Journal of Physics:
Conference Series

PURPOSE-LED
PUBLISHING™

PAPER • OPEN ACCESS

A method to reduce model uncertainty by fusing the output from multiple stochastic simulators

To cite this article: Imad Abdallah *et al* 2020 *J. Phys.: Conf. Ser.* **1618** 042019

View the article online for updates and enhancements.

You may also like

- Domain-randomised instance-segmentation benchmark for soot in PIV images
Basil Jose, Klaus Peter Geigle and Fabian Hampp
- Uncertainty quantification of grain morphology in laser direct metal deposition
Paromita Nath, Zhen Hu and Sankaran Mahadevan
- Functional derivatives for uncertainty quantification and error estimation and reduction via optimal high-fidelity simulations
Alejandro Strachan, Sankaran Mahadevan, Vadiraj Hombal et al.

This content was downloaded from IP address 173.177.195.252 on 26/02/2026 at 16:13

A method to reduce model uncertainty by fusing the output from multiple stochastic simulators

Imad Abdallah, Konstantinos Tatsis, Charilaos Mylonas, Vasilis Ntertimanis and Eleni Chatzi

Structural Mechanics & Monitoring, ETH Zürich, Zürich, Switzerland

E-mail: abdallah@ibk.baug.ethz.ch

Abstract. In virtualizing engineered systems, it is essential to come up with simulators that are essentially capable of representing the system in its "as-deployed" state. Any attempt to this end may only be approximate given the inherent uncertainties present in the loadings and operational conditions of the system, as well as the configuration of the system itself (geometry, materials, control systems, boundary conditions, etc.). This is especially true for complex systems, such as wind turbines, where often a number of assumptions govern the setup of the engineering models. Such models are often made available at different granularities with each one offering a diversified level of precision depending on the quantity of interest (e.g. macroscopic displacements or microscopic strains) and the properties of the acting loads (e.g. amplitude and frequency content). This implies that the predictive capabilities are severely hampered when a single so-deemed best model is chosen for simulation. Building on this idea, we here present a method for fusing the outputs from multiple simulators (e.g. aero-servo-hydro-elastic simulators) for estimating a quantity of interest (QoI) with higher precision. The proposed ensemble learning approach comprises two main building blocks. Firstly, a clustering step by means of a Variational Bayesian Gaussian mixture model, employed for the weighing of each available simulator. Clustering is performed on the basis of the binned input space, which allows for extraction of a probability map for each local region of the binned input space. This delivers an adaptive scheme, which allows different simulators to more or less prominently contribute to the prediction of the QoI, depending on the range of the input parameters. Local weighted Bootstrap Aggregation is then executed in a second step for combining the clustered ensemble of outputs from the individual simulators. A simulated toy example and a wind turbine blade fatigue case study are herein exploited to demonstrate the efficacy of the suggested ensemble learning scheme. The approach is compared against alternatives typically adopted in existing literature, such as Stacking, classical Bagging, and Bayesian Model Averaging. The results confirm an improvement in predictive capabilities as expressed via the reduction in the generalization error and the narrowing of the associated confidence interval.

1. Introduction

The predictions from stochastic numerical simulators used in the design of machines and structures are subject to uncertainties. There are several sources which contribute to these uncertainties: physical (input), parametric, statistical, measurement and model-form uncertainties. Model-form uncertainty refers to the uncertainty in the mathematical and computational structure of the simulator itself and the assumptions behind it. The uncertainty due to model-form, is not usually quantified and sometimes it is not even acknowledged. Some reasons for this are the absence of consensus on how to quantify and represent model


Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

uncertainty, and lack of data to do proper benchmarking [1]. However, when making critical decisions concerning the design and safety of the mechanical and structural components of wind turbines, there is a great need to deal with model uncertainty. If the model uncertainty is not taken into account, the resulting decision may be erroneous or sub-optimal. In many instances, stochastic predictions from multiple simulators for the same quantity of interest are available. In this paper, the term "simulator" is used to encompass any numerical computer model simulating a physical phenomena. In the wind energy domain, aero-servo-hydro-elastic simulator-to-simulator comparisons have been performed with relevant results presented in reports and publications, for instance [2, 3]. Often, the simulators deployed in these comparison and benchmarking activities have been calibrated and validated via experimentation, and have been subsequently adopted by manufacturers in the design and deployment of wind turbines. The problem arises, however, when one is unable to definitively prove that a single simulator is ubiquitously superior to the others across the full range of acting environmental and operational conditions and for the various materials and components making up the various sub-structures of the wind turbines. We contextualize the problem by offering an example where the stochastic blade root bending moments of a wind turbine are predicted via several distinct aero-servo-elastic simulators, which implement multiple solution methodologies to solve the same problem. For instance, the structural dynamics in one simulator may implement a multi-body/Finite Element formulation, and another simulator may implement a classical modal-superposition approach. The aerodynamics in one simulator may implement a Blade Element Moment theory, whereas the actuator line based computational fluid dynamics (CFD) or a blade resolved CFD formulation may be adopted in another [4]. Hence, whether one is performing reliability-based design optimizations or calibrating safety factors, or calculating design loads, etc., having the means to evaluate the model-form uncertainty and possibly decreasing it becomes crucial. One popular approach in the wind energy field is based on the multi-fidelity sampling or surrogate modelling approaches [5, 6, 7]. Another approach approach lies in fusion of the predicted outputs from alternate simulators, thus exploiting their individual predictive traits, as opposed to seeking a single optimal-fit option [8]. The idea is akin to model averaging; an approach often adopted within the context of statistical inference. However, typical multi-fidelity or averaging or aggregation schemes, as reported in existing literature, do not make distinction on the influencing region of the input space, or might further assume one of the models in the ensemble as reference, i.e., the data-generating model. Realizing that the range of inputs is a salient parameter to what concerns modeling of wind turbine systems, we here present an unsupervised ensemble learning method, which treats individual simulator outputs as components contributing to an ensemble estimate, as opposed to competing predictions. The proposed scheme succeeds in addressing both the epistemic uncertainty owed to lack of training data, as well as the inherently present model form uncertainty. We call our method, originally introduced in recent work of the authors [9], *Unsupervised Local Cluster-Weighted Bootstrap Aggregation* (**UnLoCWeB**).

2. Method

The method exploits the probabilistic output form a diversified set of individual simulators (engineering models of different granularity). It proceeds to firstly extract the local weights of these simulators via an unsupervised variational Bayesian Gaussian mixture (VBGM) scheme, followed by estimation of a weighted ensemble, delivered by Bootstrap Aggregation (or weighted Bagging).

We initiate with discretization of the input space domain into a finite number of bins ($\Delta \boldsymbol{x}$). Each bin, or rather the data contained therein, is further segregated into a training and validation set. The output from the simulators is clustered via use of VBGM to the training

data from each bin:

$$p\left(\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{\mathcal{Y}} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{1}$$

where K is the unknown number of clusters, each having a mixing coefficient π_k and parameters of Gaussian distributions: mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. The posterior distribution of π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are computed using Variational Bayesian inference. An adaptive solution (VBGM) is selected since the number of clusters needs to be adaptively defined, and may not be considered as fixed a priori. Furthermore the application of the clustering procedure in bins of the input space, as opposed to the full range, favors flexibility and allows a possible varying weighing of each simulator, conditioned on the acting range of the influencing inputs. As an added safety, we include an intermediate step, which examines the stability of the cluster on perturbed versions of the data included within each bin.

The probability map (weights) of the simulators' output is then extracted as the joint probability $P\left(\mathcal{S}_i, C_j \mid \Delta\boldsymbol{x}\right)$ of a cluster $\{C_j\}$ and a simulator $\{\mathcal{S}_i\}$:

$$P\left(\mathcal{S}_i, C_j \mid \Delta\boldsymbol{x}\right) = P\left(S_i \mid C_j, \Delta\boldsymbol{x}\right) \cdot P\left(C_j \mid \Delta\boldsymbol{x}\right) \tag{2}$$

where $P\left(C_j \mid \Delta\boldsymbol{x}\right) = {}^{N_{C_j|\Delta\boldsymbol{x}}}/_{N}$, $N_{C_j|\Delta\boldsymbol{x}}$ denotes the number of local data points contained in cluster C_j and N corresponds to the number of available output samples (from all simulators). $P\left(\mathcal{S}_i \mid C_j, \Delta\boldsymbol{x}\right) = \frac{N_{S_i|C_j,\Delta\boldsymbol{x}}}{N_{C_j|\Delta\boldsymbol{x}}}$ where, $N_{S_i|C_j,\Delta\boldsymbol{x}}$ is the number of data points in cluster C_j corresponding to simulator $\mathcal{S}_i$.

