The authors sincerely thank both reviewers for their very thorough reading of the manuscript and the insightful comments. We believe that based on the reviewers' inputs our article has significantly improved in quality. We have tried to address all the reviewer suggestions, please find our response in the table below.

Reviewer comment	Response by authors	Changes in manuscript
Reviewer 1		
The paper for this reviewer is too comprehensive and it is suggested that the focus/scope of the paper is revised. The paper is too lenghty and can and should be shortened in order to allow a clear delivery of the key messages. Generally, due to the large amount of theory explained, it could be an idea to assume the baseline theory known to the reader (e.g. PCE, Kriging, in particular Bootstrapping, and sensitivity indices). then it is possible to focus more on the differences between different modeling approaches	This paper includes use and comparison of a number of methods, and the details can be relevant; thus we shorten the article to be clearer, by using an appendix to collect relevant theory and details.	The structure of the paper has been modified to shorten the main body and have a better logical structure. Theory for PCE, Kriging and sensitivity analysis has been moved to Appendix A.
The introduction should include a broader overview of what has been done in the field, especially wind energy. The motivation of the chosen procedures in this paper could be more clear, i.e. how they add to and are distinct from previous research. More publications in this direction are (e.g.).....	Thanks to the reviewer for the very relevant additions to the reference list. Especially the recent publications from Muller et al. and Teixeira et al. help to enrich the state-of-the-art discussion. We have included the suggested publications in the reference list and have discussed them in the introduction.	We have included the suggested publications in the reference list and have discussed them in the introduction.
Little information is given on the direct comparison of the models and their baseline data. A comprehensive overview in form of a table is strongly recommended. Clear overviews of the procedures are necessary. written form is not enough. Also, not sufficient information is given on the number of used samples for the different models. The baseline data must be clear to proof that a fair comparison between models is performed.	Indeed, some clarifications were missing both in terms of the procedure followed, and in the model comparison, where the latter was also pointed out by another reviewer. We have added a schematic explanation of the procedure (Figure 1, accompanied by text in Section 2.1), and have added a table in section 6.2 listing the training set and evaluation set sizes for each model, as well as model execution time.	Figure 1 and a paragraph in Section 2.1 are added. Table in section 6.2 is added.

The section and description of importance sampling may need revision.	Some of the details about the way we implement Importance Sampling (the way we handle the problem with choosing optimal sampling points) were missing in the original manuscript. We also agree with the reviewer's other comment that this is a non-optimal way of using Importance Sampling. We have now indicated this in the text and have added explanations.	Text in Section 4.1 has been modified.
A clear focus should be given to the shortcomings of some of the applied methods. E.g. the use of rosenblatt transformation implies discretized jpdf's which may lead to nonconverged results if the grid is too coarse. Then, the mentioned shortcomings need to be addressed (i.e. no information is given on the applied resolution!). Similar for Kriging / PCE: it is mentioned that Kriging is computationally more expensive, but not how much more time (CPU hrs) was required in this study. Again, this is one of the key performance indicators and should be implied in overall comparison.	The Rosenblatt transformation does not require discretizing of the joint pdfs. Instead, we use a cascade of continuous conditional dependencies, where the distribution parameters of dependent variables are continuous functions of the distribution parameters of other variables. This is described in Section 6.1 (updated structure): *"The conditional dependencies are described in terms of functional relationships between the governing variable and the distribution parameters of the dependent variable, e.g. the mean and standard deviation of the turbulence are modelled as linearly dependent on the wind speed as recommended by the IEC 61400-1 standard, while the mean wind shear is dependent on the mean wind speed and on the turbulence, as defined by (Kelly et al., 2014).* We agree that the computational performance of the models is an important aspect. Therefore a table was added which showed the actual evaluation speeds for a specific example.	The following text was added to Section 6.2: *"Another important aspect to consider when comparing the performance of the surrogate models is the model execution speed, and whether there is a tradeoff between speed and accuracy. A comparison of the model evaluation times for the site-specific lifetime load computation for site 0 is given in Table X. Noticeably the Kriging model requires significantly longer execution time than other approaches, which is mainly due to the requirement of populating a cross-correlation matrix."*
It seems that all approaches provide valuable estimates from this overview study. The conclusion, that one performed "better" than another lacks the presentation of more detailed investigation, which is understood to be beyond the scope of this work.	Yes, more detailed conditional investigation is beyond the scope of the current article. Here we have made compared methods in a basic way, within the context of making a usable database.	-

