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

Adrien Hirvoas, Clémentine Prieur, Élise Arnaud, Fabien Caleyron, Miguel Munoz Zuniga

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

Adrien Hirvoas[1,2], Clémentine Prieur[2], Élise Arnaud[2], Fabien Caleyron[1], and Miguel Munoz Zuniga[3]

[1]IFP Energies nouvelles, Rond-point de l'échangeur de Solaize, BP 3, 69360 Solaize, France
[2]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LJK, 38000 Grenoble, France
* Institute of Engineering Univ. Grenoble Alpes
[3]IFP Energies nouvelles, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France

Correspondence: Adrien Hirvoas (adrien.hirvoas@gmail.com)

Abstract. In this paper, we propose a procedure for quantifying and reducing uncertainties impacting numerical simulations involved in the estimation of the fatigue of a wind turbine structure. The present study generalizes a previous work carried out by the authors proposing to quantify and to reduce uncertainties affecting the properties of a wind turbine model by combining a global sensitivity analysis and a recursive Bayesian filtering approach. We extend the procedure to include the uncertainties involved in the modeling of a synthetic wind field. Unlike the model properties having a static or slow time-variant behavior, the parameters related to the external sollicitation have a non-explicit dynamic behavior which must be taken into account during the recursive inference. A non-parametric data-driven approach to approximate the non-explicit dynamic of the inflow related parameters is used. More precisely, we focus on data assimilation methods combining a nearest neighbor or analog sampler with a stochastic filtering method such as the ensemble Kalman filter. This so-called data-driven data assimilation approach is evaluated on an industrial case of a wind turbine in operation using in situ measurements from an operating structure. The measured data are used by the method to recursively reduce the uncertainties that affect the parameters related to both model properties and wind field.

1 Introduction

A major challenge in wind energy industry is to propose robust designs withstanding unknown environmental conditions. Design standards (IEC, 2019) 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 operating wind turbine lifetime 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 introduced in this paper generalizes the one in (Hirvoas et al.) by taking into account the uncertainties affecting the parameters related to the wind inflow. It relies on a complete framework including a global sensitivity analysis, an identifiability analysis, and a recursive Bayesian inference approach. First, a surrogate based global sensitivity analysis through the estimation of Sobol' indices 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.

For design certification, offshore wind turbines in pilot farms are more and more monitored thanks to a large amount of sensors. In that context, the measured data can be efficiently used in order to learn the non-explicit dynamic behavior of the wind parameters needed for numerical simulation. 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 the principle of the nearest neighbors (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, developed in (Hirvoas et al.), interfacing Python library AnDA[1] combining analog forecasting with ensemble data assimilation. The algorithm we propose takes profit of the parallelization capabilities of 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 2 describes the different uncertainties involved in the framework of this study. In Section 3, the theoretical 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 4.

2 Context

2.1 Uncertainty in wind turbine modeling

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, 2019), 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 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

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

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 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, 2019)

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. 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 having a one-sided power spectral density defined as:

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

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 wind inflow over the swept area is generated based on a grid of points thanks to an exponential spatial coherence method (Jonkman, 2009). The related coherence 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:

$$\mathrm{coh}_{i,j}(f) = \exp\left(-a\left(\frac{\Delta r}{z_m}\right)^{\gamma}\sqrt{\left(\frac{f\,\Delta r}{\overline{u}_m}\right)^2 + \left(\frac{b'\,\Delta r}{L_u}\right)^2}\right), \tag{1}$$

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

Eight input parameters related to the wind field have been identified to be tainted by uncertainties, see Table 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.

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 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 free wind speed can be obtained from the 10-minute data obtained from the anemometer modulo an additive error term. So that the mean free wind speed at hub height can be obtained from the anemometer as:

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

where $\overline{u}_{scada}$ is the 10-minute mean wind speed obtained from the anemometer mounted on the wind turbine nacelle and $\Delta\overline{u}$ is an additive error assumed to follow the distribution defined in Table 2.

In a similar manner, the free wind speed standard deviation can be obtained from the measurement obtained by the 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 anemometer of the wind turbine nacelle and $\Delta\sigma_u$ is an additive error assumed to follow the distribution defined in Table 2.

Unless having high frequency SCADA data, no information can be obtained on the other parameters. An investigation of the distribution of the uncertainty affecting these remaining wind inflow parameters has to be properly made. The vertical wind shear is modeled with the following power law that uses a shear coefficient α:

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

where $\overline{u}$ is the prescribed hub-height mean wind velocity, z is the vertical distance from the ground surface, z_{hub} is the hub height, and α is the vertical wind shear coefficient. We adapt the Gaussian distribution proposed by Dimitrov et al. (2015) for the 10-minute vertical wind shear exponent, such as:

$$\begin{aligned}\mu_\alpha &= 0.088(\ln(\overline{u}_{scada}) - 1)\\ \sigma_\alpha &= 1/\overline{u}_{scada} \qquad .\end{aligned} \tag{2}$$

Table 2 summarizes the wind-inflow parameters that we consider unknown and their respective uncertainty modeling. In particular, we defined the probability distributions of the parameters used in the exponential coherence model defined in Equation (1).