Local cluster-weighted Bootstrap Aggregating (Bagging) follows next [10], to fuse the ensemble of outputs from each bin - based on the previously extracted weights - into one single aggregated predictor. The bagging algorithm is comprised of the following steps:

(i) construct a weighted bootstrap sample $\left(\boldsymbol{x}_*^{(1)}, \mathcal{Y}_*^{(1)}\right), \ldots, \left(\boldsymbol{x}_*^{(n_{\Delta\boldsymbol{x}})}, \mathcal{Y}_*^{(n_{\Delta\boldsymbol{x}})}\right)$ by randomly drawing $n_{\Delta\boldsymbol{x}}$ times with replacement from the local clustered data,

(ii) compute the bootstrapped estimator $h_n(\cdot)$: $\hat{g}_* = h_n\left(\left(\boldsymbol{x}_*^{(1)}, \mathcal{Y}_*^{(1)}\right), \ldots, \left(\boldsymbol{x}_*^{(n_{\Delta\boldsymbol{x}})}, \mathcal{Y}_*^{(n_{\Delta\boldsymbol{x}})}\right)\right)$; $h_n(\cdot)$ defines an estimator as a function of the data,

(iii) repeat steps 1 and 2 M times to infer $\left\{\hat{g}_*^k, k = 1, \ldots, M\right\}$. The final local cluster-weighted bagged estimator is obtained as: $\hat{g}_{w,bagg} = \frac{\sum_{k=1}^{M} \hat{g}_*^k}{M}$.

3. Analytical computational experiment

We present a synthetic analytical example to motivate the approach in the context of response prediction for wind energy infrastructure, and demonstrate the benefits and limitations. Assume that a target QoI (blade or tower response) is prescribed by the following analytical expression:

$$y_{target}(x) = (6x - 2)^2 \sin(12x - 4) + 12 \tag{3}$$

We further make an assumption on the QoI being normally distributed:

$$P\left(\mathcal{Y}_{target} \mid X\right) \sim \mathcal{N}\left(y_{target}(x), COV = 0.05\right) \tag{4}$$

The coefficient of variation COV is assumed to be homoscedastic, while the random variable X is uniformly distributed $X \sim U(0,1)$. Five alternative simulators are assumed available for predicting $\mathcal{Y}_{target}$. For the purpose of this illustrative example it is assumed that all simulators may be attributed the same level of confidence, which is often the case in absence of experimental results or field measurements, serving for validation. To further highlight the

workings of the proposed scheme, we assume all simulators to be in mutual disagreement and to further deviate from the target response (as illustrated in Figure 1). The form of each simulator is listed below

$$\begin{aligned}
y_1(x) &= y_{target}(x) - 0.1; P\left(\mathcal{Y}_1 \mid X\right) \sim \mathcal{N}\left(y_1(x), COV = 0.05\right) \\
y_2(x) &= y_{target}(x) + 1; P\left(\mathcal{Y}_2 \mid X\right) \sim \mathcal{N}\left(y_2(x), COV = 0.05\right) \\
y_3(x) &= A_3 y_{target}(x^2) + B_3\left(x^3 - 0.5\right) + C_3; P\left(\mathcal{Y}_3 \mid X\right) \sim \mathcal{N}\left(y_3(x), COV = 0.05\right) \\
y_4(x) &= A_4 y_{target}(x) + B_4\left(x - 0.5\right) + C_4; P\left(\mathcal{Y}_4 \mid X\right) \sim \mathcal{N}\left(y_4(x), COV = 0.05\right) \\
y_5(x) &= A_5 y_{target}\left(\sqrt{x} + B_5\right) + C_5; P\left(\mathcal{Y}_5 \mid X\right) \sim \mathcal{N}\left(y_5(x), COV = 0.05\right)
\end{aligned} \tag{5}$$

The model coefficients are selected so as to induce the desired mutual disagreement, and are here defined as $\{A_3 = 0.5, B_3 = 8, C_3 = 2\}$, $\{A_4 = 0.5, B_4 = 10, C_4 = 5\}$, and $\{A_5 = 0.25, B_5 = 0.5, C_5 = 12\}$. The set of available simulators suffers different levels of approximation in terms of the overall response behavior (linear versus non-linear), the calculated expected value (systematic versus random bias), and even the position of extrema of the response with respect to the value of the inputs. As is further evident, as the generating model, $\mathcal{Y}_{target}$, is not included in the candidate model set, this example pertains to an $\mathcal{M}$-open setting. The goal lies in predicting the stochastic response conditional on Δx, upon availability of the individual outputs of the five stochastic simulators: $\mathbb{E}\left[\mathcal{Y}_{target} \mid \Delta x, \mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3, \mathcal{Y}_4, \mathcal{Y}_5\right]$. The results are compared against the commonly adopted alternative of classical bagging, in terms of the generalization error (GE), and the breadth of the 95% empirical bootstrap confidence interval (CI). The GE, which comprises a global metric of the estimation precision, is obtained as:

