uniformly scaling the distributed tower thickness. The probability distribution of this last mentioned parameter is determined by changing the first fore-aft tower frequency mode by $\pm 10\%$ of its nominal value. The uncertainties in blade structural properties have been represented using five parameters. The blade structural responses have led to the definition of the uncertainty range. Indeed, the frequency of the edge-wise (EW) and flap-wise (FW) modes are changed about 10% each from their reference value. These modifications of the frequency modes are done by uniformly scaling the associated stiffness and the distributed blade mass of all blades. Blade mass imbalance effects have been also included by applying a different mass factor value to each blade. One blade's mass property is modified to be a value that is higher than the nominal value, and another one modified to a lower value. The third blade remains unchanged at the nominal value. Finally, for the individual blade pitch error, a constant offset angle is applied to two of the blades, respectively above and below the nominal value. These different parameters are considered independent from each other. Table 7.3 gathers information about the probability distribution of each of these paremeters.

In the monitoring context of an operating wind turbine, one of the major challenges is to predict the remaining lifetime of the structure. Hence, the current study focuses on a complete framework first quantifying and then reducing in a recursive fashion the uncertainties affecting the damage loads obtained from an aero-servo-elastic simulation. Hereafter, we will focus on the estimation of the effective damage equivalent load (DEL) describing the fatigue behavior of the wind turbine at some specific locations. The DEL is obtained by considering the internal loads and is defined as a virtual load amplitude that would create, in reference regular cycles, the same damage as the considered irregular load history, see Section 2.5.

The aim of the work in this chapter is to generalize the complete methodology proposed

in [Hirvoas et al.] for quantifying and reducing the uncertainties affecting a wind turbine numerical model by handling wind inflow uncertainties additionally to model property ones. The procedure relies on a global sensitivity analysis (GSA) based on Sobol' index estimation and a recursive Bayesian inference procedure to reduce the uncertainties. In order to alleviate the computational cost of index estimation during the sensitivity analysis of the fatigue loads, the aero-servo-elastic time-consuming numerical model is approximated by a surrogate. A major challenge in building such a surrogate model relies on the fact that the turbulent wind inflow realization causes variations in the quantities of interest obtained from the model. Thus, to take into account the inherent variability on the turbine response induced by different turbulent wind field realizations, the approach focuses on the use of heteroscedastic Gaussian process regression models. Then, a recursive reduction of the influent parameter uncertainties based on an ensemble Kalman filter is proposed. This data assimilation filtering method is computationally efficient with high-performance computing tools which is a major advantage for online calibration of time-consuming codes, such as aero-servo-elastic wind turbine models. Nevertheless, a challenge in this kind of inverse problem is to determine whether the measurements are sufficient to unambiguously determine the parameters that generated the observations, i.e., identifiability properties. In that context, GSA is proposed to detect non identifiable parameters considering the current measurements.

Table 7.3 – Model parameters - uncertainties affecting the inputs of the wind turbine model. $\mathcal{U}$: uniform distribution, $\mathcal{G}$: Gaussian distribution, and $\mathcal{TG}$: Truncated Gaussian distribution.

Input	Variable	Unit	Distribution	Parameters	REF
Nacelle mass	N_{mass}	[kg]	$\mathcal{G}$	$\mu = 6.90e+04 \quad \sigma = 2.30e+03$	[Witcher, 2017]
Nacelle center of mass	N_{CM_x}	[m]	$\mathcal{G}$	$\mu = 1.00 \quad \sigma = 3.35e-02$	[Robertson et al., 2019b]
Tower thickness	e	[%]	$\mathcal{G}$	$\mu = 0 \quad 7.00$	IFPEN $\pm 10\%$ 1 FA
Tower rayleigh damping	β_{TR}	[-]	$\mathcal{TG}$	$\mu = 2.55 \quad \sigma = 0.82$	[Koukoura, 2014]
Inertial nacelle	I_{zz}	$[kg \cdot m^2]$	$\mathcal{G}$	$\mu = 7.00e+05 \quad \sigma = 2.33e+04$	IFPEN $\pm 10\%$ μ
Drive-train torsional stiffness	K_D	$[\frac{N \cdot m^2}{rad}]$	$\mathcal{G}$	$\mu = 9.08e+09 \quad \sigma = 3.03e+07$	[Holierhoek et al., 2010]
Blade flap wise stiffness	α_{BF}	$[N \cdot m^2]$	$\mathcal{G}$	$\mu = 1.00 \quad \sigma = 3.33e-02$	IFPEN $\sim \pm 10\%$ 1 FW
Blade edge wise stiffness	α_{BE}	$[N \cdot m^2]$	$\mathcal{G}$	$\mu = 1.00 \quad \sigma = 3.33e-02$	IFPEN $\sim \pm 10\%$ 1 EW
Blade mass coefficient	α_{mass}	[-]	$\mathcal{G}$	$\mu = 1.00 \quad \sigma = 1.67e-02$	[Witcher, 2017]
Blade rayleigh damping	β_{BR}	[-]	$\mathcal{TG}$	$\mu = 1.55 \quad \sigma = 4.83e-01$	[Robertson et al., 2019b]
Blade mass imbalance	η_B	[%]	$\mathcal{G}$	$\mu = 2.50 \quad \sigma = 8.33e-01$	[Robertson et al., 2019b]
Individual pitch error	Ω	[°]	$\mathcal{G}$	$\mu = 0.10 \quad \sigma = 3.33e-02$	[Simms et al., 2001]

The main contribution of the presented work is the inference of parameters involved in the model properties of the wind turbine having a static or slow evolution, and short-term wind inflow varying at each inference iteration of 10-minute. To take into account the non-explicit dynamics of the parameters related to the wind inflow in the recursive inference procedure, the study relies on a data-driven approach combining a K-nearest

neighbors with an ensemble Kalman filtering scheme. In the next section, we propose to describe this data-driven procedure used in our model calibration strategy.

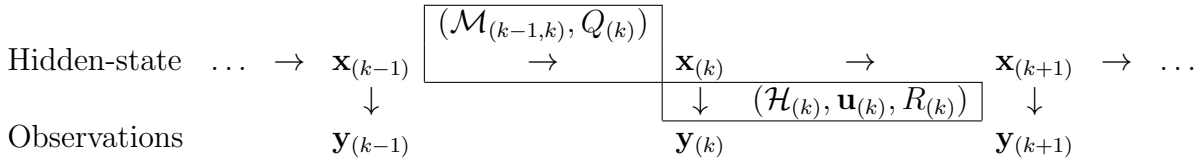
7.2 Data-driven data assimilation

State-space model (SSM) is a useful framework to perform recursive inference strategy such as sequential data assimilation techniques [Bertino et al., 2003, Durbin and Koopman, 2012, Hirvoas et al.]. In order to take into account the information obtained from the SCADA system of the wind turbine, we consider the SSM formulation involving forcing variables defined $\forall k \in \mathbb{N}^*$ as:

$$\mathbf{x}_{(k)} = \mathcal{M}_{(k-1,k)}(\mathbf{x}_{(k-1)}) + \boldsymbol{\epsilon}^m_{(k)}\,, \tag{7.2}$$

$$\mathbf{y}_{(k)} = \mathcal{H}_{(k)}(\mathbf{x}_{(k)}, \mathbf{u}_{(k)}) + \boldsymbol{\epsilon}^o_{(k)}\,. \tag{7.3}$$

where $\mathbf{y}_{(k)}$ corresponds to the observation at step k and $\mathbf{x}_{(k)}$ is a p-dimensional vector representing the hidden-state variables. The model denoted by $\mathcal{M}$ (potentially nonlinear) allows to describe the dynamic behavior of the hidden process. The model error $\boldsymbol{\epsilon}^m_{(k)}$ is supposed to be a Gaussian white noise of zero mean and of covariance $\mathbf{Q}_{(k)}$, modeling the uncertainties related to the dynamics model structure. The propagator $\mathcal{H}$ relates the hidden-state vector to the measured observations and contains some forcing variables $\mathbf{u}_{(k)}$, e.g., mean wind speed obtained from the anemometer of the wind turbine. The sources of errors in the observation model defined in Equation (7.3) are reflected by the Gaussian white noise of zero mean and of covariance $\mathbf{R}_{(k)}$, denoted by $\boldsymbol{\epsilon}^o_{(k)}$, and assumed to be independent of the model error $\boldsymbol{\epsilon}^m_{(k)}$. This SSM formulation can be represented thanks to the directed graph given below.



In many situations, the dynamical model $\mathcal{M}$ is numerically intractable or unknown. In the literature different studies have been conducted to emulate this propagator, used in Equation (7.2), from historical data. Several surrogate techniques have been employed for the reconstruction of nonlinear dynamics model of chaotic system. Authors in [Tandeo et al., 2015] propose a K-nearest neighbors based method, also known as the analog strategy in meteorology or geoscience community. Nevertheless, it has been argued that methods relying on K-nearest neighbors technique are plagued by the curse-of-dimensionality, i.e., fails in very high dimensional applications [Friedman, 1997, Chen, 2009]. Consequently, other non-parametric surrogate modeling approaches have been investigated to learn the underlying dynamics by using for example regression machine learning [Brunton et al., 2016], echo state networks [Pathak et al., 2018] or more recently residual neural networks [Bocquet et al., 2020].

Due to the limited dimension of our inference problem, we have decided to investigate and to use the analog forecasting strategy coupled with data assimilation proposed in [Tandeo et al., 2015, Hamilton et al., 2016, Lguensat et al., 2017]. Analog forecasting is related to the notion of atmospheric predictability introduced by Lorenz [1969]. Later,

this approach has been widely used in several atmospheric, oceanic, and climate studies [Toth, 1989, Alexander et al., 2017, Ayet and Tandeo, 2018]. Hereafter, we detail the principle of analog forecasting technique.

The main idea of the methodology is to substitute the dynamical model in Equation (7.2) by a data-driven model relying on an analog forecasting operator, denoted by $\mathcal{A}$, such as :

$$\forall k \in \mathbb{N}^*, \begin{cases} \mathbf{x}_{(k)} = \mathcal{A}_{(k-1,k)}(\mathbf{x}_{(k-1)}) + \boldsymbol{\epsilon}^m_{(k)} \\ \mathbf{y}_{(k)} = \mathcal{H}_{(k)}(\mathbf{x}_{(k)}, \mathbf{u}_{(k)}) + \boldsymbol{\epsilon}^o_{(k)} \end{cases} .$$

Analog forecasting principle consists in searching for one or several similar situations of the current hidden-state vector that occurred in historical trajectories of the system of interest, then retrieve the corresponding successors of these situations, and finally assume that the forecast of the hidden-state can be retrieved from these successors. Consequently, this strategy requires the existence of a representative catalog of historical data, denoted by $\mathcal{C}$. The reference catalog is formed by pairs of consecutive hidden-state vectors, separated by the same lag [Fablet et al., 2017]. The first component of each pair is named as the analog (denoted by $\mathbf{a}$) while the corresponding state is referred to as the successor (noted as $\mathbf{s}$). The corresponding representative dataset of hidden-state sequences can be written as:

$$\mathcal{C} = \{(\mathbf{a}_i, \mathbf{s}_i), i = [1 \dots P]\}, \text{ with } P \in \mathbb{N}^*.$$

This historical catalog can be constructed using observational data recorded using in-situ sensors but as well as using numerical simulations. Based on this database, the analog forecasting operator $\mathcal{A}$ is a non-parametric data-driven sampling of the state from iteration $k-1$ to iteration k. Three analog forecasting operators have been originally proposed by the authors in [Lguensat et al., 2017]. They are all based on nearest neighbors of the hidden-state in the reference catalog $\mathcal{C}$ weighted thanks to a kernel function. Among the different kernels, Chau et al. [2021] propose to use a tricube kernel which has a compact support and is smooth at its boundary. Throughout this chapter, as suggested by Lguensat et al. [2017], a radial basis function (also known as Gaussian kernel, squared exponential kernel, or exponentiated quadratic) is considered and defined as:

$$g(\mathbf{u}, \mathbf{v}) = \exp\left(-\lambda \left|\left|\mathbf{u} - \mathbf{v}\right|\right|^2\right) , \tag{7.4}$$

where $(\mathbf{u}, \mathbf{v})$ are two distinct variables in the hidden-state space, λ is a scale parameter, and $||\cdot||$ is the Euclidean distance or any other relevant distance function for our application.

Let us denote by $\{\mathbf{a}_n\}_{n\in\mathcal{I}}$ the K-nearest neighbors (also known as analog situations) of a given hidden-state at iteration $k-1$, where $\mathcal{I} = \{i_1, \dots, i_K\}$ contains the K indices of these situations. From the reference catalog $\mathcal{C}$, one can retrieve the corresponding successors $\{\mathbf{s}_n\}_{n\in\mathcal{I}}$. Then for every pair of analog and successor $(\mathbf{a}_n, \mathbf{s}_n)_{n\in\mathcal{I}}$, a normalized kernel weight $(\omega_n)_{n\in\mathcal{I}}$ can be assigned:

$$\forall n \in \mathcal{I},\ \omega_n = \frac{g(\mathbf{x}_{(k-1)}, \mathbf{a}_n)}{\sum_{j=1}^{K} g(\mathbf{x}_{(k-1)}, \mathbf{a}_{i_j})} .$$

This term provides more importance to pairs that are best suited according to the kernel function for the estimation of the hidden-state $\mathbf{x}_{(k)}$ in the K-nearest neighbor ones obtained from the catalog. Nevertheless, the parametrization of this weight is highly dependent of the kernel function. Moreover in the context of Gaussian kernel as defined in

Equation (7.4), the normalized kernel weight involves the choice of the number of nearest neighbors K and the scale parameter λ. Two common strategies in the statistic field are used for the K-nearest neighbors estimation: either a distance threshold in order to consider the nearest neighbors which respect it, or an arbitrary number of analogs [Peterson, 2009]. In our work, we consider the last strategy for simplicity. As proposed by Lguensat et al. [2017], the scale parameter can be fixed following the adaptive rule defined as:

$$\lambda = \frac{1}{\mathrm{md}(\mathbf{x}_{(k-1)})},$$

where $\mathrm{md}(\mathbf{x}_{(k-1)})$ is the median distance between the hidden-state at iteration $k-1$ and its K nearest neighbors. Nevertheless, a more sophisticated procedure, based on a cross-validation procedure, can be employed to optimize the choice of these hyper-parameters.

Three analog forecasting operators $\mathcal{A}$ have been defined in Lguensat et al. [2017]. Firstly, the locally-constant analog operator which consists in forecasting the hidden-state by only using the successors. Let us denote by $\mathbf{x}^f_{(k)}$ the forecast of the state at iteration k. The idea of the locally-constant operator is to sample this forecasted hidden-state from a Gaussian distribution defined as :

$$\mathbf{x}^f_{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{LC}}, \Sigma_{\mathrm{LC}}),$$

where the mean forecast $\boldsymbol{\mu}_{\mathrm{LC}} = \sum_{j=1}^{K} \omega_{i_j} \mathbf{s}_{i_j}$ is the weighted mean of the K successors, and $\Sigma_{\mathrm{LC}} = \mathrm{cov}_\omega((\mathbf{s}_n)_{n\in\mathcal{I}})$ is the weighted empirical covariance of the successors of the K-nearest neighbors.

The second proposed analog operator is called the locally-incremental which considers the analogs and the successors of the state $\mathbf{x}_{(k-1)}$ to obtain $\mathbf{x}^f_{(k)}$. In the same way as for the locally-constant analog operator, the principle is to sample the forecasted state from a Gaussian distribution. Nevertheless, instead of only considering a weighted mean based on the K-nearest neighbors, the procedure uses a weighted mean of the differences between these K analogs and their respective successors plus the value of the current hidden-state. The derived Gaussian distribution is defined as:

$$\mathbf{x}^f_{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{LI}}, \Sigma_{\mathrm{LI}}),$$

where the mean forecast is $\boldsymbol{\mu}_{\mathrm{LI}} = \mathbf{x}_{(k-1)} + \sum_{j=1}^{K} \omega_{i_j}(\mathbf{s}_{i_j} - \mathbf{a}_{i_j})$, and $\Sigma_{\mathrm{LI}} = \mathrm{cov}_\omega\left((\mathbf{x}_{(k-1)} + (\mathbf{s}_n - \mathbf{a}_n))_{n\in\mathcal{I}}\right)$ is the weighted empirical covariance of the increments, i.e., differences between analogs and successors.