Table 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 (2)	(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: 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 3 gathers information about the probability distribution of each of these paremeters.

2.2 Methodology for uncertainty quantification

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.

The aim of the work in this article 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 turbine model properties in addition to wind inflow uncertainties, respectively denoted by $\mathbf{x}^1$ and $\mathbf{x}^2$ in Figure 1.

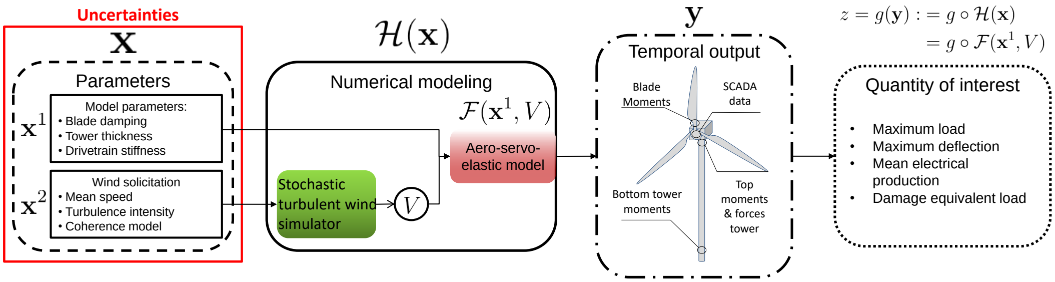

Figure 1. Wind turbine modeling framework, where $\mathbf{x}$ is the extended vector gathering both uncertainties from wind inflow parameters and model properties.

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 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 also proposed to detect non identifiable parameters considering the current measurements.

Table 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_{CMx}	[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 ±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 ±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 $\sim$ ±10% 1 FW
Blade edge wise stiffness	α_{BE}	$[N \cdot m^2]$	$\mathcal{G}$	$\mu = 1.00 \quad \sigma = 3.33e-02$	IFPEN $\sim$ ±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 both the model properties of the wind turbine having a static or slow evolution and the 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.

3 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:

$$\mathbf{x}_{(k)} = \mathcal{M}_{(k-1,k)}(\mathbf{x}_{(k-1)}) + \boldsymbol{\epsilon}^m_{(k)}\,, \tag{3}$$

$$\mathbf{y}_{(k)} = \mathcal{H}_{(k)}(\mathbf{x}_{(k)}, \mathbf{v}_{(k)}) + \boldsymbol{\epsilon}^o_{(k)}\,. \tag{4}$$

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 which depict the wind inflow conditions denoted by $\mathbf{x}^2$ in the next section. 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 $\mathbf{v}_{(k)}$, e.g., mean wind speed obtained from the anemometer of the wind turbine. The sources of errors in the observation model defined in Equation (4) 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.

$$\begin{tikzcd}
\text{Hidden-state} & \cdots \arrow[r] & \mathbf{x}_{(k-1)} \arrow[r, "(\mathcal{M}_{(k-1,k)},\mathbf{Q}_{(k)})"] \arrow[d] & \mathbf{x}_{(k)} \arrow[r] \arrow[d, "(\mathcal{H}_{(k)},\mathbf{v}_{(k)},\mathbf{R}_{(k)})"'] & \mathbf{x}_{(k+1)} \arrow[r] \arrow[d] & \cdots \\
\text{Observations} & & \mathbf{y}_{(k-1)} & \mathbf{y}_{(k)} & \mathbf{y}_{(k+1)} &
\end{tikzcd}$$

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 (3), 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 (3) by a data-driven model relying on an analog forecasting operator, denoted by $\mathcal{A}$, such as :

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

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:

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

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 article, as selected 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:

$$g(\mathbf{u}, \mathbf{v}) = \exp\left(-\lambda \left|\left|\mathbf{u} - \mathbf{v}\right|\right|^2\right), \tag{5}$$

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. The kernel choice is case dependant. The Gaussian kernel used hereafter is isotropic and parameterized allowing to easily control the bandwidth.