Page 1, line 20: this list could be more complete. especially in offshore and floating there has been some research very similar to what you are presenting. looking more closely for e.g. a recent study by teixera (2017) on the analysis of offshore structures using kriging surfaces is available, polynomial chaos was applied for fatigue load calculations of blade loading by Ganesh in 2012, etc. if not already done later, it should be clarified in the introduction why the authors chose the particular models analyzed in this study. how was previous research considered in this study? what are shortfalls? what is still missing? what are good practices?	Thanks to the reviewer for the relevant studies. We have now included a short discussion of them in the introduction, and have outlined the differences in scope with our paper.	We have now included a short discussion of them in the introduction, and have outlined the differences in scope with our paper.
Page 2, line 14: which is the considered system in this work?	On page 1, the comment concerns the general design process of any wind turbine system. In our particular calculations, we consider the DTU10MW reference wind turbine, which is an open-source research platform and as such provides good opportunities for reproducibility and comparisons. This is mentioned in Section 2.5 of the manuscript (updated structure).	Clearer mention in § 2.5, with updated structure.
Page 2, line 19: the assumption here is 10min wind fields, correct? otherwise a more broad definition of the wind climate would have to be taken into account	Correct, we assume 10min wind fields. A clarification is added to the text.	The following text was added to the manuscript: "All the quantities referred to above are considered in terms of 10-minute average values."
Page 2, line 20: vertical wind profile modeled in this study by the mean wind shear exponent	Yes we use the power-law exponent α, as stated in the text.	—
Page 3, line 18: better to provide a table	There is a table (Table 1) shown on the next page. A reference to Table 1 is now also made on page 3.	A reference to Table 1 is added on page 3.
Page 3, line 22: wind field for consistency	Changed	Changed
Page 3, line 27: why is it most convenient to apply a Rosenblatt transformation?	The Rosenblatt transformation allows more complex conditional dependencies than the Nataf transformation which implies linear correlation.	This is now mentioned in the text
Page 4, line 1: leave out description of Rosenblatt in order to save space. this is a very short explanation and needs to be clear to the reader to understand it (hence no additional information) if the reader is not aware the procedure can easily be obtained from literature	We prefer to leave the Rosenblatt transformation in the manuscript, because based on later comments from the reviewers some additional explanations were added, which need reference back to the Rosenblatt transformation.	–

Page 4, table 1: this table should fully describe the environmental model that is used as basis for the lifetime fatigue calculation.	The environmental model should in principle be site-specific and is thus not necessarily relevant for inclusion in this table. Table 1 gives all relationships necessary to construct the reference database, but is not intended as a way for showing the site-specific environmental model. Instead, this is now done in a new table (Table 6).	Table 6 has been added to the manuscript
Page 4, table 1: please also indicate the resolution of each variable and its probability function used for the rosenblatt transformation, as well as the applied hierarchy	The applied hierarchy is already defined just after the definition of the Rosenblatt transformation, and it follows the order used in Table 1. This is the text used: *"For the currently considered set of variables, the Rosenblatt transformation can be applied in the order defined in Table 1 - i.e., the wind speed is considered independent of other variables, the turbulence is dependent on the wind speed, the wind shear is conditional on both wind speed and turbulence, etc.".* As already described in the earlier comments, there is no need to give resolution numbers for each variable as the conditional dependencies are modelled as continuous functions.	–
Page 4, table 1: above 3m/s is stated for U	This is a typo, we've used 4m/s as lower limit throughout the paper	3m/s is changed to 4m/s on page 4
Page 5, line 9: this chapter starts out with the right motivation but basically only describes the sampling procedure used, which is only covered superficially. => rephrase chapter.	The section name is changed to "Sampling procedure"	Changed section title to "Sampling procedure"
Page 5, line 14: i.e. surrogate models / response surfaces	The suggested text was added to the manuscript	Added suggested text to the manuscript.
Page 6, figure1: use same format for all points	We have decided to remove Figure 1 as it did not contribute sufficiently to the story.	Removed Fig. 1
Page 6, line 1: not clear how this is different from point 2)	Indeed this bullet-point was confusing and we have removed it.	Removed this bullet-point.