The last operator, developed by Lguensat et al. [2017], is named as the locally-linear forecasting operator. It consists in performing a weighted least square linear regression between the K-nearest neighbors and their corresponding successors in the catalog $\mathcal{C}$. The multivariate linear regression provides slope $\boldsymbol{\alpha}$, intercept $\boldsymbol{\beta}$, and residuals defined as $\forall j \in [1, \ldots, K]$, $\boldsymbol{\xi}_j = \mathbf{s}_{i_j} - \left(\boldsymbol{\alpha}\mathbf{a}_{i_j} + \boldsymbol{\beta}\right)$. The Gaussian sampling resorts to:

$$\mathbf{x}^f_{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{LL}}, \Sigma_{\mathrm{LL}}),$$

where the mean forecast is $\boldsymbol{\mu}_{\mathrm{LL}} = \boldsymbol{\alpha}\mathbf{x}_{(k-1)} + \boldsymbol{\beta}$, and $\Sigma_{\mathrm{LL}} = \mathrm{cov}_\omega((\boldsymbol{\xi}_j)_{j\in[1,\ldots,K]})$ is the weighted empirical covariance of the residuals.

The complexity of the application and the available computational resources are the two main constraints that will drive the choice of one forecasting operator over the others.

For example in situations facing some rare events, the locally-constant gives poor results due to the fact that the forecasting estimate is held in the range of K-nearest neighbors. In that context, the locally-incremental and the locally-linear forecasting operators are much more efficient. A graphical representation of the locally-constant, locally-incremental, and locally-linear analog forecasting operators for a 2-dimensional hidden-state is given in Figure 7.1. In this example, the underlying dynamics model has a simple polynomial form and the analogs are obtained by using a normal distribution sampling centered on the real value of the hidden-state at iteration $k-1$.

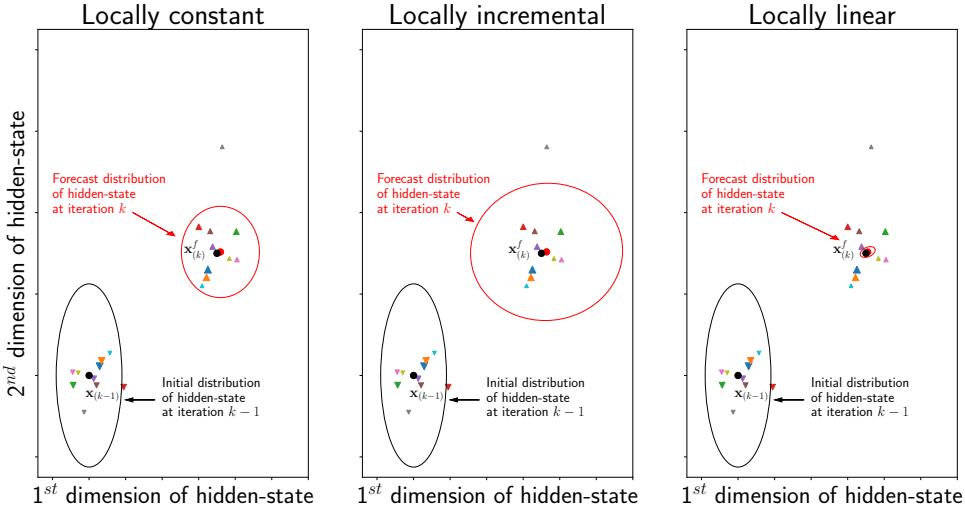


Figure 7.1 – Analog forecasting operator strategies. The real values of the hidden-state $\mathbf{x}_{(k-1)}$ and its forecast $\mathbf{x}_{(k)}$ are represented by full circles. Analogs are displayed in colored down-pointing triangles and successors in up-pointing triangles with their equivalent colors. The size of each triangle is proportional to the normalized kernel weight. The ellipsoids in black and red represent respectively the 95 % confidence intervals of the hidden state distribution before and after the analog forecasting strategy.

Hereafter, we propose to describe the data assimilation framework coupled with the analog forecasting method firstly proposed by Tandeo et al. [2014] and further detailed in [Lguensat et al., 2017]. Data assimilation methods allow us to combine all the sources of information obtained from a physical model and observations. In particular, sequential data assimilation techniques, also known as filtering approaches, which consist in estimating the filtering posterior distribution of the current hidden-state knowing past and present observations $p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k)})$, see Chapter 5.

As highlighted in the mentioned chapter, different methods are available in order to compute the filtering distribution of interest. In the context of linear Gaussian state-space models, Kalman filter methods can be considered to provide the exact filtering methods [Kalman, 1960, Brown, 1986, Harvey, 1990, Haykin, 2004, Wells, 2013]. Nevertheless in real applications, this nonlinear assumption is often unrealistic and more sophisticated Kalman-based approaches have to be used [Julier and Uhlmann, 1997, Evensen, 2009]. In particular, the ensemble Kalman filter (EnKF) which is a Monte Carlo variant relying on an ensemble of members to represent the statistics. This sequential Monte Carlo filter, introduced by Evensen [1994], is widely used in data assimilation application to take into account the nonlinearities in the state-space formulation and to handle the

high dimensional problems [Houtekamer and Mitchell, 2001, Snyder and Zhang, 2003, Aanonsen et al., 2009]. The EnKF principle is to sequentially update the ensemble of members by means of a correction term relying on the Kalman gain which allows to blend the model responses and the observations at a given iteration, see Evensen [2003] or Section 5.2. Due to the fact that this approach is based on an ensemble, it is hence inherently well-adapted to parallelization which is a crucial advantage with the current high-performance computing architectures for the inference of time-consuming numerical models [Houtekamer et al., 2014].

Thus, we present the formulation of a non-parametric EnKF method, also known as analog EnKF (AnEnKF), see [Tandeo et al., 2014, Lguensat et al., 2017]. The procedure is similar to the stochastic ensemble Kalman recursion [Evensen, 2009]. Nevertheless, the main difference of the AnEnKF occurs for the forecast step where the non-parametric data-driven sampling, i.e., the analog forecasting operator, is used instead of the dynamic model $\mathcal{M}$ in Equation (7.2). The Analog ensemble Kalman filter consists at each iteration to apply one of the three analog forecast sampling strategies to each analysis member of the ensemble to generate a forecast term. Then, the equations used in the procedure are equivalent to the EnKF strategy. At each iteration during the analysis step, each forecast member of the ensemble is corrected by computing $\mathbf{x}_{(k)}^{a(i)} = \mathbf{x}_{(k)}^{f(i)} + \mathbf{K}_{(k)} \left(\mathbf{y}_{(k)}^{(i)} - \mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}, \mathbf{u}_{(k)}) \right)$ where $\mathbf{K}_{(k)} = \mathbf{P}_{(k)}^{f} \ \mathbf{H}_{(k)}^{T} \left(\mathbf{R}_{(k)} + \mathbf{H}_{(k)} \mathbf{P}_{(k)}^{f} \mathbf{H}_{(k)}^{T} \right)^{-1}$ is named as the Kalman Gain. Due to the nonlinearity of the model $\mathcal{H}_{(k)}$, the terms $\mathbf{P}_{(k)}^{f} \ \mathbf{H}_{(k)}^{T}$ and $\mathbf{H}_{(k)} \mathbf{P}_{(k)}^{f} \mathbf{H}_{(k)}^{T}$ are respectively empirically estimated based on the ensemble members. The ensemble Kalman filter coupled with the analog forecasting strategy is detailed in Algorithm 6.

Algorithm 6: Ensemble Kalman Filter with analog forecast methodology, so-called AnEnKF.

Data:
number of members in the ensemble N_{ens};
catalog $\mathcal{C}$ and number of nearest neighbors K;
prior guess of the parameter vector $\mathbf{x}_b$ and prior parameter covariance matrix $\mathbf{P}_b$;
some forcing variables $\{\mathbf{u}_{(k)}\}_{k=1,\dots,T}$ and measurements $\{\mathbf{y}_{(k)}\}_{k=1,\dots,T}$;
error covariance matrix $\{\mathbf{R}_{(k)}\}_{k=1,\dots,T}$ and artificial error covariance matrix $\{\mathbf{Q}_{(k)}\}_{k=0,\dots,T}$.

Initialisation step:

for $i = 1$ *to* N_{ens} **do**

$$\mathbf{x}_{(0)}^{a(i)} = \mathbf{x}_b + \boldsymbol{\epsilon}^b \text{ with, } \boldsymbol{\epsilon}^b \sim \mathcal{N}(0, \mathbf{P}_b)$$

for $k = 1$ *to* T **do**

Forecast step:

for $i = 1$ *to* N_{ens} **do**

Locally-constant forecasting analog operator:

$$\mathbf{x}_{(k)}^{f(i)} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{LC}}, \Sigma_{\mathrm{LC}}) \text{ with, } \boldsymbol{\mu}_{\mathrm{LC}} = \sum_{j=1}^{K} \omega_{i_j} \mathbf{s}_{i_j}$$
$$\text{and } \Sigma_{\mathrm{LC}} = \mathrm{cov}_{\omega}((\mathbf{s}_n)_{n\in\mathcal{I}})$$

or **Locally-incremental forecasting analog operator:**

$$\mathbf{x}_{(k)}^{f(i)} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{LI}}, \Sigma_{\mathrm{LI}}) \text{ with, } \boldsymbol{\mu}_{\mathrm{LI}} = \mathbf{x}_{(k-1)}^{a(i)} + \sum_{j=1}^{K} \omega_{i_j} (\mathbf{s}_{i_j} - \mathbf{a}_{i_j})$$
$$\text{and } \Sigma_{\mathrm{LI}} = \mathrm{cov}_{\omega}\left((\mathbf{x}_{(k-1)}^{a(i)} + (\mathbf{s}_n - \mathbf{a}_n))_{n\in\mathcal{I}}\right)$$

or **Locally-linear analog operator:**

$$\mathbf{x}_{(k)}^{f(i)} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{LL}}, \Sigma_{\mathrm{LL}}) \text{ with, } \boldsymbol{\mu}_{\mathrm{LL}} = \boldsymbol{\alpha}\mathbf{x}_{(k-1)} + \boldsymbol{\beta}$$
$$\text{and } \Sigma_{\mathrm{LL}} = \mathrm{cov}_{\omega}((\boldsymbol{\xi}_j)_{j\in[1,\dots,K]})$$

where $(\mathbf{a}_n, \mathbf{s}_n)_{n\in\mathcal{I}}$ (with $\mathcal{I} = \{i_1, \dots, i_K\}$) are the K-pairs of analog and successor for the i-th analysis member of the ensemble and cov_{ω} is the weighted covariance.

Update step:

$$\mathbf{P}_{(k)}^{f}\ \mathbf{H}_{(k)}^{T} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathbf{x}_{(k)}^{f(i)} - \bar{\mathbf{x}}_{(k)}^{f}\right) \left(\mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}, \mathbf{u}_{(k)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}_{(k)}^{f}, \mathbf{u}_{(k)})\right)^{T}$$

$$\mathbf{H}_{(k)}\mathbf{P}_{(k)}^{f}\mathbf{H}_{(k)}^{T} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}, \mathbf{u}_{(k)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}_{(k)}^{f}, \mathbf{u}_{(k)})\right) \left(\mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}_{(k)}^{f})\right)^{T}$$

$$\mathbf{K}_{(k)} = \mathbf{P}_{(k)}^{f}\ \mathbf{H}_{(k)}^{T} \left(\mathbf{R}_{(k)} + \mathbf{H}_{(k)}\mathbf{P}_{(k)}^{f}\mathbf{H}_{(k)}^{T}\right)^{-1}$$

for $i = 1$ *to* N_{ens} **do**

$$\mathbf{y}_{(k)}^{(i)} = \mathbf{y}_{(k)} + \boldsymbol{\epsilon}_{(k)}^{o(i)} \text{ with, } \boldsymbol{\epsilon}_{(k)}^{o(i)} \sim \mathcal{N}(0, \mathbf{R}_{(k)})$$
$$\mathbf{x}_{(k)}^{a(i)} = \mathbf{x}_{(k)}^{f(i)} + \mathbf{K}_{(k)} \left(\mathbf{y}_{(k)}^{(i)} - \mathcal{H}_{(k)}(\mathbf{x}_{(k)}^{f(i)}, \mathbf{u}_{(k)})\right)$$

7.3 Numerical results

In this section, the numerical results of the proposed methodology to quantify and reduce the uncertainties based on global sensitivity analysis and a data-driven data assimilation approach are presented in the context of an industrial operating wind turbine. The two categories of parameters investigated in this application case are the wind turbine model properties and the wind-inflow conditions. In the sensitivity analysis of the fatigue loads of the wind turbine, we assume that the 10-minute mean and standard deviation obtained from the SCADA are respectively equal to 10 m/s and 1.4 m/s.

7.3.1 Case description

For the purpose of this work, the considered model is a numerical representation of a reference 2MW onshore horizontal-axis wind turbine based on the open-source aero-servo-elastic software FAST developed by the National Renewable Energy Laboratory (NREL) [Jonkman et al., 2005]. This numerical code employs a combined modal and multibody dynamics formulation which allows to consider a limited number of degree of freedom for the structure. Moreover, the aerodynamic model relies on the blade-element momentum theory coupled with some corrections, e.g., dynamic stall. The generation of the synthetic turbulent wind field solicitation uses a Kaimal turbulence model with an exponential spatial coherence method thanks to the TurbSim software [Jonkman, 2009]. Some specifications of the turbine are presented in Table 7.4.

Table 7.4 – Reference wind turbine specifications

Quantity	Value
Number of blades	3
Rated power	2.0 MW
Rotor speed range	8.5 – 17.1 rpm ($\pm$16 %)
Rated wind speed	13 m/s
Cut-in wind speed	3.0 m/s
Cut-out wind speed	25 m/s
Rotor radius	41 m
Hub height	80 m

The in situ data used to assess the performances of our procedure are based on a specific measurement campaign of eight months from the national project SMARTEOLE [2]. For that purpose, the wind turbine has a supervisory control and data acquisition system (SCADA) gathering 10-minute statistics about the external conditions at the nacelle hub, e.g., wind speed or direction, and also information on the turbine operation, e.g., generator speed, generated power. Alongside, a nacelle mounted Light Detection And Ranging (LIDAR) system is placed on top of the wind turbine nacelle in order to measure the upstream wind flow conditions. A graphical representation of the monitoring system configuration is proposed in Figure 7.2. In the study, we suppose that the wind speed at hub height reconstructed from the LIDAR system is the free wind to be applied on

2. The author acknolewdges SMARTEOLE project partners for the use of experimental data from national project SMARTEOLE (ANR-14-CE05-0034) measurement campaigns.

the servo-aero-elastic model through the synthetic turbulence wind field. Lastly, bi-axial measuring devices are located at mid and top tower height position. From these sensors, we can record four functional acceleration time series. Then, the power spectral density (PSD) of each measured acceleration time series is computed using Welch's method.

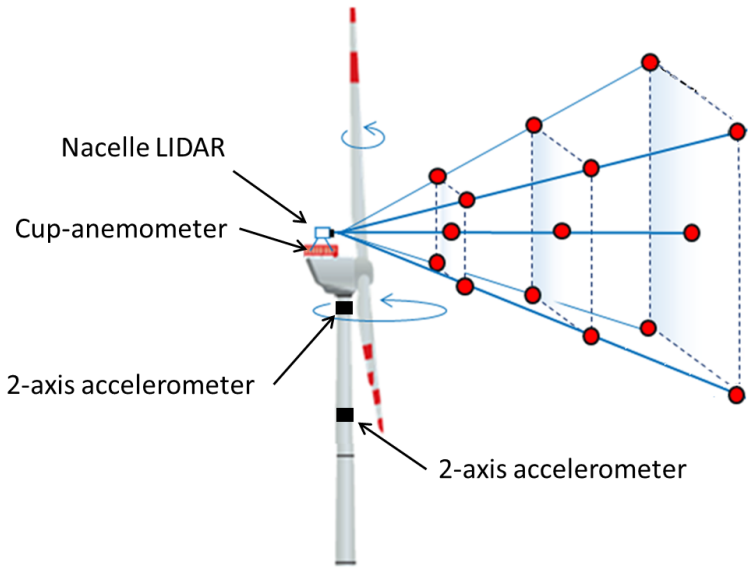


Figure 7.2 – Monitoring system configuration for the reference wind turbine.