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 $\mathcal{I} = \{i_1, \cdots, i_K\}$ 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 such that:

$$\omega_n = \frac{g(\mathbf{x}_{(k-1)}, \mathbf{a}_n)}{\sum_{j=1}^{K} g(\mathbf{x}_{(k-1)}, \mathbf{a}_{i_j})} \cdot$$

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 (5), 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 K 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 not used hereafter, 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 :

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

where the mean forecast $\boldsymbol{\mu}^{\mathrm{LC}}_{(k)} = \sum_{j=1}^{K} \omega_{i_j} \mathbf{s}_{i_j}$ is the weighted mean of the K successors, and $\Sigma^{\mathrm{LC}}_{(k)}$ 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:

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

where the mean forecast is $\boldsymbol{\mu}^{\mathrm{LI}}_{(k)} = \mathbf{x}_{(k-1)} + \sum_{j=1}^{K} \omega_{i_j} (\mathbf{s}_{i_j} - \mathbf{a}_{i_j})$, and $\Sigma^{\mathrm{LI}}_{(k)}$ 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 a slope matrice of size $p \times p$ denoted by $\boldsymbol{\alpha}$, a vector intercept of size $p \times 1$ designated hereafter by $\boldsymbol{\beta}$, and residuals defined as the following vectors $\forall j \in [1, \cdots, K]$, $\mathbf{s}_{i_j} - \big(\boldsymbol{\alpha}\mathbf{a}_{i_j} + \boldsymbol{\beta}\big)$. The Gaussian sampling resorts to:

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

where the mean forecast is $\boldsymbol{\mu}^{\mathrm{LL}}_{(k)} = \boldsymbol{\alpha}\mathbf{x}_{(k-1)} + \boldsymbol{\beta}$, and $\Sigma^{\mathrm{LL}}_{(k)}$ 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 extreme values of the hidden-state based on the available catalog, 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 2. 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 $k-1$.

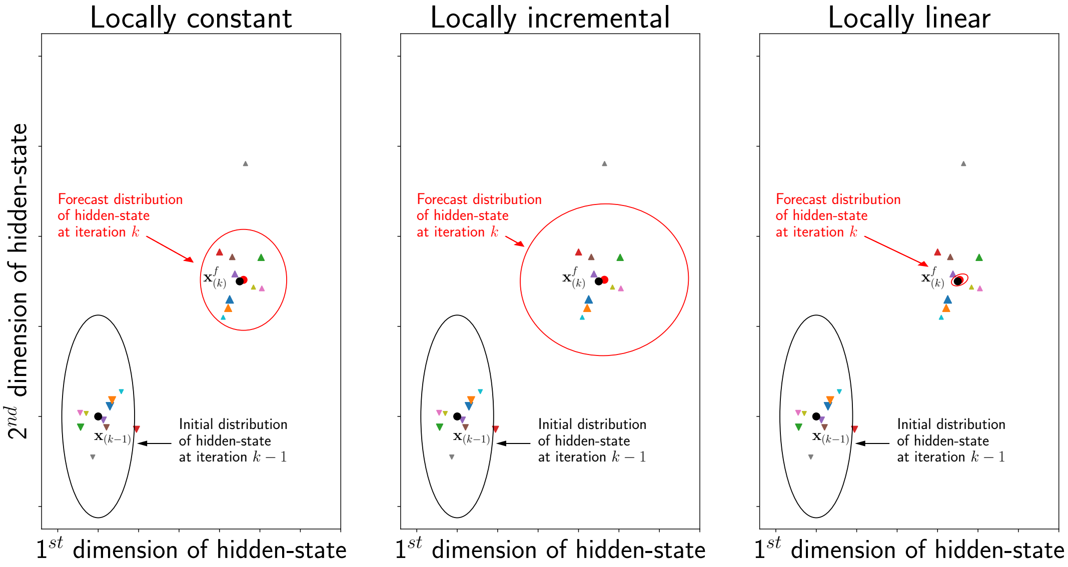

Figure 2. Analog forecasting operator strategies. The real values of the hidden-state $\mathbf{x}_{(k-1)}$ and its forecast $\mathbf{x}_{(k)}$ are represented by full circles. Analogs are displayed in colored down-pointing triangles and successors in up-pointing triangles with their equivalent colors. 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 $p_{\mathbf{X}_{(k)}|\mathbf{Y}_{(1:k)}}(\mathbf{x}_{(k)}|\mathbf{y}_{(1:k)})$ where $\mathbf{Y}_{(1:k)} = [\mathbf{Y}_{(1)},\cdots,\mathbf{Y}_{(k)}]$. 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, the linear 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 applications to take into account the non-linearities 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 principle of the EnKF is to sequentially update all members of the ensemble 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). Due to the fact that this approach is based on an ensemble, it is 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 (3). 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 $\mathbf{x}^{a(i)}_{(k)} = \mathbf{x}^{f(i)}_{(k)} + \mathbf{K}_{(k)} \left(\mathbf{y}^{(i)}_{(k)} - \mathcal{H}_{(k)}(\mathbf{x}^{f(i)}_{(k)}, \mathbf{v}_{(k)})\right)$ where $\mathbf{K}_{(k)} = \mathbf{P}^{f}_{(k)}\, \mathbf{H}^{T}_{(k)} \left(\mathbf{R}_{(k)} + \mathbf{H}_{(k)}\mathbf{P}^{f}_{(k)}\mathbf{H}^{T}_{(k)}\right)^{-1}$ is named as the Kalman Gain with $\mathbf{P}^{f}_{(k)}$ the forecast covariance matrix and $\mathbf{H}_{(k)}$ the observation operator. Due to the nonlinearity of the model $\mathcal{H}_{(k)}$, the terms $\mathbf{P}^{f}_{(k)}\, \mathbf{H}^{T}_{(k)}$ and $\mathbf{H}_{(k)}\mathbf{P}^{f}_{(k)}\mathbf{H}^{T}_{(k)}$ are respectively empirically estimated based on the ensemble members. The ensemble Kalman filter coupled with the analog forecasting strategy is detailed in Algorithm 1.