Page 6, line 7: what are the disadvantages of quasi-random numbers and what is the implication for this study?	A disadvantage of the quasi-random sequences is that their properties typically deteriorate in high-dimensional problems, where periodicity and correlation between points in different dimensions may appear. However, such behaviour typically occurs when more than 20-25 dimensions are used. In the present problem the dimensionality is limited by the computational requirements of the load mapping models and the aeroelastic simulations used to train them. Therefore the behaviour of quasi-random sequences in high dimensions does not have implications for the present study.	This explanation is added to the manuscript.
Page 6, line 7: why halton and not sobol, which is much more typical in literature?	The Sobol sequence is characterized with some grouping of point locations in higher dimensions. The Halton sequence does not show such grouping, but, on the other hand, has quite regular (i.e. not sufficiently random) behaviour in high dimensions, so there is a tradeoff in properties. We initially tried Halton, Sobol and Hammersley sequences and found very little effect on the results. We think the choice of a specific pseudorandom sequence is beyond the scope of this paper and have simply chosen one of three possibilities which work equally well for the present problem.	
Page 6, line 8: what is the difference between the three?	Since we don't use any Latin Hypercube designs in the study, we removed Figure 1 and have deleted the sentence referring to it.	Removed Fig.1 and associated reference.
Page 6, line 8: which implementation was used of the sequence? direct sequence? any postprocessing of the points applied? it is important to be able to let the reader reproduce the quasi random series as they may not be well distributed in high dimensions.	The Halton sequence was applied as a direct sequence taking all points consequentially, but discarding the first point in the sequence as this point contains zeros in all dimensions and is associated with zero joint probability. This information is now added to the manuscript.	Added explanation about discarding first point.
Page 6, line 10: what about LHS? even of interest? then it may as well be left out entirely	Indeed, all references to LHS were removed.	Removed references to LHS.

Page 6, line 14: there are more studies on comparing crude monte carlo to quasi random sequences. in these studies high dimensionality relates to dimensions much higher than what is used here. please highlight this when indicating that quasi-random sequences may not be optimal for the current problem an option of this could be to apply a different set of quasi-random numbers on the obained model and perform a convergence study that fits the problem	As discussed above, the number of dimensions is limited by the computational requirements for the models, and not by the properties of the quasi-random series, so we haven't experienced any specific issues with the use of quasi-random series. This is now made clearer and we have added a note that the high dimensionality where issues could appear is typically above 20.	Clarified issue regarding computational requirements vs. quasi-random series type; noted limit for onset of related issues.
Page 7, figure 2: the distribution of the samples seems probability weighted for wind shear as well, not uniform as indicated in the description. is this related to the wind distribution? can the procedure on this be described?	The shear distribution is uniform, however the uniform interval bounds are conditional on the wind speed and turbulence, which gives the impression that the shear is probability-weighted. This is clarified in the caption of Figure 2	Following text was added to the caption of Figure 2: *Solid lines show the sampling space bounds which are curved due to conditional dependencies.*
Page 7, line 1: this is the reference data set?	This is the data set used for model training.	Following was added to the text: *A large-scale generic load database is generated in order to serve as a training data set for the load mapping functions.*
Page 7, line 1: except wind speed and wind shear	Correct, the wind speed is not uniformly distributed. The wind shear though is uniformly distributed within the conditional bounds. A new bulletpoint is added to clarify this	New text: *The physical values of the stochastic variables for all quasi-MC samples are obtained by applying a Rosenblatt transformation using the conditional distribution bounds given in Table 1 and using uniform distribution density, except for the wind speed for which a Beta distribution is used.*
Page 7, line 3: i assume different wind seeds? what about run-in time?	Yes by varying sample points the wind speed is also varied from cut-in to cut-out. The run-in time was 200s, which is excluded from the output time series. This is now indicated in the text.	Included info about run-in time.
Page 7, line 4: please indicate for which parameters this is the case	It's the Mann model turbulence parameters (L, Γ, $\alpha\epsilon^{2/3}$) which determine the turbulence intensity (this is added to the manuscript)	Re-introduced Mann-model & turbulence aspect into paper.
Page 7, line 9: this information should be given in abstract and introduction	–	Information was added both in the abstract and in the introduction.