7.3.2 Global sensitivity analysis on fatigue loads

To quantify the importance of each input parameter on the variability of the fatigue loads obtained from the aero-servo-elastic numerical model, a global sensitivity analysis (GSA) based on Sobol' index estimation [Sobol', 1993, Saltelli et al., 2000] has been investigated. We focus our interest on total Sobol' sensitivity indices. The total Sobol' index associated to each input parameter represents the amount of the quantity of interest variance due to this parameter alone or in interaction with any other subset of parameters. It allows to quantify the part of variation in the damage equivalent load that could be reduced if the parameter was to be fixed in a single value. To alleviate the computational cost in the sensitivity index estimation, heteroscedastic Gaussian process (GP) models are built independently for each DEL. Fitting such surrogate model to the load behavior of a wind turbine requires a design of experiments covering the range of variation in all parameters. In that context, we rely on a Latin Hypercube Sampling (LHS) of size 996 with a geometrical criterion maximizing the minimum distance between the design points. To testify the accuracy of the fitted surrogate model for each output of interest, an augmented LHS of size 200 has been generated. Then, ten different turbulent inflow realizations are generated using the Kaimal spectrum with an exponential spatial coherence model for each point, for which the empirical mean and standard deviation of the fatigue loads are estimated. The heteroscedastic property of the GP allows to capture the global fatigue behavior of the turbine but also to estimate the inherent variability due to different turbulent wind field realizations. This study leads to a total number of 11,960 aero-servo-elastic numerical model evaluations.

Eight different model quantities of interest are considered for describing the fatigue behavior of the wind turbine, see Table 7.5. For each output, the total effect Sobol indices are estimated using the corresponding heteroscedastic Gaussian process metamodel based on the estimator proposed by Jansen [1999] and implemented in the function sobolGP of the R package sensitivity [Iooss et al., 2019]. The estimation approach relies on the complete conditional predictive distribution of the metamodel which allows to evaluate the uncertainty in the estimation due to the Monte Carlo procedure or the surrogate approximation, see Algorithm in [Hirvoas et al.].

Table 7.5 – Wind turbine model fatigue load outputs with their corresponding negative inverse slope coefficient m.

Quantity of interest	m
DEL blade root in-plane bending moment	10
DEL blade root out-of-plane bending moment	10
DEL tower bottom fore-aft bending moment	3
DEL tower bottom side-to-side bending moment	3
DEL tower top side-to-side bending moment	3
DEL tower top fore-aft force	3
DEL shaft torsional moment	3

For the estimation procedure, two distinct LHSs with a maximin criterion of size 9,946 have been generated. The uncertainty related to the kriging approximation is quantified by using 100 samples. Moreover, the uncertainty due to Monte Carlo integration was estimated with a bootstrap procedure with a sample size of 100, see Annex A for futher details in bootstrapping strategy. The estimated total Sobol' indices, denoted by S^T, for the considered quantities of interest with their corresponding 95% confidence intervals are presented in Figure 7.3. Most of the outputs have a large total Sobol' index for the errors relative to the wind speed $\Delta\overline{u}$ and $\Delta\sigma_u$. These input parameters have an important impact on the variability of fatigue loads obtained from our aero-servo-elastic numerical model. The vertical wind shear coefficient α has also a clear impact in particular for the torsional moment of the shaft and the out-of-plane bending moment of the blade. The noticeable effect of the wind shear for rotating components can be explained by the fact that they will face cyclic changes in wind velocity if wind shear is considered. Six other parameters describing the wind inflow conditions or the wind turbine model properties have total Sobol' indices higher to the arbitrary threshold (set to $5e-02$) and can be considered as influential. The arbitrary threshold is used to discriminate efficiently sensitive and insensitive input parameters. For simplicity, these parameters are underlined in Figure 7.3. In particular, we can notice that model property parameters related to tower thickness, lineic mass and mass imbalance related to the blades (e, α_{mass}, and η_B) have a non-negligible influence on fatigue load variance of the considered wind turbine components. The remaining parameters can be fixed to any specific value in their range of variability without affecting the considered fatigue loads.

After assessing the sensitivity analysis of the fatigue load of some critical components of the wind turbine structure, one major challenge is to reduce the uncertainties affecting the most influential input parameters.

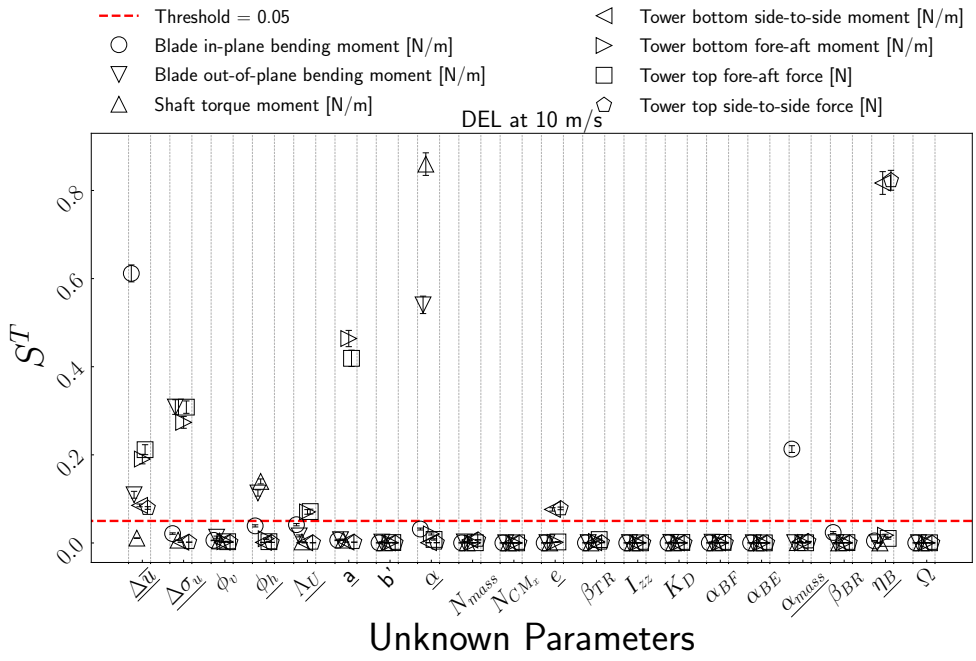


Figure 7.3 – Estimation of total Sobol' indices (y-axis) with their 95% confidence interval corresponding to each of the 20 parameters (x-axis) for the different fatigue loads. The dashed line corresponds to a threshold arbitrarily chosen to 5e-2. Confidence intervals (CI) are obtained by taking into account the uncertainties due to both the metamodel and the Monte Carlo estimation. The number of samples for the conditional Gaussian process, in order to quantify the uncertainty of the kriging approximation, was set to 100. The uncertainty due to Monte Carlo integration was computed with a bootstrap procedure with a sample size of 100.

7.3.3 Identifiability study

A major issue for parameter estimation problem is the identifiability. In this context, Dobre et al. [2012] highlight that nullity of total sensitivity index for a specific input parameter implies its non-identifiability from the measured output. Consequently, we perform a GSA on the measured outputs in order to determine which parameters cannot be inferred with the current sensors on the wind turbine. In our industrial application, six measured outputs are considered, see Table 7.6.

For the acceleration outputs, we are mainly interested in their response in the frequency-domain by using the power spectral density (PSD). When performing GSA, discretized PSD series involve a substantial dimensionality and a high degree of redundancy. To overcome this issue, the different discretized PSD outputs have been reduced using a Principal Component Analysis (PCA) [Wold et al., 1987]. This dimensionality reduction approach allows the functional output expansion in a new reduced space spanned by the most significant directions in terms of variance. Then, a method based on PCA and GSA with a GP model is used to compute an aggregated Sobol' index for each input parameter of the model [Lamboni et al., 2011]. The proposed index synthesizes the influence of the parameter on the whole discretized functional output. Table 7.7 summarizes the estimated total aggregated Sobol' indices. In this sensitivity analysis, the input parameters having total Sobol' index values under a threshold set at $1e-02$ are considered as non-identifiable from the measured output.

Table 7.6 – Observations performed in our reference wind turbine.

Observation	Unit
10-minute mean power production	[kW]
10-minute mean rotor speed	[rpm]
Tower middle fore-aft acceleration's PSD	[dB]
Tower middle side-to-side acceleration's PSD	[dB]
Tower top fore-aft acceleration's PSD	[dB]
Tower top side to side acceleration's PSD	[dB]

Table 7.7 – Total Sobol' and aggregated total Sobol' indices for each output used during the recursive inference procedure. Estimated total Sobol' indices higher than the arbitrary threshold are underlined.

Measured outputs	$\Delta\overline{u}$ $[m/s]$	$\Delta\sigma_u$ $[m/s]$	ϕ_h $[°]$	Λ_u $[m]$	a $[-]$	α $[-]$	e $[\%]$	α_{mass} $[\%]$	η_B $[\%]$
10-minute mean power production	9.81e-01	4.29e-04	1.71e-02	1.30e-04	3.70e-04	1.50e-02	3.84e-05	3.83e-04	5.23e-05
10-minute mean rotor speed	9.75e-01	3.30e-03	1.87e-02	9.43e-04	1.61e-03	1.62e-02	1.03e-04	7.56e-04	7.34e-05
Tower middle fore-aft acceleration's PSD	1.44e-01	2.49e-01	1.00e-02	1.77e-01	3.70e-01	1.33e-02	4.58e-02	5.82e-03	3.48e-03
Tower middle side-to-side acceleration's PSD	2.04e-01	2.51e-01	1.09e-02	1.92e-01	3.00e-01	1.33e-02	4.49e-02	4.86e-03	3.42e-03
Tower top fore-aft acceleration's PSD	3.12e-01	2.16e-01	1.87e-02	1.75e-01	2.69e-01	9.59e-03	3.36e-02	8.49e-03	7.01e-03
Tower top side to side acceleration's PSD	2.84e-01	1.87e-01	1.18e-02	1.76e-01	2.50e-01	1.21e-02	8.33e-02	5.52e-03	2.38e-02

According to the GSA, the coefficient related to the distributed blade mass α_{mass} is not identifiable with the current observations. Consequently, the model parameter properties remaining for the inference procedure are the tower thickness coefficient e, and the mass imbalance factor η_B. Moreover, all the influent parameters related to the wind field remain candidates for the recursive inference strategy.

7.3.4 Recursive inference strategy based on AnEnKF approach

With the current monitoring configuration, data availability or quality does not allow a proper extraction of the mean flow angle ϕ_h, the longitudinal turbulence length scale Λ_u, and the decrement parameter of the coherence model a. Consequently, only the six remaining parameters having an influential effect on the fatigue behavior of the structure and potentially identifiable are considered during the recursive inference procedure. These input parameters and their corresponding prior Gaussian distributions are detailed in Table 7.8. Their corresponding reference variable in the augmented state vector is also

specified.

Table 7.8 – A-priori Gaussian distribution $\mathcal{G}$ for each of the considered input parameters.

Input parameter	Variable	Distribution	Initial prior	State
Tower thickness	e	$\mathcal{G}$	$\mu = -10.00 \quad \sigma = 7.00$	$\mathbf{x}^1$
Blade mass imbalance	η_B	$\mathcal{G}$	$\mu = 4.00 \quad \sigma = 8.33e-01$	
Error mean of the wind speed at hub height	$\Delta\overline{u}$	$\mathcal{G}$	$\mu = 0.00 \quad \sigma = 9.11e-01$	$\mathbf{x}^2$
Error standard deviation of the wind speed at hub height	$\Delta\sigma_u$	$\mathcal{G}$	$\mu = 0.00 \quad \sigma = 9.70e-02$	
Vertical wind shear exponent	α	$\mathcal{G}$	$\mu = 1.30e-01 \quad \sigma = 2.90e-01$	

For assessing the performance of the AnEnKF for our recursive inference procedure, we rely on pseudo-experimental numerical tests. They consist in performing forward aero-servo-elastic simulations considering known values of the input parameters, and then adding a Gaussian noise of known variance to the simulated measurements. In our study, the simulated data are perturbed by considering a covariance matrix such as the obtained standard deviation is equivalent to a 10% signal-to-noise ratio. The pseudo-simulated responses of the wind turbine structure are generated using the wind inflow conditions obtained from the nacelle mounted LIDAR for a specific day and the mean values of the model properties described in Table 7.3. The noisy pseudo-experimental outputs used to recursively update the wind turbine model are 10-minute mean power production and rotor speed, and the PSD of the acceleration time series obtained for side to side and fore-aft at the two different tower positions. Our recursive inference problem using a filtering-based estimation procedure can be considered as a state estimation problem for the following augmented system:

$$\forall k \in \mathbb{N}^*, \begin{cases} \mathbf{x}_{(k)} = \begin{bmatrix} \mathbf{x}^1_{(k)} \\ \mathbf{x}^2_{(k)} \end{bmatrix} = \begin{pmatrix} \mathbf{x}^1_{(k-1)} \\ \mathcal{A}_{(k-1,k)}(\mathbf{x}^2_{(k-1)}) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}^m_{(k)} \\ 0 \end{pmatrix} \\ \mathbf{y}_{(k)} = \mathcal{H}_{(k)}(\mathbf{x}_{(k)}, \mathbf{u}_{(k)}) + \boldsymbol{\epsilon}^o_{(k)} \end{cases} .$$

where $\mathbf{x}^1_{(k-1)}$ and $\mathbf{x}^2_{(k-1)}$ are respectively the uncertain parameters for the model properties and the wind inflow conditions at iteration $k-1$ as described in Table 7.8, $\mathcal{A}_{(k-1,k)}$ is the analog forecasting operator as detailed in Section 7.2, $\mathbf{u}_{(k)}$ is the forcing vector corresponding to the 10-minute mean and standard deviation wind speed obtained from the SCADA system, and $\mathcal{H}_{(k)}$ is the combination of the aero-servo-elastic model FAST and the turbulent wind field generation software Turbsim.

For the initialization of the EnKF approach, independent Gaussian distributions are assumed to be the initial prior for each of the input parameters, see Table 7.8. The initial error covariance matrix of the input parameters, denoted by $\boldsymbol{P}_b$, is thus assumed to be diagonal. To create the catalog, we rely on the measurements obtained from both the SCADA system and the LIDAR installed on the onshore wind turbine. A data pre-treatment has been performed in order to find any corrupted observations. The obtained database consists in both 4,735 analog situations to be compared to the current parameters related to the wind inflow and their corresponding successors at a 10-minute interval.

Figure 7.4 shows the results of the identification of the considered input parameters by applying the AnEnKF approach with the locally-linear forecasting operator using $N = 500$ members and $K = 50$ nearest-neighbors. It can be noticed that the augmented state vector is well reconstructed by using this non-parametric data assimilation procedure which allows to emulate the dynamical model from a dataset. Indeed, the mean of the empirical distribution obtained from the members of the ensemble is close to the true hidden-state for every parameter. A major advantage of the procedure is the confidence intervals obtained at each inference iteration allowing us to give information about the difficulty to retrieve the value of the input parameters from the measured outputs.