Algorithm 1 Ensemble Kalman Filter with analog forecast methodology, so-called AnEnKF.

```
*I*nput:
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 $\mathbf{P}_b$.
Initialisation step:
**for** $i = 1$ to $N_{ens}$ **do**
    $\mathbf{x}^{a(i)}_{(0)} = \mathbf{x}_b + \boldsymbol{\epsilon}^b$ with, $\boldsymbol{\epsilon}^b \sim \mathcal{N}(0, \mathbf{P}_b)$
**end for**
**for** $k = 1$ to $T$ **do**
    ***Forecast step:***
    **for** $i = 1$ to $N_{ens}$ **do**
        $\mathbf{x}^{f(i)}_{(k)} = \mathcal{A}_{(k-1,k)}(\mathbf{x}^{f(i)}_{(k-1)}) + \boldsymbol{\epsilon}^{m(i)}_{(k)}$, where,
        **Locally-constant forecasting analog operator:** $\mathcal{A}_{(k-1,k)}(\mathbf{x}^{f(i)}_{(k-1)}) := \boldsymbol{\mu}^{\mathrm{LC}}_{(k)} = \sum_{j=1}^{K} \omega_{i_j} \mathbf{s}_{i_j}$
            and $\boldsymbol{\epsilon}^{m(i)}_{(k)} \sim \Sigma^{\mathrm{LC}}_{(k)}$
        **Locally-incremental forecasting analog operator:** $\mathcal{A}_{(k-1,k)}(\mathbf{x}^{f(i)}_{(k-1)}) := \boldsymbol{\mu}^{\mathrm{LI}}_{(k)} = \mathbf{x}^{a(i)}_{(k-1)} + \sum_{j=1}^{K} \omega_{i_j} (\mathbf{s}_{i_j} - \mathbf{a}_{i_j})$
            and $\boldsymbol{\epsilon}^{m(i)}_{(k)} \sim \Sigma^{\mathrm{LI}}_{(k)}$
        **Locally-linear analog operator:** $\mathcal{A}_{(k-1,k)}(\mathbf{x}^{f(i)}_{(k-1)}) := \boldsymbol{\mu}^{\mathrm{LL}}_{(k)} = \boldsymbol{\alpha}\mathbf{x}_{(k-1)} + \boldsymbol{\beta}$
            and $\boldsymbol{\epsilon}^{m(i)}_{(k)} \sim \Sigma^{\mathrm{LL}}_{(k)}$
        where $(\mathbf{a}_n, \mathbf{s}_n)_{n \in \mathcal{I}}$ (with $\mathcal{I} = \{i_1, \cdots, i_K\}$) are the $K$-pairs of analog and successor for the $i$-th analysis member of the ensemble at iteration $k-1$ and $\mathrm{cov}_\omega$ is the weighted covariance.
    **end for**
    ***Update step:***
    $\mathbf{P}^f_{(k)}\,\mathbf{H}^T_{(k)} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathbf{x}^{f(i)}_{(k)} - \bar{\mathbf{x}}^f_{(k)}\right) \left(\mathcal{H}_{(k)}(\mathbf{x}^{f(i)}_{(k)}, \mathbf{v}_{(k)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}^f_{(k)}, \mathbf{v}_{(k)})\right)^T$
    $\mathbf{H}_{(k)}\mathbf{P}^f_{(k)}\mathbf{H}^T_{(k)} = \frac{1}{N_{ens}-1} \sum_{i=1}^{N_{ens}} \left(\mathcal{H}_{(k)}(\mathbf{x}^{f(i)}_{(k)}, \mathbf{v}_{(k)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}^f_{(k)}, \mathbf{v}_{(k)})\right) \left(\mathcal{H}_{(k)}(\mathbf{x}^{f(i)}_{(k)}) - \mathcal{H}_{(k)}(\bar{\mathbf{x}}^f_{(k)})\right)^T$
    $\mathbf{K}_{(k)} = \mathbf{P}^f_{(k)}\,\mathbf{H}^T_{(k)} \left(\mathbf{R}_{(k)} + \mathbf{H}_{(k)}\mathbf{P}^f_{(k)}\mathbf{H}^T_{(k)}\right)^{-1}$
    **for** $i = 1$ to $N_{ens}$ **do**
        $\mathbf{y}^{(i)}_{(k)} = \mathbf{y}_{(k)} + \mathbf{e}^{o(i)}_{(k)}$ with, $\mathbf{e}^{o(i)}_{(k)} \sim \mathcal{N}(0, \mathbf{R}_{(k)})$
        $\mathbf{x}^{a(i)}_{(k)} = \mathbf{x}^{f(i)}_{(k)} + \mathbf{K}_{(k)} \left(\mathbf{y}^{(i)}_{(k)} - \mathcal{H}_{(k)}(\mathbf{x}^{f(i)}_{(k)}, \mathbf{v}_{(k)})\right)$
    **end for**
**end for**
```