$$GE = \frac{\sum_{i=1}^{n_v}\left[\mathcal{Y}^{(i)} - \hat{g}_{w,bagg}\left(\boldsymbol{x}^{(i)}\right)\right]^2}{\sum_{i=1}^{n_v}\left[\mathcal{Y}^{(i)} - \mu_{\mathcal{Y}}\right]^2}, \quad \mu_{\mathcal{Y}} = \frac{1}{n}\sum_{i=1}^{n_v}\mathcal{Y}^{(i)} \tag{6}$$

where n_v denotes the validation set dimension, and $\mu_{\mathcal{Y}}$ corresponds to the mean of the computed simulator response for the validation set. The local predictive capability is reflected is the computed 95% CI. In order to assess the performance of the proposed scheme, simulations are varied with respect to the number of stochastic samples per simulator and the assigned coefficient of variation (COV). Weighted Bagging is executed for 100 Bootstraps, with the outcome averaged across 50 repetitions, to ensure independence from a specific split of the sample space. Figure 2 indicates improvement of the prediction for the proposed approach, captured in $\mathcal{Y}$, versus classical Bagging. Specifically, reductions of 39% and 7% are obtained for the the generalization error (Figure 3) and the 95% confidence interval (Figure 4), averaged across the binned input space.

A further assessment is offered with respect to the variation of the output COV. Assignment of a higher output dispersion is expected to harden the discrimination of unique clusters. Indeed an increasing output COV, according to Figures 3-4 result in deterioration in the average reduction in the GE and CI of our proposed scheme versus a classical Bagging approach. This is particularly true for a $COV \geq 10\%$. This is tackled by increasing the number of utilized samples to 5000 (per simulator) which results in improvement of the proposed approach versus classical Bagging. This behavior is explained by the fact that when clustering with a mixture of Gaussian distributions, the majority of the data is assumed to lie close to the center of the mixture component. A higher number of samples enables the discrimination of unique clusters, and therefore more unique weights.

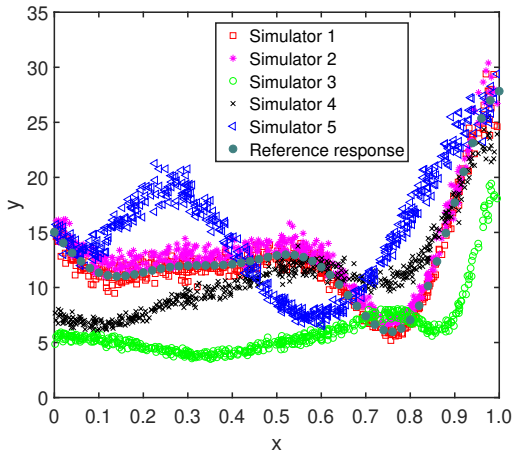


Figure 1: Illustration of the stochastic output y from five simulators.

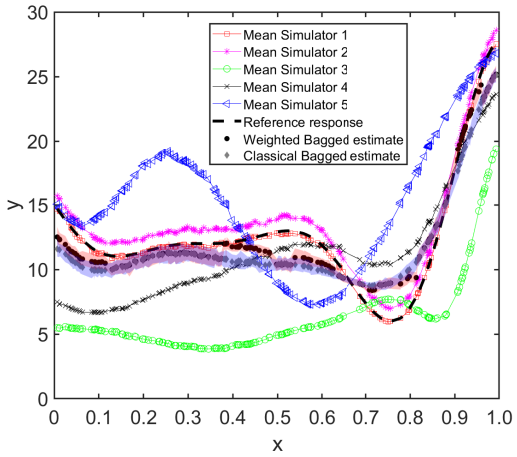


Figure 2: UnLoCWeB versus classical Bagging estimator.

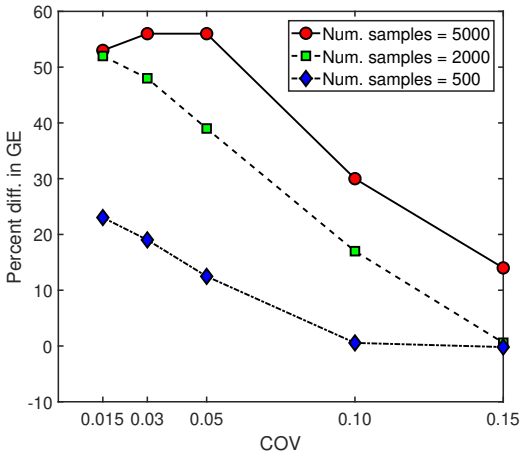


Figure 3: Generalization error: comparing UnLoCWeB to classical Bagging for varying number of samples and COV.