Page 7, line 9: please explain how HAWC2 is considered high-fidelity. spontaneously i would assume something CFD-based as high-fidelity.	Hawc2 is a nonlinear, dynamic, finite element-based load calculation tool providing high-frequency load time series. Indeed it does not use high-fidelity atmospheric representations, but its load output can be considered high-fidelity due to the time dependency which is absent in the surrogate model approaches.	–
Page 8, line 1: have you used the mean DEL of the 8 1 hour seeds or another value?	We have used the mean DEL from the 48 10-minute periods obtained by splitting the 1h periods into 6 parts. In order to avoid confusions, we changed some text on this page to refer to 10-minute periods instead of 1h.	changed some text on this page to refer to 10-minute periods instead of 1h.
Page 9, line 1: not clear the motivation of this chapter at this point of the paper.	This chapter was moved together with other load-mapping approaches to form chapter 4 in the revised paper.	moved chapter along with other load-mapping approaches to form new chapter 4
Page 9, line 2: section could be left out for brevity	Some of the theory was taken out of the main body of the paper which hopefully should help to improve the readability; however for the sake of completeness we would like to maintain at least small explanations of the basic concepts we use.	Removed some theoretical parts
Page 9, line 2: which	Figure 2 shows the distributions of the first 6 variables	–
Page 9, line 19: i dont understand what is the difference here. the xi can come also from pseudo-MC sampling?	The idea was that applying the IS weights directly on the high-fidelity database points would require using more points to get a converged result compared to directly running a MC/IS simulation with the target distribution. Nevertheless this paragraph is left out of the revised paper for brevity.	Removed paragraph
Page 9, line 20: the database for the base-line data here is based on uniform & importance sampling (wind speed, wind shear)! as i understand importance sampling assumes that the sampling is already based on the occurence probability of the independent variables. hence, a different data base would have to be defined for this comparison (may be extracted from the surrogate/response surface/simplified model). the weighting as described in 7 then adjusts for bias in the created samples.	Here we use a non-standard approach to IS, with the idea that since we have generated a large number of uniformly distributed points for our high-fidelity database, some of these points will also have high density in the site-specific (target) distribution. So we compute the target distribution weights for all points in the database and pick thouse with highest weights as our IS sample. This is now described in the manuscript.	Added description of our IS distribution-weights computation
Page 10, figure 3: this is based on a surrogate model or raw data?	This is based on raw data. We have now indicated that in the text when referring to the figure.	Reference to figure now indicates raw data.

Page 10, line 6: then, your result depends highly on the resolution of your jpdf. how is ensured that this does not lead to biased results? e.g. convergence study?	As mentioned earlier, our Rosenblatt transformation uses continuous functions and we don't expect any issues with the resolution of the joint pdf.	–
Page 10, line 8: then, the definition of the evaluation point would be dependent of the model output, which likely will lead to biased results, no?	Yes the results will most likely suffer a bias from using such an approach. On the other hand, in this way we tend to pick points which are a closer match for the target point in the variable dimensions which have the highest impact. This may work towards reducing the bias as we increase the error with respect to variables which have smaller impact, but reduce the error with respect to variables with higher impact. In our experience the net result was reduction in bias.	–
Page 10, line 10: not really covered. could be left out.	The length of the section was reduced significantly - only the short description of bootstrapping is left as this is the only CI estimation method actually used in the paper.	Removed/left out most things around CI estimation
Page 11, line 25: indicate which method was chosen in this study. if not both are used, it may be sufficient to only present one and briefly mention the alternative	Indeed, only bootstrapping was used and we have only present bootstrapping in the revised paper.	–
Page 12, line 2: low-fidelity? same turbine / model used?	"Low fidelity" was added. The "site-specific" loads are computed using the surrogate models. A full quasi-MC simulation was also carried out for each site as reference, and using the same DTU10MW model. This explanation is added to section 6.2	Added "low-fidelity", and explanation for reference quasi-MC simulations.
Page 12, table 2: have these calculations been performed in other work?	No, these calculations are done specifically for the present study although the measurement data sets may have been used in previous studies for other purposes.	–
Page 12, table 2: if only IA is used in this study, what are the different turbulence classes useful for?	We do not use only class IA, the study is not connected or limited to a specific class. We predict the site-specific loads for several hypothetical sites each corresponding exactly to certain IEC-class conditions.	–
Page 12, line 15: please provide the functional relationships	This is done in a new table (Table 6)	Added a table for functional relationships
Page 12, line 16: why pseudo monte carlo?	Quasi-MC (the "pseudo" term in the manuscript is now corrected) is used because it converges faster and allows using a smaller sample size.	corrected to "quasi-"
Page 12, line 18: so lifetime damage not calculated according to eq (6)?	It is in fact eq.6 but with equal weights, this is now indicated in the text.	Now indicate use of (6) with equal weights.