4 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 $1.4\ m/s$.

4.1 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 degree of freedom number 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 4.

Table 4. Reference wind turbine specifications.

Quantity	Value
Number of blades	3
Rated power	2.0 MW
Rotor speed range	8.5 – 17.1 rpm (±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 French 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 level of 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

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

the upstream wind flow conditions. A graphical representation of the monitoring system configuration is proposed in Figure 3. 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 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.

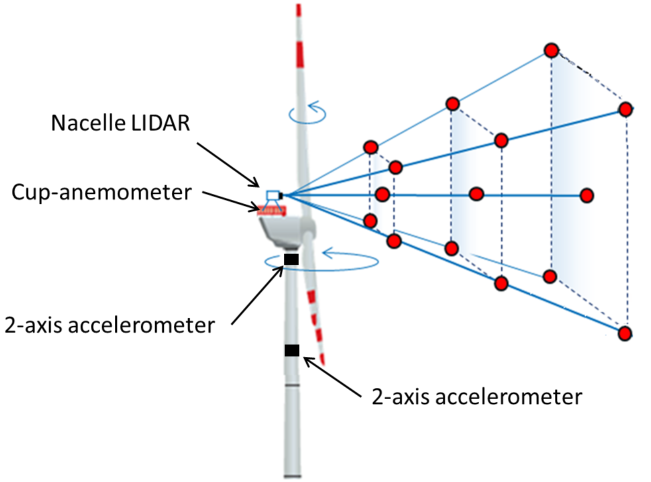

Figure 3. Monitoring system configuration for the reference wind turbine.

4.2 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 has been investigated. We focus our interest on total Sobol' sensitivity indices (Sobol', 1990). 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 to a single value. To alleviate the computational cost in the sensitivity index estimation, heteroscedastic Gaussian process (GP) models (Ginsbourger et al., 2008) are built independently for each Damage Equivalent Load (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 (Damblin et al., 2013). 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 of the DOE, from 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 5. For each output, the total effect Sobol indices are estimated using the corresponding heteroscedastic Gaussian process metamodel, with an estimated predictivity on validation set-over at least 0.8, 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 (Hirvoas et al.).

Table 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 from the conditional distribution of the predictor based on the learning sample. Moreover, the uncertainty due to Monte Carlo integration was estimated with a bootstrap procedure with a sample size of 100, see Efron (1981) for futher details in bootstrapping strategy. The estimated total Sobol' indices for the considered quantities of interest with their corresponding 95% confidence intervals are presented in Figure 4. 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. Eight other parameters describing the wind inflow conditions or the wind turbine model properties have total Sobol' indices higher to the arbitrary threshold (set to $2.5e-02$) 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 4. 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.