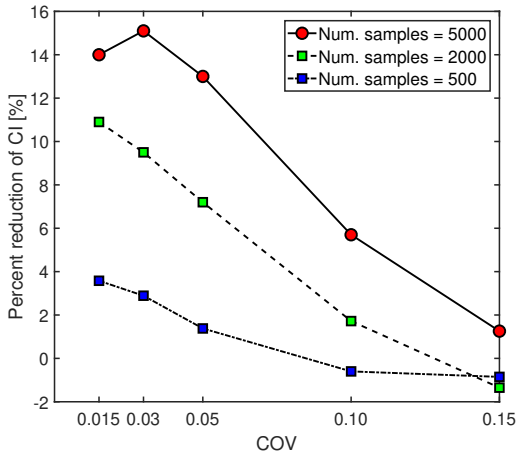


Figure 4: 95% confidence interval: comparing UnLoCWeB to classical Bagging for varying number of samples and COV.

3.1. Comparison to other methods

For the comparison to other methods, we modify the previous example and use 4 simulators given as:

$$
\begin{aligned}
y_1(x) &= A_1 y_{target}(x) + B_1\,(x-0.5) + C_1; P\left(\mathcal{Y}_1 \mid X\right) \sim \mathcal{N}\left(y_1(x), COV = 0.05\right) \\
y_2(x) &= A_2 y_{target}(x^2) + B_2\,(x^3-0.5) + C_2; P\left(\mathcal{Y}_2 \mid X\right) \sim \mathcal{N}\left(y_2(x), COV = 0.05\right) \\
y_3(x) &= A_3 y_{target}(x) + B_3\,(x-0.5) + C_3; P\left(\mathcal{Y}_3 \mid X\right) \sim \mathcal{N}\left(y_3(x), COV = 0.05\right) \\
y_4(x) &= A_4 y_{target}\left(\sqrt{x} + B_4\right) + C_4; P\left(\mathcal{Y}_4 \mid X\right) \sim \mathcal{N}\left(y_4(x), COV = 0.05\right)
\end{aligned}
\tag{7}
$$

The constants are now set to $\{A_1 = 1, B_1 = 10, C_1 = 2\}$, $\{A_2 = 0.5, B_2 = 8, C_2 = 2\}$, and $\{A_3 = 0.5, B_3 = 10, C_3 = 5\}$, and $\{A_4 = 0.25, B_4 = 0.5, C_4 = 12\}$.

In further providing a comparison to the state of the art, we contrast the predictions delivered by our proposed UnLoCWeB scheme to (i) Bayesian Model Averaging, and (ii) Stacking of predictive distributions (Stacking). The BMA and Stacking methods are implemented according to the work of [11] using *PyMC3* [12]. The simulations are conducted with 500 samples per each of the 4 simulators, a discretized input bin size $\Delta x = 0.02$, and $COV = 0.05$. The target is once again the expected value of the response under availability of outputs from all individual simulators. The stacking weights are determined equal to 0.88 for simulator 1 and 0.12 for simulator 3. The BMA select a single optimal simulator via assignment equal to 1 for simulator 1. The generalization errors for each method are listed in Table 1. Simulator 1 is in general not an adequate simulator for the target QoI, but performs better than the individual remaining simulators. In the BMA setting this leads to a higher marginal likelihood assignment when compared against the further candidates. This result pertains across the whole input space. The Stacking scheme further assigns significant weight to simulator 3, yielding an improved generalization error of 0.602 (as opposed to 0.723 for the BMA). To the contrary, the adaptive scheme proposed in the UnLoCWeB approach, achieved by the local assignment of weights with respect to the input space, delivers a lower generalization error of 0.511, and thereby a higher precision amounting to a 15% reduction, as compared against Stacking.

Table 1: Comparing UnLoCWeB to Stacking and BMA: Generalization Error (50 repeats, 100 Bootstraps).

Method	GE
UnLoCWeB	0.511
Stacking	0.602
BMA	0.723

4. Practical Case Study: Wind turbine blade fatigue

The relevance of this approach for engineering simulations in the context of wind infrastructures is demonstrated on evaluation of the Fatigue Damage Equivalent Load (DEL) for a wind turbine blade. A 30 m long wind turbine blade is simulated by means of 10 alternative Finite Element (FE) simulators, of different assumptions and granularity, as summarized in Table 2. In terms of acting loads, the turbulent wind field (Figure 5) is obtained via use of an exponential coherence model and the Kaimal turbulence auto-spectrum following Annex B in [13]. A quasi-static Blade Element Momentum model is used to calculate the normal F_N and tangential F_T aerodynamic forces along the span of the blade (Figure 6).