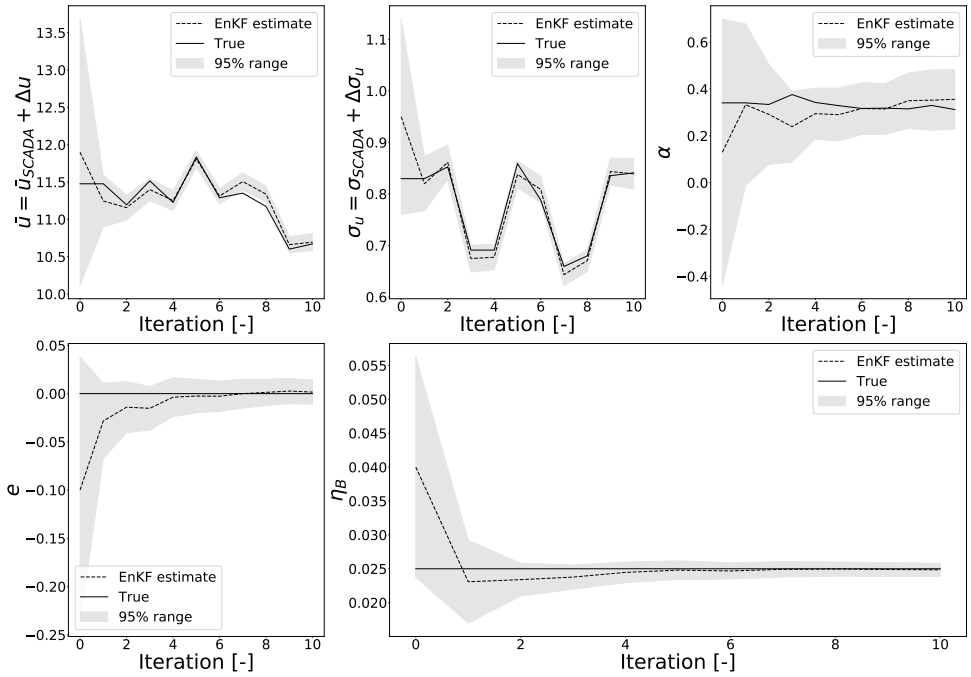


Figure 7.4 – Iteration evolution of the posteriori estimates of the input parameters. Results obtained by running the AnEnKF procedure with $N = 500$ members of the ensemble used for the estimation and considering pseudo-experimental numerical observations.

Conclusion

In the present work, we extend a procedure to quantify and reduce the uncertainties affecting the fatigue load estimation of a wind turbine numerical model. The fatigue loads encountered by a wind turbine structure are function of the parameters describing the turbulent wind field, the structural properties, and the control system. The study aims at taking into account these parameters used as input to aero-servo-elastic fatigue load simulations of an operating wind turbine. The procedure relies on a global sensitivity analysis and a recursive Bayesian inference method. A major challenge during the recursive inference procedure is the dynamic behavior of the inflow-related parameters. Unfortunately, the underlying dynamic behavior of these parameters is not explicitly known. To overcome this issue, a combination of the implicit analog forecasting of the dynamics with the ensemble Kalman filtering scheme is investigated.

Finally, we demonstrate the applicability and performance of the procedure using a numerical representation of a reference wind turbine. The study leads to the following main conclusions. The global sensitivity analysis based on heteroscedastic Gaussian processes for the estimation of Sobol' indices shows that parameters related both to the wind and the structure have an influence on the fatigue loads of a wind turbine structure. The presented metamodeling approach is an efficient way to capture the inherent stochasticity of aero-servo-elastic simulations due to the turbulent inflow realization leading to variations in the quantities of interest. After determining the most influential parameters in terms of fatigue loads variability, an identifiability study based on a global sensitivity analysis is performed to assess if these parameters can be inferred from the current sensors. The sensitivity analysis is based on the estimation of the so-called aggregated Sobol' indices involving a principal component analysis in order to take into account the functional behavior of the measured outputs. Finally, the ensemble Kalman filtering method coupled with the analog forecasting strategy used in this study is very suitable for carrying the recursive inference of parameters related to the wind field solicitation and the wind turbine numerical description.

Further research should focus on the quality of the catalog used for the analog forecasting strategy. Additionally, other types of kernels in the forecasting operator have to be studied. Lastly, the hyperparameters used in the K-nearest neighbors method and the chosen kernel function could be optimized for each member of the ensemble Kalman filtering procedure by using a cross-validation approach. From an industrial perspective, the proposed AnEnKF methodology has to be performed using measured acceleration time-series obtained from the sensor devices of the onshore wind turbine.

Conclusion and perspectives

Ainsi s'écoule toute la vie. On cherche le repos en combattant quelques obstacles ; et si on les a surmontés, le repos devient insupportable.

Blaise Pascal

The main focus of this thesis was the development of a data assimilation method for the calibration and continuous update of a wind turbine numerical model based on in situ observations. We proposed a complete framework for quantifying and recursively reducing the uncertainties affecting the input parameters of a numerical model. For that purpose, this dissertation explored various concepts related to the field of uncertainty quantification. This scientific discipline focuses on the characterization and the reduction of uncertainties in numerical applications. Chapter 1 provided a literature review of the methodologies widely used by the uncertainty quantification community, involving different topics such as sensitivity analysis, surrogate modeling, and parameter inference. Our research work was mainly guided by the industrial application in which this doctoral project fits. As mentioned in Chapter 2, the numerical modeling of wind turbines requires the involvement of different physics in order to properly represent the different phenomena of interest. The application of uncertainty quantification to such numerical applications is challenging. Indeed, the aero-servo-elastic simulations currently in use are stochastic and time-consuming. That led us to propose a complete procedure for quantifying and reducing the uncertainties while respecting the different constraints we were confronted with. The relevance and efficiency of the different statistical approaches proposed in this work were illustrated on two numerical cases. The development of wind turbines models required for these application cases was part of this thesis. The general frame of our research work was mainly divided into three parts.

The first part was devoted to the quantification of the ubiquitous uncertainties affecting the parameters of the aero-servo-elastic numerical model as well as the external sollicitations. In Chapter 3, we presented a global sensitivity analysis based on Sobol' indices to rank the input parameters according to their impact on the output variability. This probabilistic approach consists in modeling the uncertain parameters by independent random variables characterized by their probability distribution. In this study, the estimation of these sensitivities was performed with a Monte Carlo based procedure. Such

technique avoids any regularity assumption on the model but requires a lot of calls to it. Nevertheless, in our context of stochastic and time-consuming simulations, this statistical strategy cannot be used directly. We suggested in Chapter 4 to alleviate the computational burden of the Monte Carlo estimation by surrogating the mean of the output, obtained from replications of stochastic aero-servo-elastic simulation, by a Gaussian process with heteroscedastic noise. This metamodeling strategy also made possible to take into account the inherent variability of our stochastic simulator. Global sensitivity analysis is important for the determination of the most important input parameters that have to be properly calibrated and the less influential ones that can be fixed to reference values.

Then, the second part was related to the reduction of the uncertainties affecting the influential parameters. Chapter 5 highlighted the use of data assimilation procedures for recursive model calibration. In particular, we investigated the ensemble Kalman filter for the inference of model parameters. This filtering scheme relies on measurements to retrieve the unknown input parameter distribution based on the Bayesian paradigm. In our application study, we proposed to use a latin hypercube sampling for the generation of the prior ensemble of members. Nevertheless, such inverse problems can only be achieved assuming that several conditions of well-posedness and identifiability are achieved. In that context, we made use of the relationship between non-identifiability of input parameters and total Sobol' indices. As highlighted by Dobre et al. [2012], if all the total Sobol' indices associated to a prescribed input parameter are "small" for the different measured outputs, it means that this parameter is non-identifiable.

The last part of our research work focused on two industrial applications of the proposed framework in the context of an onshore wind turbine numerical model. For the first application in Chapter 6, we developed a high-fidelity numerical representation of the turbine based on a full finite element analysis. This first study was focused on the parameters describing the model properties having a static or slow time-variant behavior. A global sensitivity analysis on the fatigue loads at some critical parts of the structure was performed in order to determine the most important inputs in terms of variability. The results we obtained showed that a subset of the model property parameters happens to influence these responses. Then, the question of parameter identifiability was investigated. Due to the functional nature of the observations, a dimension reduction preliminary step was performed thanks to a principal component analysis and then an aggregated Sobol' index for each model parameter was estimated. Finally, the proposed inference strategy based on the ensemble Kalman filter was able to recursively estimate the parameters related to the wind turbine properties from the synthetic measured data. Due to the encouraging results, in the second application presented in Chapter 7, we increased the complexity of our research work by extending the framework to the parameters related to the synthetic wind field. A major challenge during the recursive inference procedure is the dynamic aspect of these inflow-related parameters which is not explicitly known. For the reconstruction of this dynamics in the ensemble Kalman filter, we proposed to use a non-parametric data-driven approach relying on an analog forecasting strategy based on K-nearest neighbors.

The research work ends at this stage for the thesis but opens several extensions in terms of theoretical, practical, and application perspectives. Some of these promising ways of extensions are now discussed.

Concerning the theoretical perspectives, a further investigation of the relationship between global sensitivity analysis based on the functional analysis of the variance and

identifiability should be conducted [Dobre et al., 2012]. Moreover, the benefits of using a latin hypercube sampling or other space filling designs in the ensemble Kalman filter should be quantified more properly as it seems promising as empirical results.

In this thesis a global sensitivity analysis of the fatigue loads obtained from an aero-servo-elastic wind turbine numerical model has been performed to an under-rated wind speed configuration. Several practical extensions should be tested, some of them are listed hereafter. It should be relevant to consider other values of wind speed in order to take into account the blade pitch controller response in the global sensitivity analysis. Furthermore, the joint-metamodeling of the mean and dispersion of heteroscedastic data could be investigated for the global sensitivity analysis of stochastic computer codes [Marrel et al., 2012]. In this context, Murcia et al. [2018] proposed to take into account the inherent variability due to the different realizations of the turbulent wind field by fitting independent polynomial chaos expansions for the mean and standard deviation of each quantity of interest obtained from an aero-servo-elastic numerical model. These different strategies could be implemented and compared. Concerning the recursive inference procedure, we could study the use of dimension reduction techniques for the functional measurement considered in the ensemble Kalman filter. For the analog forecasting operator used in our application, we focused on a Gaussian kernel function. It would be pertinent to study the use of different covariance functions [Hofmann et al., 2008, Duvenaud, 2014]. Additionally, the hyper-parameters used in the K-nearest neighbor methods and the kernel design could be optimized for each member of the ensemble Kalman filtering procedure by using a cross-validation approach, although we could face numerical burdens. Moreover, recent years have been marked by a development of machine and deep learning techniques. A direction for future research work lies in the use of these innovative approaches to determine the hidden-state dynamics [Talmon et al., 2015, Krishnan et al., 2015, Fraccaro et al., 2016, Bocquet et al., 2020].

For application perspectives, the quality of the catalog used for the analog forecasting strategy has to be improved. Indeed in this study, we performed a coarse pretreatment of the database and only a limited part of the SMARTEOLE measurement campaign was considered. Last but not least, another application research path would be the use of the proposed recursive inference strategy with measured time-series obtained from the sensor devices of the operating onshore wind turbine.

Appendices

A

Bootstrap sampling

Bootstrap confidence intervals with percentile bias correction

The sampling error, due to the Monte Carlo evaluation of the variances in the Sobol' index estimators defined in Section 3.3, can be quantified by using confidence intervals.

One can approximate confidence intervals by employing a non parametric bootstrap method with the bias percentile correction. We first present the principle of the method introduced by Efron [1981], Efron and Tibshirani [1986]. Let us consider an estimator $\widehat{T}$ of an unknown quantity of interest T function of a random variable $X \in \mathcal{P}$. In order to obtain a point estimate of T, we can consider a s independent and identically distributed random sample $\{x_1, \cdots, x_s\}$ from $\mathcal{P}$, and then compute $\widehat{T}(x_1, \cdots, x_s)$. In non-parametric bootstrap strategy, the idea is to consider an integer $B > 0$ and repeatedly, for $b = 1, \cdots, B$, create a bootstrap sample $\{x_1[b], \cdots, x_s[b]\}$ by sampling with replacement from the sample $\{x_1, \cdots, x_s\}$ in order to obtain a replication of $\widehat{T}$ by estimating $\widehat{T}[b] = \widehat{T}(x_1[b], \cdots, x_n[b])$.

Let us denote by $\mathcal{R} = \{\widehat{T}[1], \cdots, \widehat{T}[B]\}$ the set of replications of $\widehat{T}$. This set can be used to estimate a bootstrap confidence interval for the quantity of interest. By considering the standard normal cumulative distribution function defined as:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t \cdot$$

A bias correction constant term z_0 can be estimated such as:

$$\widehat{z_0} = \Phi^{-1}\left(\frac{\#\{\widehat{T}[b] \in \mathcal{R} \text{ s.t. } \widehat{T}[b] \leq \widehat{T}\}}{B}\right) .$$

Then, we can express the corrected quantile estimate $\widehat{q}(\beta)$ for $\beta \in]0;1[$ as $\widehat{q}(\beta) = \Phi(2\widehat{z_0} + z_\beta)$, where z_β satisfies $\Phi(z_\beta) = \beta$. The bias corrected confidence interval of level $1-\alpha$ is obtained by estimating the $\widehat{q}(\alpha/2)$ and $\widehat{q}(1-\alpha/2)$ quantiles of $\mathcal{R}$. In order to justify this previous confidence interval, Efron [1981] highlights that an increasing transformation g, $z_0 \in \mathbb{R}$ and $\sigma > 0$ has to exist.

The bias corrected bootstrap confidence interval for the closed Sobol' index described in Equation (3.7) is detailed in Algorithm 7. The major advantage of bootstrapping

our sensitivity estimators is that we do not require supplementary model evaluations to estimate a confidence interval.

Algorithm 7: Bootstrapping procedure for confidence interval of the closed Sobol' index $\tilde{S}_{\mathbf{u}}$ adapted from [Janon et al., 2011].

1. For $\mathbf{u} \subseteq \{1, \ldots, p\}$, generate two designs $\mathbf{P}$ and $\mathbf{P}^{\mathbf{u}}$ from independent random vectors distributed according to the input parameter vector (see Section 3.3).
2. Create a third design $\mathbf{P}^{\mathbf{u}} = \{\mathbf{x}_i^{\mathbf{u}}\}_{i=1}^s$ with $\forall j \in [1, p] \begin{cases} x_j^{\mathbf{u}} = x_j & \text{if } j \in \mathbf{u} \\ x_j^{\mathbf{u}} = x_j' & \text{otherwise} \end{cases}$.
3. Compute $\forall i = 1, \cdots, s$, $y_i = \mathcal{M}(\mathbf{x}_i)$ and $y_i^{\mathbf{u}} = \mathcal{M}(\mathbf{x}_i^{\mathbf{u}})$
4. Estimate $\widehat{\tilde{s}}_{\mathbf{u}}$:
$$\widehat{\tilde{S}}_{\mathbf{u}} = \frac{\frac{1}{s}\sum_{i=1}^s y_i\, y_i^{\mathbf{u}} - (\frac{1}{s}\sum_{i=1}^s y_i)\,(\frac{1}{s}\sum_{i=1}^s y_i^{\mathbf{u}})}{\frac{1}{s}\sum_{i=1}^s y_i^2 - (\frac{1}{s}\sum_{i=1}^s y_i)^2}.$$
5. **for** $b = 1, \cdots, B$ **do**
    1. Draw at random a list L of length s, with replacement from $\{1, \cdots, s\}$.
    2. Estimate replication $\widehat{\tilde{S}}_{\mathbf{u}}[b]$:
$$\widehat{\tilde{S}}_{\mathbf{u}}[b] = \frac{\frac{1}{s}\sum_{k\in L} y_i\, y_i^{\mathbf{u}} - \left(\frac{1}{s}\sum_{k\in L} y_i\right)\left(\frac{1}{s}\sum_{i\in L} y_i^{\mathbf{u}}\right)}{\frac{1}{s}\sum_{i\in L} y_i^2 - \left(\frac{1}{s}\sum_{k\in L} y_i\right)^2}.$$
6. Estimate $\widehat{z}_0$:
$$\widehat{z}_0 = \Phi^{-1}\left(\frac{\#\{b \in \{1, \cdots, B\} \text{ s.t. } \widehat{\tilde{S}}_{\mathbf{u}}[b] \leq \widehat{\tilde{S}}_{\mathbf{u}}\}}{B}\right)$$
where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$.
7. Search for $z_{\alpha/2}$ so that:
$$\Phi(z_{\alpha/2}) = \alpha/2$$
and take $z_{1-\alpha/2} = -z_{\alpha/2}$, satisfying: $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$.
8. Compute $\widehat{q}(\alpha/2)$ and $\widehat{q}(1-\alpha/2)$:
$$\widehat{q}(\alpha/2) = \Phi(2\widehat{z}_0 + z_{\alpha/2}) \quad \text{and} \quad \widehat{q}(1-\alpha/2) = \Phi(2\widehat{z}_0 + z_{1-\alpha/2})$$
9. Compute $\widehat{\tilde{S}}_{\mathbf{u},\alpha/2}$ and $\widehat{\tilde{S}}_{\mathbf{u},1-\alpha/2}$, the $\widehat{q}(\alpha/2)$ and $\widehat{q}(1-\alpha/2)$ quantiles of $\{\tilde{S}_{\mathbf{u}}[1], \cdots, \tilde{S}_{\mathbf{u}}[B]\}$.

return $\left[\widehat{\tilde{S}}_{\mathbf{u},\alpha/2}; \widehat{\tilde{S}}_{\mathbf{u},1-\alpha/2}\right]$

Bibliography

Sigurd I. Aanonsen, Geir Nævdal, Dean S. Oliver, Albert C. Reynolds, Brice Vallès, et al. The ensemble kalman filter in reservoir engineering–a review. *Spe Journal*, 14(03): 393–412, 2009. (Cited on pages 69 and 116.)