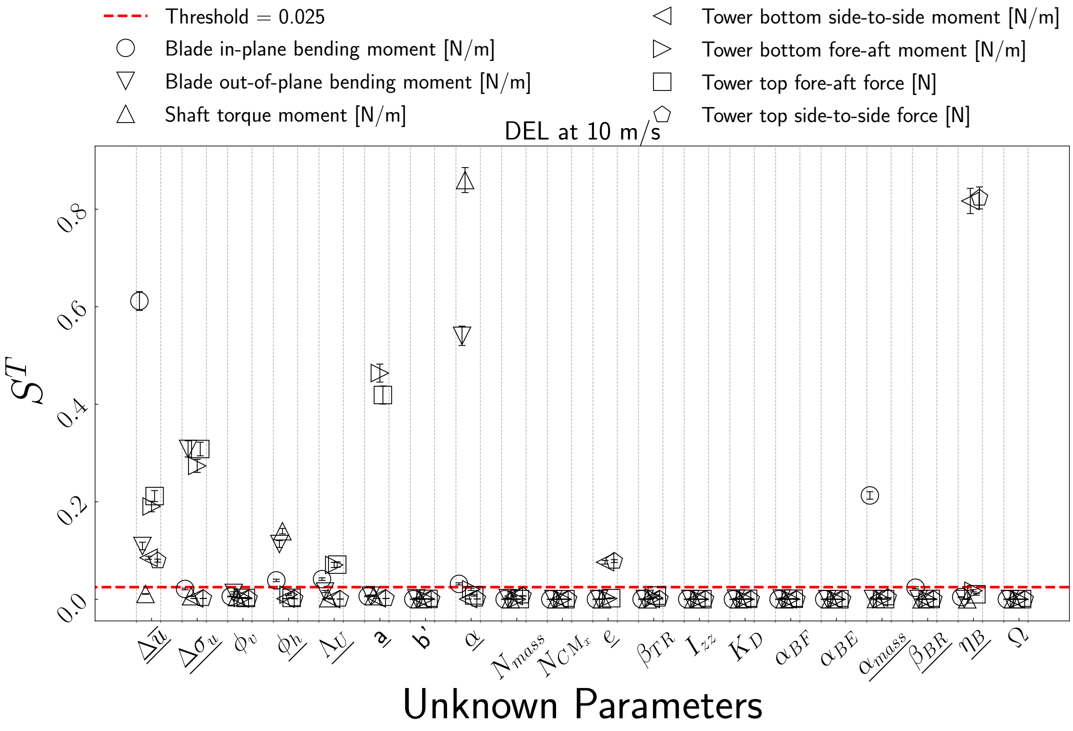

Figure 4. Estimation of total Sobol' indices (y-axis) with their 95% confidence interval corresponding to each of the 20 parameters (x-axis) for the different fatigue loads. The dashed line corresponds to a threshold arbitrarily chosen to 5e-2. Confidence intervals (CI) are obtained by taking into account 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.

4.3 Identifiability study

A major issue for parameter estimation problem is the identifiability. In this context, Dobre et al. (2012) highlights 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 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 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 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]

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.

Table 7. Total Sobol' and aggregated 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

4.4 Recursive inference strategy based on AnEnKF approach

The current in situ wind data availability or quality from the LIDAR system 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 8. Their corresponding reference variable in the augmented state vector is also specified.

Table 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 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:

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

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 8, $\mathcal{A}_{(k-1,k)}$ is the analog forecasting operator as detailed in Section 3, $\mathbf{v}_{(k)}$ 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 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 pretreatment 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 5 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.

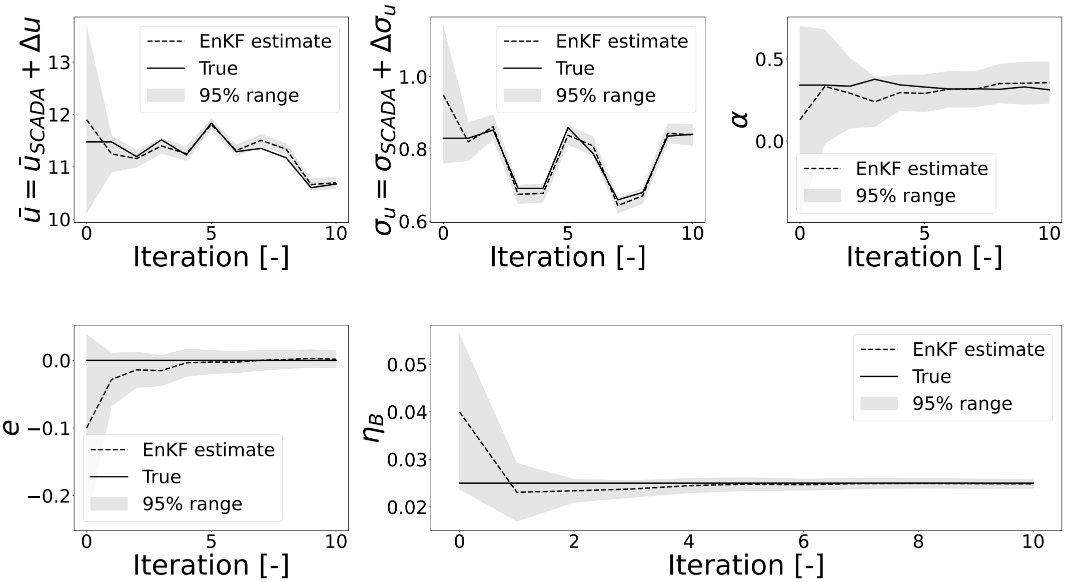

Figure 5. Iteration evolution of the posteriori estimates of the input parameters. Results obtained by running the AnEnKF procedure with $N = 500$ members of the ensemble used for the estimation and considering pseudo-experimental numerical observations.

5 Conclusions

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.

Acknowledgements. The author would like to thank the framework agreement between IFP Energies Nouvelles (IFPEN) and the French Institute for Research in Computer Science and Automation (INRIA) which has funded the present work. Moreover, the author gratefully thanks all researchers who directly or indirectly contributed to the achievement of the work mentioned in this paper.