For the structural modeling, a common engineering approximation is adopted. The blade is approximated as an equivalent clamped-free tapered beam, which represents the main load bearing spar caps and shear-web configuration along the blade span. The target QoI in this case is the time varying blade root in-plane bending moment M_X, which is induced as a result of the action of the nodal forces F_N and F_T along the span of the blade. Ten finite element simulators are used as candidate models for reproducing this QoI. The configurational characteristics of the FEM simulators are overviewed in Table 2, with a further illustration for three characteristic such simulators included in Figure 7. 500 times series of F_N and F_T, are generated as inputs. Each time series is of 600 seconds in duration, at a time step of 0.01 *sec*, and the forces are

assumed to be exerted onto the cross-sectional aerodynamic center.

$$DEL = \left(\frac{1}{N_{eq}} \sum_i n_i \left(\Delta M_{X,i} \right)^m \right)^{1/m} \tag{8}$$

where n_i designates the number of load cycles with range $\Delta M_{X,i}$, i denotes the fatigue cycle index, and N_{eq} specifies the equivalent number of load cycles, typically 10^7 cycles. A Whöler exponent $m = 10$ is assumed, as typical for a composite structure.

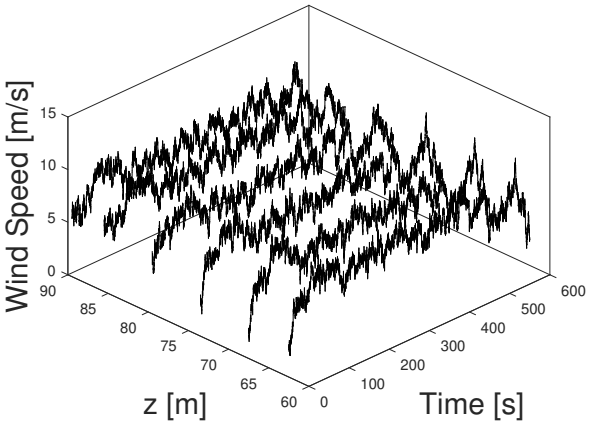


Figure 5: Example of the free stream turbulent wind speed at various blade radial positions.

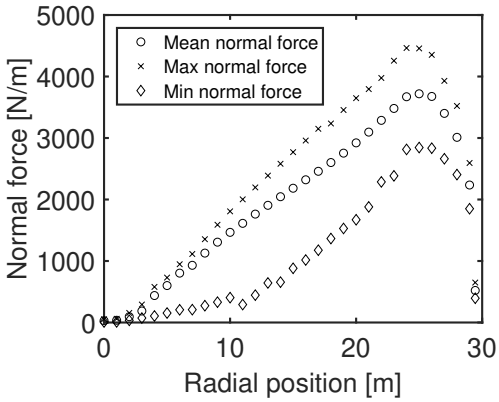


Figure 6: Example of the normal force F_N along the span of the blade.

Table 2: Configuration of the FE-based simulators of the blade.

Simulator	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Dimensions	1D	1D	3D	3D	2D	2D	3D	3D	3D	3D
Element type	Euler-Bernoulli Beam	Euler-Bernoulli Beam	Euler-Bernoulli Beam	Timoshenko Beam	Plane-Stress (2D Elas.)	Plane-Stress (2D Elas.)	Solid (3D Elas.)	Solid (3D Elas.)	Solid (3D Elas.)	Solid (3D Elas.)
Shape function	Linear	Linear	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Quadratic	Quadratic
Modes	2	4	4	8	4	8	4	8	10	12
#elements	16	48	48	48	16	48	2x2x96	4x4x192	8x8x384	16x16x768
Load nodes	6	14	10	14	6	14	6	10	14	28
**Nodal force orientation	In-plane projection	In-plane projection	According to γ	According to γ	In-plane projection	In-plane projection	In-plane projection	According to γ	According to γ	According to γ
Torsion DOF	off	off	on	on	off	off	off	off	on	on

**"In-plane projection": the resultant of the forces is projected onto the $YZ-plane$. **"According to γ": the forces are oriented according to the angle γ, which depends on the relative wind speed, the blade pitch and twist angles.