Imad Abdallah, Christos Lataniotis, and Bruno Sudret. Parametric hierarchical kriging for multi-fidelity aero-servo-elastic simulators-application to extreme loads on wind turbines. *Probabilistic Engineering Mechanics*, 55:67–77, 2019. (Cited on page 21.)

Fasihul M. Alam, Ken R. McNaught, and Trevor J. Ringrose. A comparison of experimental designs in the development of a neural network simulation metamodel. *Simulation Modelling Practice and Theory*, 12(7-8):559–578, 2004. (Cited on page 18.)

Romeo Alexander, Zhizhen Zhao, Eniko Székely, and Dimitrios Giannakis. Kernel analog forecasting of tropical intraseasonal oscillations. *Journal of the Atmospheric Sciences*, 74(4):1321–1342, 2017. (Cited on page 113.)

Jaison T. Ambadan and Youmin Tang. Sigma-point kalman filter data assimilation methods for strongly nonlinear systems. *Journal of the Atmospheric Sciences*, 66(2):261–285, 2009. (Cited on page 92.)

Jeffrey L. Anderson. An ensemble adjustment kalman filter for data assimilation. *Monthly weather review*, 129(12):2884–2903, 2001. (Cited on page 69.)

Jeffrey L. Anderson. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A: Dynamic Meteorology and Oceanography*, 59(2):210–224, 2007. (Cited on page 78.)

Jeffrey L. Anderson and Stephen L. Anderson. A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. *Monthly Weather Review*, 12(127):2741–2758, 1999. (Cited on page 78.)

Theodore W. Anderson. An introduction to multivariate statistical analysis. Technical report, 1958. (Cited on pages 43, 45, and 49.)

Graeme E. Archer, Andrea Saltelli, and Il'ya M. Sobol'. Sensitivity measures, anova-like techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation*, 58(2):99–120, 1997. (Cited on pages 44, 48, and 56.)

Richard C. Aster, Brian Borchers, and Clifford H. Thurber. *Parameter estimation and inverse problems*. Elsevier, 2018. (Cited on page 69.)

Benjamin Auder and Bertrand Iooss. Global sensitivity analysis based on entropy. In *Safety, reliability and risk analysis-Proceedings of the ESREL 2008 Conference*, pages 2107–2115, 2008. (Cited on pages 17 and 44.)

Alex Ayet and Pierre Tandeo. Nowcasting solar irradiance using an analog method and geostationary satellite images. *Solar Energy*, 164:301–315, 2018. (Cited on page 113.)

François Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of gaussian processes with model misspecification. *Computational Statistics and Data Analysis*, 66:55–69, 2013. (Cited on pages 55, 60, 61, and 89.)

Michaël Baudin, Anne Dutfoy, Bertrand Iooss, and Anne-Laure Popelin. Open turns: An industrial software for uncertainty quantification in simulation, 2017. (Cited on pages xv and 11.)

Laurent Bertino, Geir Evensen, and Hans Wackernagel. Sequential data assimilation techniques in oceanography. *International Statistical Review*, 71(2):223–241, 2003. (Cited on page 112.)

Marc Berveiller. *Eléments finis stochastiques : approches intrusive et non intrusive pour des analyses de fiabilité*. PhD thesis, Université Blaise Pascal, Clermont-Ferrand, 2005. (Cited on page 18.)

Philippe Besse. PCA stability and choice of dimensionality. *Statistics and Probability Letters*, 13(5):405–410, 1992. (Cited on pages 43, 45, and 49.)

Christopher M. Bishop. *Pattern recognition and machine learning*. springer, 2006. (Cited on page 70.)

Éric Blayo, Marc Bocquet, Emmanuel Cosme, and Leticia F. Cugliandolo. *Advanced Data Assimilation for Geosciences: Lecture Notes of the Les Houches School of Physics: Special Issue, June 2012*. Oxford University Press, USA, 2014. (Cited on pages 20 and 68.)

Frédéric Blondel, Ronan Boisard, Malika Milekovic, Gilles Ferrer, Caroline Lienard, and David Teixeira. Validation and comparison of aerodynamic modelling approaches for wind turbines. In *Journal of Physics: Conference Series*, volume 753, page 24, 2016. (Cited on page 32.)

Marc Bocquet, Alban Farchi, and Quentin Malartic. Online learning of both state and dynamics using ensemble kalman filters. *arXiv preprint arXiv:2006.03859*, 2020. (Cited on pages 107, 112, and 129.)

Cord Böker. *Load simulation and local dynamics of support structures for offshore wind turbines*. Shaker, 2010. (Cited on page 34.)

Emanuele Borgonovo. A new uncertainty importance measure. *Reliability Engineering and System Safety*, 92(6):771–784, 2007. (Cited on pages 17 and 44.)

Emanuele Borgonovo and Elmar Plischke. Sensitivity analysis: a review of recent advances. *European Journal of Operational Research*, 248(3):869–887, 2016. (Cited on page 44.)

Emanuele Borgonovo, William Castaings, and Stefano Tarantola. Moment independent importance measures: new results and analytical test cases. *Risk Analysis: An International Journal*, 31(3):404–428, 2011. (Cited on page 44.)

Carlo Luigi Bottasso, Alessandro Croce, Carlo E.D. Riboldi, Gunjit S. Bir, et al. Spatial estimation of wind states from the aeroelastic response of a wind turbine. *The Science of Making Torque from Wind (TORQUE 2010), Heraklion, Crete, Greece*, pages 28–30, 2010. (Cited on page 22.)

George E.P. Box and Norman R. Draper. *Empirical model-building and response surfaces.* John Wiley and Sons, 1987. (Cited on pages 18, 54, and 55.)

Emmanuel Branlard, Jason M. Jonkman, Scott Dana, and Paula Doubrawa. A digital twin based on openfast linearizations for real-time load and fatigue estimation of land-based turbines. In *Journal of Physics: Conference Series*, volume 1618, page 022030. IOP Publishing, 2020. (Cited on pages 22 and 85.)

Steven D. Brown. The kalman filter in analytical chemistry. *Analytica chimica acta*, 181: 1–26, 1986. (Cited on pages 73 and 115.)

Thomas Browne, Bertrand Iooss, Loïc Le Gratiet, Jérôme Lonchampt, and Emmanuel Remy. Stochastic simulators based optimization by gaussian process metamodels–application to maintenance investments planning issues. *Quality and Reliability Engineering International*, 32(6):2067–2080, 2016. (Cited on page 64.)

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016. (Cited on pages 107 and 112.)

Gerrit Burgers, Peter Jan van Leeuwen, and Geir Evensen. Analysis scheme in the ensemble kalman filter. *Monthly weather review*, 126(6):1719–1724, 1998. (Cited on page 76.)

Tony Burton, David Sharpe, Nick Jenkins, and Ervin Bossanyi. *Wind energy handbook*, volume 2. Wiley Online Library, 2001. (Cited on pages xv, 27, 31, and 32.)

Dan G. Cacuci. *Sensitivity and Uncertainty Analysis, Volume 1: Theory*, volume 1. CRC Press, 2003. (Cited on page 14.)

Russel E. Caflisch. Monte Carlo and Quasi-Monte Carlo methods. *Acta Numerica*, 7: 1–49, 1998. (Cited on page 13.)

Katherine Campbell, Michael D. McKay, and Brian J. Williams. Sensitivity analysis when model outputs are functions. *Reliability Engineering and System Safety*, 91(10-11):1468–1472, 2006. (Cited on pages 45 and 49.)

Francesca Campolongo, Andrea Saltelli, and Jessica Cariboni. From screening to quantitative sensitivity analysis. a unified approach. *Computer Physics Communications*, 182 (4):978–988, 2011. (Cited on page 15.)

Mathieu Carmassi, Pierre Barbillon, Merlin Keller, Eric Parent, and Matthieu Chiodetti. Bayesian calibration of a numerical code for prediction. *arXiv preprint arXiv:1801.01810*, 2018. (Cited on page 68.)

Rob Carnell. lhs: Latin hypercube samples. *R package version 0.10, URL http://CRAN. R-project. org/package= lhs*, 2012. (Cited on pages xv, 51, and 52.)

Thi Tuyet Trang Chau, Pierre Ailliot, and Valérie Monbet. An algorithm for non-parametric estimation in state–space models. *Computational Statistics & Data Analysis*, 153:107062, 2021. (Cited on page 113.)

Lei Chen. *Curse of Dimensionality*, pages 545–546. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_133. URL https://doi.org/10.1007/978-0-387-39940-9_133. (Cited on page 112.)

Wei Chen, Ruichen Jin, and Agus Sudjianto. Analytical metamodel-based global sensitivity analysis and uncertainty propagation for robust design. In *SAE 2004 World Congress and Exhibition*. SAE International, mar 2004. doi: https://doi.org/10.4271/2004-01-0429. URL https://doi.org/10.4271/2004-01-0429. (Cited on page 19.)

Wei Chen, Ruichen Jin, and Agus Sudjianto. Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty. *Transactions-American Society of Mechanical Engineers Journal of Mechanical Design*, 127(5):875, 2005. (Cited on page 56.)

Seung-Kyum Choi, Ramana V. Grandhi, Robert A. Canfield, and Chris L. Pettit. Polynomial chaos expansion with latin hypercube sampling for estimating response variability. *AIAA journal*, 42(6):1191–1198, 2004. (Cited on page 19.)

Nicolas Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3): 539–552, 2002. (Cited on page 20.)

Andy Clifton, Levi Kilcher, Julie K. Lundquist, and Paul Fleming. Using machine learning to predict wind turbine power output. *Environmental research letters*, 8(2):024009, 2013. (Cited on page 21.)

Andy Clifton, Megan H. Daniels, and Michael Lehning. Effect of winds in a mountain pass on turbine performance. *Wind Energy*, 17(10):1543–1562, 2014. (Cited on page 21.)

Joel P. Conte, Rodrigo Astroza, and Hamed Ebrahimian. Bayesian methods for nonlinear system identification of civil structures. In *MATEC Web of Conferences*, volume 24, page 03002. EDP Sciences, 2015. (Cited on page 20.)

Nicolai Cosack. *Fatigue load monitoring with standard wind turbine signals*. PhD thesis, 01 2011. (Cited on pages 37 and 96.)

Roy R. Craig Jr. and Mervyn C.C. Bampton. Coupling of substructures for dynamic analyses. *AIAA journal*, 6(7):1313–1319, 1968. (Cited on page 35.)

Robert I. Cukier, CM Fortuin, Kurt E Shuler, AG Petschek, and JH Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory. *The Journal of chemical physics*, 59(8):3873–3878, 1973. (Cited on page 43.)

Robert I. Cukier, Howard B. Levine, and Kurt E. Shuler. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of computational physics*, 26(1):1–42, 1978. (Cited on page 17.)

Sebastien Da Veiga. Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305, 2015. (Cited on page 17.)

JC Dai, YP Hu, DS Liu, and X Long. Aerodynamic loads calculation and analysis for large scale wind turbine based on combining bem modified theory with dynamic stall model. *Renewable Energy*, 36(3):1095–1104, 2011. (Cited on page 31.)

Etienne De Rocquigny, Nicolas Devictor, and Stefano Tarantola. *Uncertainty in industrial practice: a guide to quantitative uncertainty management*. John Wiley and Sons, 2008. (Cited on pages 10 and 85.)

Gregory Deman, Katerina Konakli, Bruno Sudret, Jaouhar Kerrou, Pierre Perrochet, and Hakim Benabderrahmane. Using sparse polynomial chaos expansions for the global sensitivity analysis of groundwater lifetime expectancy in a multi-layered hydrogeological model. *Reliability Engineering and System Safety*, 147:156–169, 2015. (Cited on page 18.)

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. (Cited on page 10.)

Nikolay Dimitrov, Anand Natarajan, and Mark Kelly. Model of wind shear conditional on turbulence and its impact on wind turbine loads. *Wind Energy*, 18(11):1917–1931, 2015. (Cited on pages 21, 109, and 110.)

Nikolay Dimitrov, Anand Natarajan, and Jakob Mann. Effects of normal and extreme turbulence spectral parameters on wind turbine loads. *Renewable Energy*, 101:1180–1193, 2017. (Cited on page 110.)

DNV GL. Guideline for the certification of wind turbines 2010 edition. *DNV GL*, 2010. (Cited on page 26.)

DNV GL. Bladed: Wind turbine design software, 2013. (Cited on pages 34, 84, and 95.)

DNV GL. Loads and site conditions for wind turbines. Standard, November 2016. (Cited on page 97.)

Simona Dobre, Thierry Bastogne, Christophe Profeta, Muriel Barberi-Heyob, and Alain Richard. Limits of variance-based sensitivity analysis for non-identifiability testing in high dimensional dynamic models. *Automatica*, 48(11):2740–2749, 2012. (Cited on pages 84, 87, 101, 121, 128, and 129.)

Gérard Dreyfus. *Neural networks: methodology and applications*. Springer Science and Business Media, 2005. (Cited on pages 54 and 55.)

Sylvain Dubreuil, Marc Berveiller, Frank Petitjean, and Michel Salaün. Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion. *Reliability Engineering and System Safety*, 121:263–275, 2014. (Cited on page 18.)

Serhat Duran. Computer-aided design of horizontal-axis wind turbine blades. Master's thesis, 2005. (Cited on page 31.)

James Durbin and Siem J. Koopman. *Time series analysis by state space methods*. Oxford university press, 2012. (Cited on page 112.)

David Duvenaud. The kernel cookbook: Advice on covariance functions, 2014. (Cited on pages 55, 89, and 129.)

Bradley Efron. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981. (Cited on page 132.)

Bradley Efron and Charles L. Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981. (Cited on page 45.)

Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986. (Cited on page 132.)

Bradley Efron and Robert J. Tibshirani. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436, 1993. (Cited on page 44.)

Mette Eknes and Geir Evensen. An ensemble kalman filter with a 1-d marine ecosystem model. *Journal of Marine Systems*, 36(1-2):75–100, 2002. (Cited on page 102.)

Alexandre A. Emerick and Albert C. Reynolds. History matching time-lapse seismic data using the ensemble kalman filter with multiple data assimilations. *Computational Geosciences*, 16(3):639–659, 2012. (Cited on pages 20 and 68.)

Pierre Etoré, Clémentine Prieur, Dang Khoi Pham, and Long Li. Global sensitivity analysis for models described by stochastic differential equations. working paper or preprint, November 2018. URL https://hal.archives-ouvertes.fr/hal-01926919. (Cited on page 88.)

Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994. (Cited on pages 67, 75, and 115.)

Geir Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003. (Cited on page 116.)

Geir Evensen. *Data assimilation: the ensemble Kalman filter*. Springer Science and Business Media, 2009. (Cited on pages 20, 69, 84, 87, 92, 93, 102, 115, and 116.)

Geir Evensen and Peter Jan Van Leeuwen. Assimilation of geosat altimeter data for the agulhas current using the ensemble kalman filter with a quasigeostrophic model. *Monthly Weather Review*, 124(1):85–96, 1996. (Cited on pages 69 and 75.)

Ronan Fablet, Phi Viet, Redouane Lguensat, and Bertrand Chapron. Data-driven assimilation of irregularly-sampled image time series. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4302–4306. IEEE, 2017. (Cited on page 113.)

Kai-Tai Fang, Runze Li, and Agus Sudjianto. *Design and modeling for computer experiments*. CRC press, 2005. (Cited on pages 18 and 61.)