References

Aanonsen, S. I., Nævdal, G., Oliver, D. S., Reynolds, A. C., Vallès, B., et al.: The ensemble Kalman filter in reservoir engineering–a review, Spe Journal, 14, 393–412, 2009.

Alexander, R., Zhao, Z., Székely, E., and Giannakis, D.: Kernel analog forecasting of tropical intraseasonal oscillations, Journal of the Atmospheric Sciences, 74, 1321–1342, 2017.

Ayet, A. and Tandeo, P.: Nowcasting solar irradiance using an analog method and geostationary satellite images, Solar Energy, 164, 301–315, 2018.

Bertino, L., Evensen, G., and Wackernagel, H.: Sequential data assimilation techniques in oceanography, International Statistical Review, 71, 223–241, 2003.

Bocquet, M., Farchi, A., and Malartic, Q.: Online learning of both state and dynamics using ensemble Kalman filters, arXiv preprint arXiv:2006.03859, 2020.

Brown, S. D.: The Kalman filter in analytical chemistry, Analytica chimica acta, 181, 1–26, 1986.

Brunton, S. L., Proctor, J. L., and Kutz, J. N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proceedings of the national academy of sciences, 113, 3932–3937, 2016.

Chau, T. T. T., Ailliot, P., and Monbet, V.: An algorithm for non-parametric estimation in state–space models, Computational Statistics & Data Analysis, 153, 107 062, 2021.

Chen, L.: Curse of Dimensionality, pp. 545–546, Springer US, Boston, MA, https://doi.org/10.1007/978-0-387-39940-9_133, 2009.

Damblin, G., Couplet, M., and Iooss, B.: Numerical studies of space-filling designs: optimization of Latin Hypercube Samples and subprojection properties, Journal of Simulation, 7, 276–289, 2013.

Dimitrov, N., Natarajan, A., and Kelly, M.: Model of wind shear conditional on turbulence and its impact on wind turbine loads, Wind Energy, 18, 1917–1931, 2015.

Dimitrov, N., Natarajan, A., and Mann, J.: Effects of normal and extreme turbulence spectral parameters on wind turbine loads, Renewable Energy, 101, 1180–1193, 2017.

Dobre, S., Bastogne, T., Profeta, C., Barberi-Heyob, M., and Richard, A.: Limits of variance-based sensitivity analysis for non-identifiability testing in high dimensional dynamic models, Automatica, 48, 2740–2749, 2012.

Durbin, J. and Koopman, S. J.: Time series analysis by state space methods, Oxford university press, 2012.

Efron, B.: Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods, Biometrika, 68, 589–599, 1981.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, Journal of Geophysical Research: Oceans, 99, 10 143–10 162, 1994.

Evensen, G.: The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean dynamics, 53, 343–367, 2003.

Evensen, G.: Data assimilation: the ensemble Kalman filter, Springer Science and Business Media, 2009.

Fablet, R., Viet, P., Lguensat, R., and Chapron, B.: Data-driven assimilation of irregularly-sampled image time series, in: 2017 IEEE International Conference on Image Processing (ICIP), pp. 4302–4306, IEEE, 2017.

Friedman, J. H.: On bias, variance, 0/1-loss, and the curse-of-dimensionality, Data mining and knowledge discovery, 1, 55–77, 1997.

Ginsbourger, D., Picheny, V., Roustant, O., and Richet, Y.: 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), 2008.

Hamilton, F., Berry, T., and Sauer, T.: Ensemble Kalman filtering without a model, Physical Review X, 6, 011 021, 2016.

Harvey, A. C.: Forecasting, structural time series models and the Kalman filter, Cambridge university press, 1990.

Haykin, S.: Kalman filtering and neural networks, vol. 47, John Wiley & Sons, 2004.

Hirvoas, A., Prieur, C., Arnaud, E., Caleyron, F., and Munoz Zuniga, M.: 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, https://doi.org/https://doi.org/10.1002/nme.6630.

Holierhoek, J., Korterink, H., van de Pieterman, R., Rademakers, L., and Lekou, D.: Recommended Practices for Measuring in Situ the ‘Loads’ on Drive Train, Pitch System and Yaw System., Energy Research Center of the Netherlands (ECN), 2010.

Houtekamer, P. L. and Mitchell, H. L.: A sequential ensemble Kalman filter for atmospheric data assimilation, Monthly Weather Review, 129, 123–137, 2001.

Houtekamer, P. L., He, B., and Mitchell, H. L.: Parallel implementation of an ensemble Kalman filter, Monthly Weather Review, 142, 1163–1182, 2014.

IEC, I. E. C.: IEC 61400-1: 2019: Wind energy generation systems-Part 1: Design requirements, 2019.