Figure 8 reports on the computed stochastic DEL for each of the 10 FE-based simulators, as well as on the UnLoCWeB-computed DEL, as a function of wind speed. The divergence among individual simulators appears insignificant below 15 m/s. This is due to the nature of the target

QoI, i.e., the damage equivalent fatigue of the blade root bending moment. For low amplitudes, the linear nature of the response results in the simulators exhibiting similar predictions. This observation justifies the nature of our proposed adaptive calculation scheme, which employs binning to isolate regions of effective discrepancy between candidate simulators, ignoring a priori information on their nominal precision. According to Figure 8, the UnLoCWeB predictor produces an ensemble estimate, which is characterized by a lower variance, with respect to the dispersion exhibited by the individual simulators. This indicates lowering of the uncertainty in the estimation of the short-term fatigue DEL. The clustering step effectively delivers a map of probabilities, in the form of the joint probability $P\left(\mathcal{S}_i, C_j \mid \Delta \boldsymbol{x}\right)$ of a cluster C_j and a simulator S_i. The exploitation of these probabilities in the bootstrapping step results in a shift of the ensemble predictor towards those clusters which comprise a higher density. This observation is verified when observing the estimation result at wind speeds above 15 m/s, where simulators 5 and 6 produce drifting estimates, in comparison to the main cluster.

In a final investigation, we explore the effect of removal of the highest fidelity simulator 10 from the candidate set. As reported in Figure 9, the UnLoCWeB estimation is hardly affected. This implies that the proposed scheme ensures a robust estimation, which is maintained even in absence of the higher precision (and most computationally taxing) simulator. The computational time for simulator 10 is approximately one order of magnitude higher than the simpler simulators. In this respect, we remind the reader that this trait is conditional to the examined QoI. A reliability or failure analysis focusing on extreme responses, or non-linear effects could instead require an alternate weighting of the higher precision simulator.

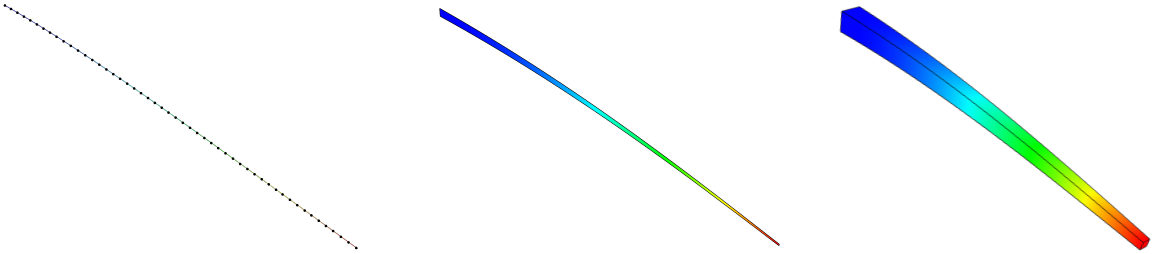


Figure 7: Illustration of the FEM simulators: 3D beam model (simulator 4), FEM formulation of 2D elasticity, i.e. plane stress/strain (simulator 6), and FEM formulation of 3D elasticity, i.e. 3D Elastic Solids (simulator 10).

5. Concluding Discussion

We proposed an ensemble learning framework, termed UnLoCWeB, which relies on unsupervised variational Bayesian Gaussian mixture clustering and local weighted bootstrap aggregation of outputs from individual simulators.

- Clustering is used to infer a probability map, serving for weighing of the simulators.
- An adaptive scheme is proposed with clustering executed with respect to bins of the input space.
- In contrast to existing methods, this scheme assigns weights to clusters of the predicted output, within each bin of the input space, not to the individual simulators.
- A cluster-weighting bootstrap aggregation is then carried out for each bin of the input space for rendering an ensemble prediction.

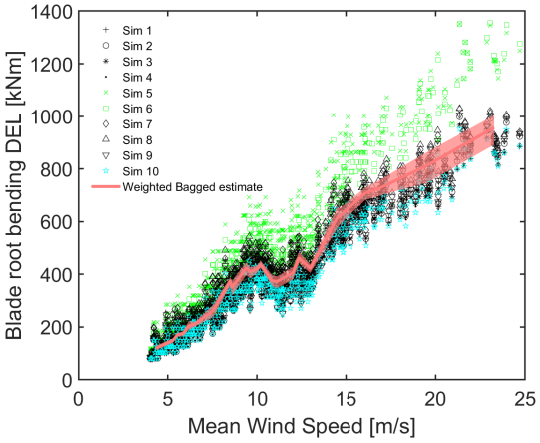


Figure 8: Comparing the DEL of the blade root bending from 10 FEM simulators, and the local cluster-weighted Bagged predictor with the 95% CI.

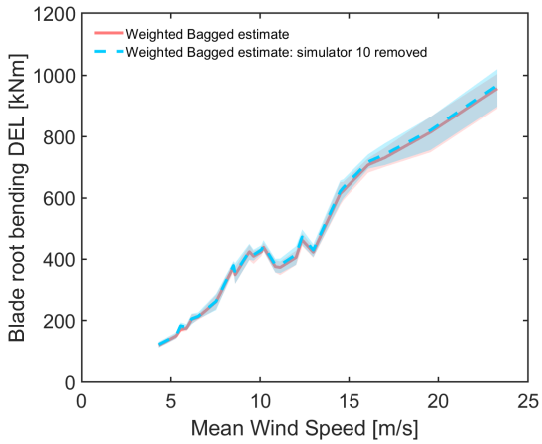


Figure 9: Comparing the local cluster-weighted Bagged predictor when the highest fidelity simulator 10 is removed from the ensemble.