Jean-Pierre Fargues. *Modélisation dynamique des risers pétroliers en grands déplacements*. PhD thesis, Châtenay-Malabry, Ecole centrale de Paris, 1995. (Cited on page 35.)

Robert Fitzgerald. Divergence of the kalman filter. *IEEE Transactions on Automatic Control*, 16(6):736–747, 1971. (Cited on page 75.)

Alexander I.J. Forrester and Andy J. Keane. Recent advances in surrogate-based optimization. *Progress in aerospace sciences*, 45(1-3):50–79, 2009. (Cited on page 87.)

Alexander I.J. Forrester, András Sóbester, and Andy J. Keane. Multi-fidelity optimization via surrogate modelling. *Proceedings of the royal society a: mathematical, physical and engineering sciences*, 463(2088):3251–3269, 2007. (Cited on page 86.)

Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. *Advances in neural information processing systems*, 29: 2199–2207, 2016. (Cited on page 129.)

Gregg Freebury and Walter Musial. Determining equivalent damage loading for full-scale wind turbine blade fatigue tests. In *2000 ASME Wind Energy Symposium*, page 50, 01 2000. doi: 10.2514/6.2000-50. (Cited on page 86.)

Kai Freudenreich and Kimon Argyriadis. Wind turbine load level based on extrapolation and simplified methods. *Wind Energy*, 11(6):589–600, 2008. doi: https://doi.org/10.1002/we.279. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/we.279. (Cited on page 26.)

Jerome H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77, 1997. (Cited on page 112.)

Fabrice Gamboa, Alexandre Janon, Thierry Klein, and Agnès Lagnoux. Sensitivity indices for multivariate outputs. *arXiv preprint arXiv:1303.3574*, 2013. (Cited on page 49.)

Roger Ghanem and Pol D. Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003. (Cited on page 43.)

Roger Ghanem, David Higdon, and Houman Owhadi. *Handbook of uncertainty quantification*, volume 6. Springer, 2017. (Cited on pages 10 and 12.)

Michael Ghil and Paola Malanotte-Rizzoli. Data assimilation in meteorology and oceanography. In *Advances in geophysics*, volume 33, pages 141–266. Elsevier, 1991. (Cited on pages 20, 68, and 75.)

Michael Ghil, Stephen Cohn, John Tavantzis, K. Bube, and Eugene Isaacson. Applications of estimation theory to numerical weather prediction. In *Dynamic meteorology: Data assimilation methods*, pages 139–224. Springer, 1981. (Cited on page 75.)

Walter R. Gilks and Carlo Berzuini. Following a moving target-monte carlo inference for dynamic bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):127–146, 2001. (Cited on page 13.)

David Ginsbourger, Victor Picheny, Olivier Roustant, and Yann Richet. Kriging with Heterogeneous Nugget Effect for the Approximation of Noisy Simulators with Tunable Fidelity (Krigeage avec effet de pépite hétérogène pour l'approximation de simulateurs bruités à fidélité réglable). In *Congrès conjoint de la Société Statistique du Canada et de la SFdS*, pages –, Ottawa, Canada, May 2008. URL https://hal.archives-ouvertes.fr/hal-00409766. (Cited on page 65.)

Peter A. Graf, Gordon Stewart, Matthew Lackner, Katherine Dykes, and Paul Veers. High-throughput computation and the applicability of monte carlo integration in fatigue load estimation of floating offshore wind turbines. *Wind Energy*, 19(5):861–872, 2016. (Cited on page 21.)

Kurtis Gurley and Ahsan Kareem. Applications of wavelet transforms in earthquake, wind and ocean engineering. *Engineering structures*, 21(2):149–167, 1999. (Cited on page 28.)

Hadamard, Jacques. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton university bulletin*, pages 49–52, 1902. (Cited on pages 19 and 87.)

Lorenz Haid, Gordon Stewart, Jason Jonkman, Amy N. Robertson, Matthew Lackner, and Denis Matha. Simulation-length requirements in the loads analysis of offshore floating wind turbines. In *International Conference on Offshore Mechanics and Arctic Engineering*, volume 55423, page V008T09A091. American Society of Mechanical Engineers, 2013. (Cited on page 36.)

Franz Hamilton, Tyrus Berry, and Timothy Sauer. Ensemble kalman filtering without a model. *Physical Review X*, 6(1):011021, 2016. (Cited on page 112.)

Joseph L. Hart, Alen Alexanderian, and Pierre A. Gremaud. Efficient computation of sobol'indices for stochastic models. *SIAM Journal on Scientific Computing*, 39(4): A1514–A1530, 2017. (Cited on pages 64 and 85.)

Andrew C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge university press, 1990. (Cited on pages 73 and 115.)

Yutaka Hasegawa, Junsuke Murata, Hiroshi Imahura, Kai Karikomi, and Naoyuki Yonezawa. Aerodynamic loads calucaltion of a horizontal axis wind turbine rotor in conbined inflow condition. 01 2004. (Cited on pages xv and 28.)

W. Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970. (Cited on pages 13, 20, and 68.)

Simon Haykin. *Kalman filtering and neural networks*, volume 47. John Wiley & Sons, 2004. (Cited on page 115.)

Jon C. Helton and Freddie J. Davis. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety*, 81(1):23 – 69, 2003. ISSN 0951-8320. doi: https://doi.org/10.1016/S0951-8320(03)00058-9. URL http://www.sciencedirect.com/science/article/pii/S0951832003000589. (Cited on page 18.)

Jon C. Helton, Jay Dean Johnson, Cedric J. Sallaberry, and Curt B. Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety*, 91(10-11):1175–1209, 2006. (Cited on pages 44 and 47.)

Adrien Hirvoas, Clémentine Prieur, Elise Arnaud, Fabien Caleyron, and Miguel Munoz Zuniga. Quantification and reduction of uncertainties in a wind turbine numerical model based on a global sensitivity analysis and a recursive bayesian inference approach. *International Journal for Numerical Methods in Engineering*, n/a(n/a). doi: https://doi.org/10.1002/nme.6630. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.6630. (Cited on pages 84, 106, 107, 110, 111, 112, and 120.)

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. pages 293–325. JSTOR, 1948. (Cited on pages 44, 45, and 89.)

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008. (Cited on page 129.)

Jessica Holierhoek, Hendrik Korterink, René van de Pieterman, Luc Rademakers, and Denja Lekou. Recommended Practices for Measuring in Situ the 'Loads' on Drive Train, Pitch System and Yaw System. *Energy Research Center of the Netherlands (ECN)*, 2010. (Cited on pages 98 and 111.)

Toshimitsu Homma and Andrea Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety*, 52(1):1–17, 1996. (Cited on pages 17, 44, 48, 56, 64, and 91.)

Stephen Hora and Ronald L. Iman. A comparison of maximum/bounding and bayesian/-monte carlo for fault tree uncertainty analysis. *Report SAND85-2839, Sandia Laboratories*, 1986. (Cited on page 17.)

Peter L. Houtekamer and Herschel L. Mitchell. A sequential ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 129(1):123–137, 2001. (Cited on pages 75, 76, 92, and 116.)

Peter L. Houtekamer, Herschel L. Mitchell, Gérard Pellerin, Mark Buehner, Martin Charron, Lubos Spacek, and Bjarne Hansen. Atmospheric data assimilation with an ensemble kalman filter: Results with real observations. *Monthly weather review*, 133(3): 604–620, 2005. (Cited on page 102.)

Peter L. Houtekamer, Bin He, and Herschel L. Mitchell. Parallel implementation of an ensemble kalman filter. *Monthly Weather Review*, 142(3):1163–1182, 2014. (Cited on pages 69 and 116.)

Clemens Hübler, Cristian Guillermo Gebhardt, and Raimund Rolfes. Hierarchical four-step global sensitivity analysis of offshore wind turbines based on aeroelastic time domain simulations. *Renewable Energy*, 111:878–891, 2017. (Cited on page 21.)

IEC. Iec 61400-1. *Wind Turbines–Part*, 1, 2005. (Cited on pages xvi, xix, 26, 28, 36, 96, 97, 107, and 108.)

Marco A. Iglesias, Kody J.H. Law, and Andrew M. Stuart. Ensemble kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, 2013. doi: 10.1088/0266-5611/29/4/045001. URL https://doi.org/10.1088%2F0266-5611%2F29%2F4%2F045001. (Cited on pages 69 and 84.)

Ronald L. Iman and Stephen C. Hora. A robust measure of uncertainty importance for use in fault tree system analysis. *Risk analysis*, 10(3):401–406, 1990. (Cited on page 17.)

Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*, pages 101–122. Springer, 2015. (Cited on pages 16 and 18.)

Bertrand Iooss and Mathieu Ribatet. Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering and System Safety*, 94(7):1194–1204, 2009. (Cited on page 64.)

Bertrand Iooss and Andrea Saltelli. Introduction to Sensitivity Analysis. In *Handbook of Uncertainty Quantification*, pages 1–20. Springer International Publishing, 2016. doi: 10.1007/978-3-319-11259-6_31-1. (Cited on page 18.)

Bertrand Iooss, Christian Lhuillier, and Hélène Jeanneau. Numerical simulation of transit-time ultrasonic flowmeters: uncertainties due to flow profile and fluid turbulence. *Ultrasonics*, 40(9):1009–1015, 2002. (Cited on page 64.)

Bertrand Iooss, François Van Dorpe, and Nicolas Devictor. Response surfaces and sensitivity analyses for an environmental model of dose calculations. *Reliability Engineering and System Safety*, 91(10-11):1241–1251, 2006. (Cited on pages 48 and 54.)

Bertrand Iooss, Loïc Boussouf, Vincent Feuillard, and Amandine Marrel. Numerical studies of the metamodel fitting and validation processes. *International Journal on Advances in Systems and Measurements*, 3(1):11–21, 2010. (Cited on page 18.)

Bertrand Iooss, Alexandre Janon, Gilles Pujol, with contributions from Baptiste Broto, Khalid Boumhaout, Sebastien Da Veiga, Thibault Delage, Jana Fruth, Laurent Gilquin,

Joseph Guillaume, Loic Le Gratiet, Paul Lemaitre, Barry L. Nelson, Filippo Monari, Roelof Oomen, Oldrich Rakovec, Bernardo Ramos, Olivier Roustant, Eunhye Song, Jeremy Staum, Roman Sueur, Taieb Touati, and Frank Weber. *sensitivity: Global Sensitivity Analysis of Model Outputs*, 2019. URL https://CRAN.R-project.org/package=sensitivity. R package version 1.16.0. (Cited on pages 91 and 120.)

Alexandre Janon. *Analyse de sensibilité et réduction de dimension. Application à l'océanographie*. PhD thesis, 2012. (Cited on page 91.)

Alexandre Janon, Maëlle Nodet, and Clémentine Prieur. Confidence intervals for sensitivity indices using reduced-basis metamodels. working paper or preprint, February 2011. URL https://hal.inria.fr/inria-00567977. (Cited on page 133.)

Alexandre Janon, Maëlle Nodet, and Clémentine Prieur. Certified reduced-basis solutions of viscous burgers equation parametrized by initial and boundary values. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47(2):317–348, 2013. (Cited on page 18.)

Alexandre Janon, Thierry Klein, Agnes Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics*, 18:342–364, 2014. (Cited on pages 44, 47, 48, and 50.)

Michiel J.W. Jansen. Analysis of variance designs for model output. *Computer Physics Communications*, 117(1-2):35–43, 1999. (Cited on pages 48, 56, 64, 99, and 120.)

Tongdan Jin and Zhigang Tian. Uncertainty analysis for wind energy production with dynamic power curves. In *2010 IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems*, pages 745–750. IEEE, 2010. (Cited on page 85.)

Mark E. Johnson, Leslie M. Moore, and Donald Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2):131 – 148, 1990. ISSN 0378-3758. doi: https://doi.org/10.1016/0378-3758(90)90122-B. URL http://www.sciencedirect.com/science/article/pii/037837589090122B. (Cited on pages 51 and 92.)

Ian T. Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986. (Cited on pages 43, 45, and 49.)

Bonnie J. Jonkman. Turbsim user's guide: Version 1.50. Technical report, National Renewable Energy Lab (NREL), Golden, CO, USA, 2009. (Cited on pages xv, 26, 29, 30, 95, 96, and 118.)

Jason M. Jonkman. Modeling of the uae wind turbine for refinement of fast {_} ad. Technical report, National Renewable Energy Lab., Golden, CO (US), 2003. (Cited on page 36.)

Jason M. Jonkman. Dynamics modeling and loads analysis of an offshore floating wind turbine. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2007. (Cited on page 34.)

Jason M . Jonkman, Marshall L. Buhl Jr., et al. Fast user's guide. *Golden, CO: National Renewable Energy Laboratory*, 365:366, 2005. (Cited on pages 35 and 118.)

Jason M. Jonkman and Marshall L. Buhl Jr. Fast users guide. *Technical Report No. NREL/EL-500-38230*, 2005. (Cited on pages 84 and 85.)

Jason M. Jonkman and Bonnie J. Jonkman. Fast modularization framework for wind turbine simulation: full-system linearization. In *Journal of Physics: Conference Series*, volume 753, 2016. (Cited on pages 34, 35, and 36.)

Jason M. Jonkman and Walter Musial. Offshore code comparison collaboration (oc3) for iea wind task 23 offshore wind technology and deployment. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2010. (Cited on page 34.)

Jason M. Jonkman, Sandy Butterfield, Walter Musial, and George Scott. Definition of a 5-mw reference wind turbine for offshore system development. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2009. (Cited on page 34.)

Simon J. Julier and Jeffrey K. Uhlmann. New extension of the kalman filter to nonlinear systems. In *Signal processing, sensor fusion, and target recognition VI*, volume 3068, pages 182–193. International Society for Optics and Photonics, 1997. (Cited on page 115.)

John C. Kaimal, John C.J. Wyngaard, Y. Izumi, and OR Coté. Spectral characteristics of surface-layer turbulence. *Quarterly Journal of the Royal Meteorological Society*, 98 (417):563–589, 1972. (Cited on pages 26 and 28.)

Rudolf E. Kalman et al. Contributions to the theory of optimal control. *Bol. soc. mat. mexicana*, 5(2):102–119, 1960. (Cited on pages 20 and 91.)

Rudolph E. Kalman. A new approach to linear filtering and prediction problems. 1960. (Cited on pages 67, 72, and 115.)

Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge university press, 2003. (Cited on page 68.)

Assem Kanso, Ghassan Chebbo, and Bruno Tassin. Application of mcmc–gsa model calibration method to urban runoff quality modeling. *Reliability engineering and system safety*, 91(10-11):1398–1405, 2006. (Cited on page 68.)

Nicholas Kantas, Arnaud Doucet, Sumeetpal Sindhu Singh, and Jan M. Maciejowski. An overview of sequential monte carlo methods for parameter estimation in general state-space models. *IFAC Proceedings Volumes*, 42(10):774–785, 2009. (Cited on pages 67 and 68.)

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001. (Cited on pages 19 and 68.)

Richard E. Kopp and Richard J. Orford. Linear regression applied to system identification for adaptive control systems. *Aiaa Journal*, 1(10):2300–2306, 1963. (Cited on page 92.)

Christina Koukoura. *Validated Loads Prediction Models for Offshore Wind Turbines for Enhanced Component Reliability*. PhD thesis, Technical University of Denmark, 2014. (Cited on pages 98 and 111.)

Nikola B. Kovachki and Andrew M. Stuart. Ensemble kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, 35(9):095005, 2019. (Cited on pages 69, 84, and 91.)

Daniel G. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52 (6):119–139, 1951. (Cited on pages 19, 54, and 55.)

Daniel G. Krige, Massimo Guarascio, and Ferdi A. Camisani-Calzolari. Early South African geostatistical techniques in today's perspective. *Geostatistics*, 1:1–19, 1989. (Cited on page 89.)

Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015. (Cited on page 129.)

Bernard Krzykacz-Hausmann. Epistemic sensitivity analysis based on the concept of entropy. *Proceedings of SAMO2001*, pages 31–35, 2001. (Cited on pages 17 and 44.)

Emrah Kulunk and Nadir Yilmaz. Hawt rotor design and performance analysis. In *Energy Sustainability*, volume 48906, pages 1019–1029, 2009. (Cited on page 31.)