Iooss, B., Janon, A., Pujol, G., with contributions from Baptiste Broto, Boumhaout, K., Veiga, S. D., Delage, T., Fruth, J., Gilquin, L., Guillaume, J., Le Gratiet, L., Lemaitre, P., Nelson, B. L., Monari, F., Oomen, R., Rakovec, O., Ramos, B., Roustant, O., Song, E., Staum, J., Sueur, R., Touati, T., and Weber, F.: sensitivity: Global Sensitivity Analysis of Model Outputs, https://CRAN.R-project.org/package=sensitivity, r package version 1.16.0, 2019.

Jansen, M. J.: Analysis of variance designs for model output, Computer Physics Communications, 117, 35–43, 1999.

Jonkman, B. J.: TurbSim user’s guide: Version 1.50, Tech. rep., National Renewable Energy Lab (NREL), Golden, CO, USA, 2009.

Jonkman, J. M. ., Buhl Jr., M. L., et al.: FAST user’s guide, Golden, CO: National Renewable Energy Laboratory, 365, 366, 2005.

Julier, S. J. and Uhlmann, J. K.: New extension of the Kalman filter to nonlinear systems, in: Signal processing, sensor fusion, and target recognition VI, vol. 3068, pp. 182–193, International Society for Optics and Photonics, 1997.

Kalman, R. E.: A new approach to linear filtering and prediction problems, 1960.

Koukoura, C.: Validated Loads Prediction Models for Offshore Wind Turbines for Enhanced Component Reliability, Ph.D. thesis, Technical University of Denmark, 2014.

Lamboni, M., Monod, H., and Makowski, D.: Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models, Reliability Engineering and System Safety, 96, 450–459, https://doi.org/10.1016/j.ress.2010.12.002, 2011.

Lguensat, R., Tandeo, P., Ailliot, P., Pulido, M., and Fablet, R.: The analog data assimilation, Monthly Weather Review, 145, 4093–4107, 2017.

Lorenz, E. N.: Atmospheric predictability as revealed by naturally occurring analogues, Journal of the Atmospheric sciences, 26, 636–646, 1969.

Pathak, J., Hunt, B., Girvan, M., Lu, Z., and Ott, E.: Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach, Physical Review Letters, 120, https://doi.org/10.1103/PhysRevLett.120.024102, 2018.

Peterson, L. E.: K-nearest neighbor, Scholarpedia, 4, 1883, 2009.

Robertson, A. N., Shaler, K., Sethuraman, L., and Jonkman, J. M.: Sensitivity analysis of the effect of wind characteristics and turbine properties on wind turbine loads, Wind Energy Science (Online), 4, 2019a.

Robertson, A. N., Shaler, K., Sethuraman, L., and Jonkman, J. M.: Sensitivity of Uncertainty in Wind Characteristics and Wind Turbine Properties on Wind Turbine Extreme and Fatigue Loads, Wind Energy Science Discussions, pp. 1–41, https://doi.org/10.5194/wes-2019-2, 2019b.

Saranyasoontorn, K., Manuel, L., and Veers, P.: On estimation of coherence in inflow turbulence based on field measurements, in: 42nd AIAA Aerospace Sciences Meeting and Exhibit, p. 1002, 2004.

Simms, D., Schreck, S., Hand, M., and Fingersh, L. J.: NREL unsteady aerodynamics experiment in the NASA-Ames wind tunnel: a comparison of predictions to measurements, Tech. rep., National Renewable Energy Lab., Golden, CO (US), 2001.

Snyder, C. and Zhang, F.: Assimilation of simulated Doppler radar observations with an ensemble Kalman filter., Monthly Weather Review, 131, 2003.

Sobol', I. M.: On sensitivity estimation for nonlinear mathematical models, Matematicheskoe modelirovanie, 2, 112–118, 1990.

Solari, G. and Piccardo, G.: Probabilistic 3-D turbulence modeling for gust buffeting of structures, Probabilistic Engineering Mechanics, 16, 73–86, 2001.

Tandeo, P., Ailliot, P., Fablet, R., Ruiz, J., Rousseau, F., and Chapron, B.: The analog ensemble kalman filter and smoother, 2014.

Tandeo, P., Ailliot, P., Ruiz, J., Hannart, A., Chapron, B., Cuzol, A., Monbet, V., Easton, R., and Fablet, R.: Combining analog method and ensemble data assimilation: application to the Lorenz-63 chaotic system, in: Machine learning and data mining approaches to climate science, pp. 3–12, Springer, 2015.

Toth, Z.: Long-range weather forecasting using an analog approach, Journal of climate, 2, 594–607, 1989.

Wells, C.: The Kalman filter in finance, vol. 32, Springer Science and Business Media, 2013.

Witcher, D.: Uncertainty Quantification Techniques in Wind Turbine, 2017.

Wold, S., Esbensen, K., and Geladi, P.: Principal component analysis, Chemometrics and intelligent laboratory systems, 2, 37–52, 1987.