The framework is here illustrated in the context of predictions for engineering systems where response variability is noted across the range of operational loads, and where the assumed properties of the engineering models can influence response predictions. Wind turbine structures comprise such challenging systems due to the high loads these are exposed to, and the numerous simulators at hand (even in standard practice) for simulating these systems. Both toy examples, and a more applied case study is offered. In both cases an improvement in performance and an overall reduction in the generalization error and the 95% confidence interval are noted.

The performance of the proposed UnLoCWeB scheme is contrasted against available alternatives, namely Bayesian Model Averaging (BMA) and Stacking of predictive distributions. It is observed that BMA tends to favor a single simulator, across the whole input space. Stacking results in assigning further weight to a second additional simulator, which improves prediction with respect to the BMA, yet proves sub-par when compared against the UnLoCWeB scheme. The latter succeeds in reducing the generalization error by an additional 15% compared to Stacking. This is particularly promising for applications where available simulators tend to under-perform in different regions of the input space.

A challenging issue with the proposed method is when the correct simulator output is clustered separately (as an outlier) from the larger density clusters of the remaining simulators for certain regions of the input space. In other words, the derived weights may be inadequate because the method fails to recognize that certain simulators are more fitting than others in certain regions of the input space. We envisage three possibilities to tackle this problem: (1) include strong priors on the weights of the individual simulators. (2) verify if the so-deemed correct simulator is a systematic outlier across all regions of the input space and across several QoI. (3) a possible insight is that measurements (when they become available) may be considered as yet another physical simulator(s) in the mix with other numerical simulators, with the proposed algorithm simply re-run without any additional mathematical manipulations. We argue that such a way of performing data assimilation is critical with measurements (preferably with more than one replication), in order to update ambiguous clustering weights.

Finally, follow-up work will explore the case where the QoI follows an extreme or skewed

distribution, where the Gaussian Mixture clustering assumption will need to be revisited.

Acknowledgments
The authors acknowledge the support of the European Research Council via the ERC Starting Grant WINDMIL (ERC-2015-StG #679843) on the topic of Smart Monitoring, Inspection and Life-Cycle Assessment of Wind Turbines. The authors would further like to acknowledge the support the ERC Proof-of-Concept Grant ERC-2018-PoC WINDMIL RT-DT (funded under H2020-EU.1.1., grant agreement ID 825833).

References

[1] Pulkkinen I and Huovinen T 1996 Model uncertainty in safety assessment Tech. Rep. STUK-YTO-TR 95 Finnish Nuclear Center for Radiation and Nuclear Safety

[2] Popko W, Vorpahl F, Zuga A, Kohlmeier M, Jonkman J *et al.* 2014 *Journal of Ocean and Wind Energy* **1** 1–11

[3] Zierath J, Rachholz R and Woernle C 2016 *Wind Energy* **19** 1201–1222

[4] Ehrich S, Schwarz C M, Rahimi H, Stoevesandt B and Peinke J 2018 *Applied Sciences* **8**

[5] Eldred M S, Ng L W T, Barone M F and Domino S P 2017 *Multifidelity Uncertainty Quantification Using Spectral Stochastic Discrepancy Models* (Springer International Publishing) pp 991–1036

[6] Abdallah I, Lataniotis C and Sudret B 2018 *Probabilistic Engineering Mechanics*

[7] Quick J A, Hamlington P E, King R N and Sprague M A 2019 Multifidelity uncertainty quantification with applications in wind turbine aerodynamics Tech. Rep. NREL/CP-5000-72974 National Renewable Energy Laboratory

[8] Ranjan R and Gneiting T 2010 *J. Royal Stat. Soc., Series B* **72** 71–91

[9] Abdallah I, Tatsis K and Chatzi E 2020 *Reliability Engineering System Safety* **199**

[10] Breiman L 1996 *Machine Learning* **24** 123–140

[11] Yao Y, Vehtari A, Simpson D and Gelman A 2018 *Bayesian Analysis* **13** 917–1007

[12] Salvatier J, Wiecki T V and Fonnesbeck C 2016 *PeerJ Computer Science* **2**

[13] 2005 Wind Turbines, Part 1 Design Requirements Tech. Rep. IEC 61400-1:2005(E) International Electrotechnical Commission