Soon-Duck Kwon. Uncertainty analysis of wind energy potential assessment. *Applied Energy*, 87(3):856–865, 2010. (Cited on page 85.)

Matieyendou Lamboni, David Makowski, and Hervé Monod. Multivariate global sensitivity analysis for discrete-time models. 2008. (Cited on pages 43 and 45.)

Matieyendou Lamboni, Hervé Monod, and David Makowski. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering and System Safety*, 96(4):450–459, 2011. doi: 10.1016/j.ress.2010.12.002. (Cited on pages 18, 49, 87, 101, and 121.)

Matieyendou Lamboni, Bertrand Iooss, Anne-Laure Popelin, and Fabrice Gamboa. Derivative-based global sensitivity measures: General links with sobol' indices and numerical tests. *Mathematics and Computers in Simulation*, 87:45 – 54, 2013. ISSN 0378-4754. doi: https://doi.org/10.1016/j.matcom.2013.02.002. URL http://www.sciencedirect.com/science/article/pii/S0378475413000141. (Cited on page 16.)

Torben J. Larsen and Anders Melchior Hansen. *How 2 HAWC2, the user's manual*, 2007. (Cited on pages 34 and 95.)

Gregory W. Lawson and James A. Hansen. Implications of stochastic and deterministic filters as ensemble-based data assimilation methods in varying regimes of error growth. *Monthly weather review*, 132(8):1966–1981, 2004. (Cited on page 93.)

Cédric Le Cunff, Jean-Michel Heurtier, Loïc Piriou, Christian Berhault, Timothée Perdrizet, David Teixeira, Gilles Ferrer, and Jean-Christophe Gilloteaux. Fully coupled floating wind turbine simulator based on nonlinear finite element method: Part i methodology. In *ASME 2013 32nd International Conference on Ocean, Offshore and Arctic Engineering*, pages V008T09A050–V008T09A050. Citeseer, 2013. (Cited on pages 36, 84, and 95.)

François Le Gland, Valérie Monbet, and Vu-Duc Tran. Large sample asymptotics for the ensemble kalman filter. 2009. (Cited on page 77.)

Loic Le Gratiet, Claire Cannamela, and Bertrand Iooss. A bayesian approach for global sensitivity analysis of (multifidelity) computer codes. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):336–363, 2013. doi: 10.1137/130926869. (Cited on pages 19, 54, 56, 62, 63, 89, 90, 91, and 99.)

Loic Le Gratiet, Saltelli Marelli, and Bruno Sudret. Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes. In *Handbook of Uncertainty Quantification*, pages 1–37. Springer International Publishing, 2017. doi: 10.1007/978-3-319-11259-6_38-1. (Cited on pages 19, 55, and 87.)

Olivier Le Maître and Omar M. Knio. *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer Science and Business Media, 2010. (Cited on page 18.)

Olivier Le Maître, Omar M. Knio, Habib N. Najm, and Roger Ghanem. Uncertainty propagation using wiener–haar expansions. *Journal of computational Physics*, 197(1): 28–57, 2004. (Cited on page 19.)

Sang Hoon Lee and Wei Chen. A comparative study of uncertainty propagation methods for black-box-type problems. *Structural and Multidisciplinary Optimization*, 37(3):239, 2009. (Cited on page 12.)

J. Gordon Leishman. Principles of helicopter aerodynamics. 2000. (Cited on page 30.)

Redouane Lguensat, Pierre Tandeo, Pierre Ailliot, Manuel Pulido, and Ronan Fablet. The analog data assimilation. *Monthly Weather Review*, 145(10):4093–4107, 2017. (Cited on pages 106, 107, 112, 113, 114, 115, and 116.)

Gaoming Li, Albert C. Reynolds, et al. An iterative ensemble kalman filter for data assimilation. In *SPE annual technical conference and exhibition*. Society of Petroleum Engineers, 2007. (Cited on page 79.)

Ge Liang, Cheung Kwok Fai, and Marcelo H. Kobayashi. Stochastic solution for uncertainty propagation in nonlinear shallow-water equations. *Journal of Hydraulic Engineering*, 134(12):1732–1743, 2008. doi: 10.1061/(ASCE)0733-9429(2008)134:12(1732). (Cited on page 18.)

Huibin Liu, Wei Chen, and Agus Sudjianto. Relative entropy based method for probabilistic sensitivity analysis in engineering design. 2006. (Cited on pages 17 and 44.)

Brian A. Lockwood and Mihai Anitescu. Gradient-enhanced universal kriging for uncertainty propagation. *Nuclear Science and Engineering*, pages 1–32, 2012. (Cited on page 19.)

Edward N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric sciences*, 26(4):636–646, 1969. (Cited on pages 107 and 112.)

Jakob Mann. Wind field simulation. *Probabilistic engineering mechanics*, 13(4):269–282, 1998. (Cited on pages 26 and 28.)

James F. Manwell, Jon G. McGowan, and Anthony L. Rogers. *Wind energy explained: theory, design and application*. John Wiley & Sons, 2010. (Cited on page 31.)

Amandine Marrel. *Mise en oeuvre et exploitation du métamodèle processus gaussien pour l'analyse de modèles numériques-Application à un code de transport hydrogéologique*. PhD thesis, INSA Toulouse, 2008. (Cited on page 54.)

Amandine Marrel, Bertrand Iooss, François Van Dorpe, and Elena Volkova. An efficient methodology for modeling complex computer codes with gaussian processes. *Computational Statistics and Data Analysis*, 52(10):4731–4744, 2008. (Cited on pages 54 and 57.)

Amandine Marrel, Bertrand Iooss, Beatrice Laurent, and Olivier Roustant. Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering and System Safety*, 94(3):742 – 751, 2009. doi: http://dx.doi.org/10.1016/j.ress.2008.07.008. (Cited on pages 19, 62, and 90.)

Amandine Marrel, Bertrand Iooss, Sébastien Da Veiga, and Mathieu Ribatet. Global sensitivity analysis of stochastic computer models with joint metamodels. *Statistics and Computing*, 22(3):833–847, 2012. (Cited on page 129.)

Amandine Marrel, Nadia Perot, and Clémentine Mottet. Development of a surrogate model and sensitivity analysis for spatio-temporal numerical simulators. *Stochastic environmental research and risk assessment*, 29(3):959–974, 2015a. (Cited on page 45.)

Amandine Marrel, Nathalie Saint-Geours, and Matthias De Lozzo. Sensitivity Analysis of Spatial and/or Temporal Phenomena. In *Handbook of Uncertainty Quantification*, pages 1–31. Springer International Publishing, 2015b. ISBN 978-3-319-11259-6. doi: 10.1007/978-3-319-11259-6_39-1. (Cited on pages 19 and 55.)

Jay D. Martin and Timothy W. Simpson. Use of kriging models to approximate deterministic computer models. *AIAA journal*, 43(4):853–863, 2005. (Cited on page 57.)

Youssef M. Marzouk, Habib N. Najm, and Larry A. Rahn. Stochastic spectral methods for efficient bayesian solution of inverse problems. *Journal of Computational Physics*, 224(2):560–586, 2007. (Cited on pages 20 and 68.)

Denis Matha. Model development and loads analysis of an offshore wind turbine on a tension leg platform with a comparison to other floating turbine concepts: April 2009. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2010. (Cited on page 37.)

Michael D. McKay. Nonparametric variance-based methods of assessing uncertainty importance. *Reliability engineering and system safety*, 57(3):267–279, 1997. (Cited on page 17.)

Michael D. McKay, Richard J. Beckman, and William J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 1979. (Cited on pages 45, 50, and 90.)

MA Miner et al. Cumulative fatigue damage. *Journal of applied mechanics*, 12(3):A159–A164, 1945. (Cited on page 37.)

Philippe Moireau and Dominique Chapelle. Reduced-order unscented kalman filtering with application to parameter identification in large-dimensional systems. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(2):380–405, 2011. (Cited on page 67.)

Hervé Monod, Cédric Naud, and David Makowski. Uncertainty and sensitivity analysis for crop models. *Working with dynamic crop models: Evaluation, analysis, parameterization, and applications*, 4:55–100, 2006. (Cited on pages 44, 48, 56, and 64.)

Alexandre Morató, Srinivas Sriramula, and Nandakumar Krishnan. Kriging models for aero-elastic simulations and reliability analysis of offshore wind turbine support structures. *Ships and Offshore Structures*, 14(6):545–558, 2019. doi: 10.1080/17445302.2018.1522738. URL https://doi.org/10.1080/17445302.2018.1522738. (Cited on page 21.)

Patrick J. Moriarty and A. Craig Hansen. Aerodyn theory manual. Technical report, National Renewable Energy Lab., Golden, CO (US), 2005. (Cited on page 32.)

Max D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991. (Cited on page 15.)

Juan Pablo Murcia, Pierre-Elouan Réthoré, Nikolay Dimitrov, Anand Natarajan, John D. Sørensen, Peter Graf, and Taeseong Kim. Uncertainty propagation through an aeroelastic wind turbine model using polynomial surrogates. *Renewable Energy*, 119:910–922, 2018. (Cited on pages 21, 85, and 129.)

Kevin P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. (Cited on page 70.)

Matthew Muto and James L. Beck. Bayesian updating and model class selection for hysteretic structural models using stochastic simulation. *Journal of Vibration and Control*, 14(1-2):7–34, 2008. doi: 10.1177/1077546307079400. URL https://doi.org/10.1177/1077546307079400. (Cited on page 19.)

Vicente Negro, José-Santos López-Gutiérrez, M. Dolores Esteban, and Clara Matutano. Uncertainties in the design of support structures and foundations for offshore wind turbines. *Renewable energy*, 63:125–132, 2014. (Cited on page 20.)

Christopher Nemeth, Paul Fearnhead, and Lyudmila Mihaylova. Sequential monte carlo methods for state and parameter estimation in abruptly changing environments. *IEEE Transactions on Signal Processing*, 62(5):1245–1255, 2013. (Cited on page 67.)

Nathan M. Newmark. A method of computation for structural dynamics. *Journal of the engineering mechanics division*, 85(3):67–94, 1959. (Cited on page 35.)

Harald Niederreiter. Quasi-Monte Carlo methods and pseudo-random numbers. *Bulletin of the American Mathematical Society*, 84(6):957–1041, 1978. (Cited on page 13.)

The National Renewable Energy Laboratory NREL. Openfast. http://openfast.readthedocs.io, 2018. (Cited on page 95.)

Jeremy E. Oakley and Anthony O'Hagan. Probabilistic sensitivity analysis of complex models: a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769, 2004. doi: https://doi.org/10.1111/j.1467-9868.2004.05304.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2004.05304.x. (Cited on page 19.)

Dean S. Oliver, Albert C. Reynolds, and Ning Liu. *Inverse theory for petroleum reservoir characterization and history matching*. 2008. (Cited on page 79.)

Audrey Olivier and Andrew W. Smyth. On the performance of online parameter estimation algorithms in systems with various identifiability properties. *Frontiers in Built Environment*, 3:14, 2017. (Cited on page 67.)

Kasper Zinck Østergaard, Per Brath, and Jakob Stoustrup. Estimation of effective wind speed. In *Journal of Physics: Conference Series*, volume 75, page 012082. The Science of Making Torque from Wind Lyngby, Denmark, 2007. (Cited on page 22.)

Art B. Owen. *Monte Carlo theory, methods and examples*. 2013. (Cited on page 89.)

Art B. Owen. Sobol' indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014. doi: 10.1137/130936233. URL https://doi.org/10.1137/130936233. (Cited on page 18.)

Art B. Owen and Clémentine Prieur. On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, 2017. doi: 10.1137/16M1097717. URL https://doi.org/10.1137/16M1097717. (Cited on page 18.)

Nia E. Owen, Peter Challenor, Prathyush P. Menon, and Samir Bennani. Comparison of surrogate-based uncertainty quantification methods for computationally expensive simulators. *SIAM/ASA J. Uncertainty Quantification*, 5(1):403–435, 2017. doi: 10.1137/15M1046812. (Cited on page 19.)

Jeong-Soo Park. Optimal latin-hypercube designs for computer experiments. *Journal of statistical planning and inference*, 39(1):95–111, 1993. (Cited on page 51.)

Lucia Parussini, Daniele Venturi, Paris Perdikaris, and George E. Karniadakis. Multi-fidelity gaussian process regression for prediction of random fields. *Journal of Computational Physics*, 336:36–50, 2017. (Cited on page 87.)

Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical Review Letters*, 120, 01 2018. doi: 10.1103/PhysRevLett.120.024102. (Cited on pages 107 and 112.)

Paris Perdikaris, Daniele Venturi, Johannes O. Royset, and George Em. Karniadakis. Multi-fidelity modelling via recursive co-kriging and gaussian–markov random fields. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150018, 2015. (Cited on page 87.)

Paris Perdikaris, Daniele Venturi, and George Em. Karniadakis. Multifidelity information fusion algorithms for high-dimensional systems and massive data sets. *SIAM Journal on Scientific Computing*, 38(4):B521–B538, 2016. (Cited on page 87.)

Timothée Perdrizet, Jean-Christophe Gilloteaux, David Teixeira, Gilles Ferrer, Loïc Piriou, Delphine Cadiou, Jean-Michel Heurtier, and Cédric Le Cunff. Fully coupled floating wind turbine simulator based on nonlinear finite element method: Part ii validation results. In *ASME 2013 32nd International Conference on Ocean, Offshore and Arctic Engineering*, pages V008T09A052–V008T09A052. Citeseer, 2013. (Cited on pages 85 and 95.)

Leif E. Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009. (Cited on page 114.)

Giovanni Petrone, Carlo de Nicola, Domenico Quagliarella, Jeroen Witteveen, and Gianluca Iaccarino. Wind turbine performance analysis under uncertainty. In *49th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, page 544, 2011. (Cited on page 85.)

Dinh Tuan Pham, Jacques Verron, and Marie Christine Roubaud. A singular evolutive extended Kalman filter for data assimilation in oceanography. *Journal of Marine systems*, 16(3-4):323–340, 1998. (Cited on page 78.)

Elmar Plischke. An effective algorithm for computing global sensitivity indices (easi). *Reliability Engineering and System Safety*, 95(4):354 – 360, 2010. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2009.11.005. URL http://www.sciencedirect.com/science/article/pii/S0951832009002579. (Cited on page 17.)

Elmar Plischke, Emanuele Borgonovo, and Curtis L. Smith. Global sensitivity measures from given data. *European Journal of Operational Research*, 226(3):536 – 550, 2013. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2012.11.047. URL http://www.sciencedirect.com/science/article/pii/S0377221712008995. (Cited on page 17.)

Wojciech Popko, Fabian Vorpahl, Adam Zuga, Martin Kohlmeier, Jason M. Jonkman, Amy Robertson, Torben J. Larsen, Anders Yde, Kristian Saetertro, Knut M. Okstad, et al. Offshore code comparison collaboration continuation (oc4), phase i-results of coupled simulations of an offshore wind turbine with jacket support structure. Technical

report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2012. (Cited on page 34.)

Clémentine Prieur and Stefano Tarantola. Variance-based sensitivity analysis: Theory and estimation algorithms. *Handbook of Uncertainty Quantification*, pages 1217–1239, 2017. (Cited on page 43.)

Principia. Principia 2019 deeplines wind. URL http://www.principia-group.com/blog/product/deeplines-wind/. (Cited on pages 34 and 95.)

Julian Quick, Jennifer Annoni, Ryan King, Katherine Dykes, Paul Fleming, and Andrew Ning. Optimization under uncertainty for wake steering strategies. In *Journal of Physics: Conference Series*, volume 854, page 012036. IOP Publishing, 2017. (Cited on page 98.)

Patrick Ragan and Lance Manuel. *Statistical Extrapolation Methods for Estimating Wind Turbine Extreme Loads*. 2007. doi: 10.2514/6.2007-1221. URL https://arc.aiaa.org/doi/abs/10.2514/6.2007-1221. (Cited on page 26.)

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003. (Cited on pages 54, 55, 58, 60, and 89.)

Brian J. Reich, Eric Kalendra, Curtis B. Storlie, Howard D. Bondell, and Montserrat Fuentes. Variable selection for high dimensional bayesian density estimation: application to human exposure simulation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(1):47–66, 2012. (Cited on page 64.)

Sebastian Reich. Data assimilation: The schr\" odinger perspective. *arXiv preprint arXiv:1807.08351*, 2018. (Cited on page 79.)

Jennifer M. Rinker. Calculating the sensitivity of wind turbine loads to wind inputs using response surfaces. In *Journal of Physics: Conference Series*, volume 753, page 032057, 2016. (Cited on page 21.)

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science and Business Media, 2004. (Cited on page 13.)

Amy N. Robertson, Jason M. Jonkman, Walter Musial, Fabian Vorpahl, and Wojciech Popko. Offshore code comparison collaboration, continuation: Phase ii results of a floating semisubmersible wind system. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2013. (Cited on page 32.)

Amy N. Robertson, Kelsey Shaler, Latha Sethuraman, and Jason M. Jonkman. Sensitivity analysis of the effect of wind characteristics and turbine properties on wind turbine loads. *Wind Energy Science (Online)*, 4(NREL/JA-5000-74876), 2019a. (Cited on pages 21 and 110.)

Amy N. Robertson, Kelsey Shaler, Latha Sethuraman, and Jason M. Jonkman. Sensitivity of uncertainty in wind characteristics and wind turbine properties on wind turbine extreme and fatigue loads. *Wind Energy Science Discussions*, pages 1–41, 2019b. doi: 10.

5194/wes-2019-2. URL https://www.wind-energ-sci-discuss.net/wes-2019-2/. (Cited on pages 98 and 111.)

Mélanie C. Rochoux, Sophie Ricci, Didier Lucor, Benedicte Cuenot, and Arnaud Trouvé. Towards predictive data-driven simulations of wildfire spread - Part 1: Reduced-cost Ensemble Kalman Filter based on a Polynomial Chaos surrogate model for parameter estimation. *Nat. Hazards and Earth Syst. Sci.*, 14(11):2951–2973, 2014. (Cited on page 18.)

Olivier Roustant, David Ginsbourger, and Yves Deville. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1):1–55, 2012. URL http://www.jstatsoft.org/v51/i01/. (Cited on page 91.)

Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley and Sons, 2016. (Cited on page 13.)

Paul-Baptiste Rubio, François Louf, and Ludovic Chamoin. Fast model updating coupling bayesian inference and pgd model reduction. *Computational Mechanics*, 62(6):1485–1509, 2018. (Cited on page 68.)

Elias D. Nino Ruiz, Adrian Sandu, and Jeffrey Anderson. An efficient implementation of the ensemble kalman filter based on an iterative sherman–morrison formula. *Statistics and Computing*, 25(3):561–577, 2015. (Cited on page 69.)

Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989. (Cited on pages 19, 50, 55, and 58.)

Andrea Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer physics communications*, 145(2):280–297, 2002. (Cited on pages 17, 44, and 89.)

Andrea Saltelli, Terry Andres, and Toshimitsu Homma. Sensitivity analysis of model output: an investigation of new techniques. *Computational statistics and data analysis*, 15(2):211–238, 1993. (Cited on page 44.)

Andrea Saltelli, Karen Chan, and E. Marian Scott, editors. *Sensitivity analysis*. Wiley series in probability and statistics. J. Wiley and sons, New York and Chichester and Weinheim etc., 2000. ISBN 978-0-470-74382-9. (Cited on pages 43, 88, and 119.)

Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. *Sensitivity analysis in practice: a guide to assessing scientific models*, volume 1. Wiley Online Library, 2004. (Cited on pages 14 and 44.)

Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global Sensitivity Analysis. The Primer*. John Wiley and Sons, Ltd, dec 2007. ISBN 9780470725184. doi: 10.1002/9780470725184. (Cited on pages 14, 17, and 18.)

Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. John Wiley and Sons, 2008. (Cited on pages 44 and 50.)

Thomas J. Santner, Brian J. Williams, William I. Notz, and Brain J. Williams. *The design and analysis of computer experiments*, volume 1. Springer, 2003. (Cited on page 10.)

Korn Saranyasoontorn, Lance Manuel, and Paul Veers. On estimation of coherence in inflow turbulence based on field measurements. In *42nd AIAA Aerospace Sciences Meeting and Exhibit*, page 1002, 2004. (Cited on page 110.)

J. Gerard Schepers, J Heijdra, Dimitri Foussekis, S Øye, Robert Rawlinson Smith, M Belessis, Kenneth Thomsen, Torben J. Larsen, I Kraan, B Visser, et al. Verification of european wind turbine design codes, vewtdc; final report. *Netherlands Energy Research Foundation ECN*, 2002. (Cited on page 34.)

Claudia Schillings and Andrew M. Stuart. Analysis of the ensemble kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, 55(3):1264–1290, 2017. (Cited on pages 69 and 84.)

Roland Schöbi, Bruno Sudret, and Joe Wiart. Polynomial-Chaos-based Kriging. *International Journal for Uncertainty Quantification*, 5(2):171–193, feb 2015. (Cited on page 19.)

Lloyd S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2 (28):307–317, 1953. (Cited on page 18.)

Yulin Si. Structural control strategies for load reduction of floating wind turbines. 2015. (Cited on pages xv and 39.)

Eric Simley and Lucy Y. Pao. Evaluation of a wind speed estimator for effective hub-height and shear components. *Wind Energy*, 19(1):167–184, 2016. (Cited on page 22.)

David Simms, Scott Schreck, Maureen Hand, and Lee J. Fingersh. Nrel unsteady aerodynamics experiment in the nasa-ames wind tunnel: a comparison of predictions to measurements. Technical report, National Renewable Energy Lab., Golden, CO (US), 2001. (Cited on pages 98 and 111.)

Timothy W. Simpson, Jesse D. Poplinski, Patrick N. Koch, and Janet K. Allen. Metamodels for computer-based engineering design: survey and recommendations. *Engineering with computers*, 17(2):129–150, 2001. (Cited on page 54.)

Ralph C. Smith. *Uncertainty quantification: theory, implementation, and applications*, volume 12. Siam, 2013. (Cited on pages 10, 19, and 85.)

Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004. (Cited on pages 54 and 55.)

Chris Snyder and Fuqing Zhang. Assimilation of simulated doppler radar observations with an ensemble kalman filter. *Monthly Weather Review*, 131(8), 2003. (Cited on pages 92 and 116.)

Il'ya M. Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990. (Cited on pages 88 and 89.)

Il'ya M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp*, 1(4):407–414, 1993. (Cited on pages 44, 45, 48, 50, 64, 86, and 119.)

Il'ya M. Sobol'. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280, 2001. (Cited on pages 17 and 44.)

Il'ya M. Sobol' and Sergei Kucherenko. Derivative based global sensitivity measures and their link with global sensitivity indices. *Mathematics and Computers in Simulation*, 79(10):3009 – 3017, 2009. ISSN 0378-4754. doi: https://doi.org/10.1016/j.matcom.2009.01.023. URL http://www.sciencedirect.com/science/article/pii/S0378475409000354. (Cited on page 16.)

Christian Soize. *Uncertainty quantification*. Springer, 2017. (Cited on page 10.)

Christian Soize and Roger Ghanem. Physical systems with random uncertainties: chaos representations with arbitrary probability measure. *SIAM Journal on Scientific Computing*, 26(2):395–410, 2004. (Cited on pages 18, 19, 54, and 55.)

Giovanni Solari and Giuseppe Piccardo. Probabilistic 3-d turbulence modeling for gust buffeting of structures. *Probabilistic Engineering Mechanics*, 16(1):73–86, 2001. (Cited on page 110.)

Mohsen N. Soltani, Torben Knudsen, Mikael Svenstrup, Rafael Wisniewski, Per Brath, Romeo Ortega, and Kathryn Johnson. Estimation of rotor effective wind speed: A comparison. *IEEE Transactions on Control Systems Technology*, 21(4):1155–1167, 2013. (Cited on page 22.)

John D. Sørensen and Henrik S. Toft. Probabilistic design of wind turbines. *Energies*, 3 (2):241–257, 2010. (Cited on pages 20 and 85.)

Michael Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143–151, 1987. doi: 10.1080/00401706.1987.10488205. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1987.10488205. (Cited on page 50.)

Michael L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science and Business Media, 2012. (Cited on pages 55 and 89.)

Curtis B. Storlie, Laura P. Swiler, John C. Helton, and Cedric J. Sallaberry. Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliability Engineering and System Safety*, 94(11): 1735–1763, nov 2009. doi: 10.1016/j.ress.2009.05.007. (Cited on page 18.)

Andrew Stuart and Kostas Zygalakis. Data assimilation: A mathematical introduction. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2015. (Cited on page 79.)

Bruno Sudret. Uncertainty propagation and sensitivity analysis in mechanical models–contributions to structural reliability and stochastic spectral methods. *Habilitation à diriger des recherches, Université Blaise Pascal, Clermont-Ferrand, France*, 147, 2007. (Cited on page 10.)

Bruno Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability engineering and system safety*, 93(7):964–979, 2008. doi: 10.1016/j.ress.2007.04.002. (Cited on page 18.)

Bruno Sudret and Armen Der Kiureghian. *Stochastic finite element methods and reliability: a state-of-the-art report*. Department of Civil and Environmental Engineering, University of California, 2000. (Cited on page 12.)

Timothy John Sullivan. *Introduction to uncertainty quantification*, volume 63. Springer, 2015. (Cited on page 10.)

Herbert J. Sutherland. On the fatigue analysis of wind turbines. Technical report, Sandia National Labs., Albuquerque, NM (US); Sandia National Labs , 1999. (Cited on page 96.)

Ronen Talmon, Stéphane Mallat, Hitten Zaveri, and Ronald R. Coifman. Manifold learning for latent variable inference in dynamical systems. *IEEE Transactions on Signal Processing*, 63(15):3843–3856, 2015. (Cited on page 129.)

Pierre Tandeo, Pierre Ailliot, Ronan Fablet, Juan Ruiz, François Rousseau, and Bertrand Chapron. The analog ensemble kalman filter and smoother. 2014. (Cited on pages 115 and 116.)

Pierre Tandeo, Pierre Ailliot, Juan Ruiz, Alexis Hannart, Bertrand Chapron, Anne Cuzol, Valérie Monbet, Robert Easton, and Ronan Fablet. Combining analog method and ensemble data assimilation: application to the lorenz-63 chaotic system. In *Machine learning and data mining approaches to climate science*, pages 3–12. Springer, 2015. (Cited on pages 106, 107, and 112.)

Pierre Tandeo, Pierre Ailliot, Marc Bocquet, Alberto Carrassi, Takemasa Miyoshi, Manuel Pulido, and Yicun Zhen. A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation. *Monthly Weather Review*, 148(10):3973–3994, 2020. (Cited on pages xvi and 77.)

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005. (Cited on pages 19 and 68.)

Stefano Tarantola, Debora Gatelli, and Thierry Alex Mara. Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering and System Safety*, 91(6):717–727, 2006. (Cited on pages 17 and 43.)

Rui Teixeira, Alan O'Connor, Maria Nogal, Nandakumar Krishnan, and James Nichols. Analysis of the design of experiments of offshore wind turbine fatigue reliability design with kriging surfaces. *Procedia Structural Integrity*, 5:951 – 958, 2017. ISSN 2452-3216. doi: https://doi.org/10.1016/j.prostr.2017.07.132. URL http://www.sciencedirect.com/science/article/pii/S2452321617302445. 2nd International Conference on Structural Integrity, ICSI 2017, 4-7 September 2017, Funchal, Madeira, Portugal. (Cited on page 21.)

Marcin Tekieli and Marek Słoński. Application of monte carlo filter for computer vision-based bayesian updating of finite element model. *Mechanics and Control*, 32(4), 2013. (Cited on page 20.)

Jean-Yves Tissot. *Sur la décomposition ANOVA et l'estimation des indices de Sobol'. Application à un modèle d'écosystème marin*. PhD thesis, Grenoble, 2012. (Cited on page 45.)

Jean-Yves Tissot and Clémentine Prieur. A randomized Orthogonal Array-based procedure for the estimation of first- and second-order Sobol' indices. *Journal of Statistical Computation and Simulation*, 85(7):1358–1381, 2015. doi: 10.1080/00949655.2014.971799. URL https://hal.archives-ouvertes.fr/hal-00743964. (Cited on page 17.)

Elham Tofighi, David Schlipf, and Christopher M. Kellett. Nonlinear model predictive controller design for extreme load mitigation in transition operation region in wind turbines. In *2015 IEEE Conference on Control Applications (CCA)*, pages 1167–1172. IEEE, 2015. (Cited on pages xv and 33.)

Henrik S. Toft, Lasse Svenningsen, Wolfgang Moser, John D. Sørensen, and Morten Lybech Thøgersen. Assessment of wind turbine structural integrity using response surface methodology. *Engineering Structures*, 106:471–483, 2016a. (Cited on page 21.)

Henrik S. Toft, Lasse Svenningsen, John D. Sørensen, Wolfgang Moser, and Morten Lybech Thøgersen. Uncertainty in wind climate parameters and their influence on wind turbine fatigue loads. *Renewable Energy*, 90:352–361, 2016b. (Cited on page 21.)

Zoltan Toth. Long-range weather forecasting using an analog approach. *Journal of climate*, 2(6):594–607, 1989. (Cited on page 113.)

Kendra L. Van Buren, Mark G. Mollineaux, François M. Hemez, and Sezer Atamturktur. Simulating the dynamics of wind turbine blades: part ii, model validation and uncertainty quantification. *Wind Energy*, 16(5):741–758, 2013. (Cited on pages 21 and 85.)

Gijs A.M. Van Kuik. The lanchester–betz–joukowsky limit. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, 10(3):289–291, 2007. (Cited on page 33.)

Paul S. Veers. Three-dimensional wind simulation. Technical report, Sandia National Labs., Albuquerque, NM (USA), 1988. (Cited on page 29.)

Herman F. Veldkamp. Chances in wind energy: a probalistic approach to wind turbine fatigue design. 2006. (Cited on pages 21 and 37.)

Herman F. Veldkamp. A probabilistic evaluation of wind turbine fatigue design rules. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, 11(6):655–672, 2008. (Cited on page 21.)

John Von Neumann and Stanislav Ulam. Monte carlo method. *National Bureau of Standards Applied Mathematics Series*, 12(1951):36, 1951. (Cited on page 13.)

Warren E. Walker, Poul Harremoës, Jan Rotmans, Marjolein B.A. Van Der Sluijs, Jeroen P .and Van Asselt, Peter Janssen, and Martin P. Krayer von Krauss. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated assessment*, 4(1):5–17, 2003. (Cited on page 10.)

Gary Wang. Adaptive response surface method using inherited latin hypercube design points. *Transactions of the ASME Journal of Mechanical Design*, 125:210–220, 07 2003. doi: 10.1115/1.1561044. (Cited on page 51.)

Lijuan Wang and Ahsan Kareem. Modeling and simulation of transient winds in downbursts/hurricanes. In *Proceedings of the 10th American Conference on Wind Engineering, Baton Rouge, LA*, 2005. (Cited on page 28.)

Weijun Wang, Stéphane Caro, Fouad Bennis, and Oscar Roberto Salinas Mejia. A simplified morphing blade for horizontal axis wind turbines. *Journal of solar energy engineering*, 136(1), 2014. (Cited on pages xv and 32.)

Yimei Wang, Pierre-Elouan Réthoré, Paul van der Laan, Juan Pablo Murcia Leon, Yongqian Liu, and Licheng Li. Multi-fidelity wake modelling based on co-kriging method. In *Journal of Physics: Conference Series (Online)*, volume 753, page 032065. IOP Publishing, 2016. (Cited on page 85.)

Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, USA, 1995. (Cited on page 92.)

Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967. (Cited on pages xvi and 98.)

Curt Wells. *The Kalman filter in finance*, volume 32. Springer Science and Business Media, 2013. (Cited on pages 73 and 115.)

David Witcher. Uncertainty Quantification Techniques in Wind Turbine, 2017. (Cited on pages 98 and 111.)

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. (Cited on pages 101 and 121.)

Liang Yan and Tao Zhou. Adaptive multi-fidelity polynomial chaos approach to bayesian inference in inverse problems. *Journal of Computational Physics*, 381:110–128, 2019. (Cited on pages 20 and 68.)