Page 12, line 19: based on all samples? why use bootstrapping, why not simply the standard deviation? any results?	Bootstrapping allowed shuffling of both the selection of sample points as well as the selection of turbulence seeds at each sample point, meaning it takes into account two sources of uncertainty simultaneously. The resulting confidence intervals are shown on some of the results figures.	–
Page 13, figure 4: plot difficult to read. what information is conveyed here? the figure does not seem necessary for the line of argument of the paper.	–	This figure along with other figures depicting the sites was removed from the manuscript
Page 14, figure 5: again not clear why these figures are necessary	–	This figure along with other figures depicting the sites was removed from the manuscript
Page 14, line 1: what about the other models mentioned in the abstract? why not call this surrogate models as in the abstract?	–	We rename the section to "Load mapping functions".
Page 14, line 11: what is xi?	A variable in the range [0,1]. Clarification is added to the manuscript.	Clarification is added to the manuscript.
Page 15, figures 6 and 7: consider leaving these plots out	–	This figure along with other figures depicting the sites was removed from the manuscript
Page 16, figures 8 and 9: consider leaving these plots out	–	This figure along with other figures depicting the sites was removed from the manuscript
Page 17, line 4: if independence is to be ensured, why does dependence have to be accounted for?	–	Rephrased to "the evaluation of the cumulative distribution in general does not account for dependence between variables - this has to be addressed by applying an appropriate transformation"
Page 17, line 5: why is it convenient?	It is convenient because the joint probability distribution is defined in terms of conditional dependencies so applying the Rosenblatt transformation is straightforward. Note added to text.	Added justification/note
Page 17, line 6: normal	–	Corrected in the entire manuscript
Page 17, line 7: check consistency. either reduced order model, surrogate or response surface	Consistency was improved by changing the "reduced order model" expressions to "surrogate model". The "response surface" refers to one specific surrogate model - the quadratic response surface. The clarification "quadratic" is added where necessary.	Changed "reduced order model" to "surrogate model". Added "quadratic" where needed.
Page 17, line 8: not clear what a legendre polynomial is. can you introduce?	–	Legendre polynomials are introduced.

Page 17, equation 12: what exactly is happening here?	Each of the terms in the multivariate PCE represents a product of univariate Legendre polynomials. Equation (12) introduces the condition that the total order in each term (the sum of the orders of the univariate polynomials) does not exceed the maximum order of the expansion. Then Equation (13) shows how the multivariate polynomial terms are obtained by taking the product of the univariate polynomials.	Equations (12)–(13) are now moved to the appendix.
Page 17, line 15: this part needs more description to be understood.	The explanation for the total number of polynomials will add to the length of the paper which is already quite long. Instead, we have provided a reference where this is explained in more details. The whole discussion is now moved to the appendix.	Moved discussion to appendix.
Page 18, line 12: how was the regression performed? there seems to be a section or paragraph missing on this	Here "regression" refers to the generic process of obtaining model coefficients using least-squares minimization. In particular, we use the LASSO for regularizing the PCE model. We have thus replaced "regression" with "model" where necessary.	replaced "regression" with "model", where needed.
Page 18, line 14: standard expression is NRMSE	–	NRMS was changed to NRMSE
Page 18, line 16: how was the PCE based surrogate model established? the same set of points? clarify that you are now using data from section 2.4, if this is the case.	–	Now clarified that we are using the data from section 2.4
Page 18, line 19: a "longer" simulation here means the consideration of a larger number of seeds?	Correct, this is larger number of seeds.	Clarification added.
Page 18, line 25: is an "overfitting" possible as well?	Overfitting is theoretically possible, but only likely in cases where there are only few distinct values of a given variable. We haven't seen any overfitting (which can be easily recognized in case the model produces a higher r-squared value with the training set than with the validation set).	–
Page 18, line 29: showing some scatterplots of original and sampled data would give an intuitive view on the quality of the results	Indeed, adding a scatter plot might enhance the understanding of our statements - however we have to deal with the fact that the manuscript is already very long and detailed, and we prefer to skip this plot.	–

Page 18, line 29: this sounds like a certain uncertainty will always exist. the common understanding would be that uncertainty is reduced through additional samples and longer simulations. please take this into account in the line of argument.	Indeed the formulation was not precise. It was modified to the following: "Further increase in the number of training points or simulation length will only reduce this statistical uncertainty, but will not contribute significantly to changes in the model predictions as the flexibility of the model is limited by the maximum polynomial order."	Modified the line of argument to be more descriptive and clear
Page 19, figure 10: why this increase and decrease?	We do not have a definitive answer. One possibility is that there are numerical issues due to the size of the design matrix and hence the linear system being too small to get a well-defined solution for all the 924 PCE coefficients.	–
Page 19, figure 10: are these single-point evaluations or has the evaluation done with a varying set of samples?	Each point on the surface represents the NRMSE computed between approximately 500 quasi-MC samples generated from the joint probability distribution of site 0, and the corresponding predictions by the PCE for the same points. Each of the quasi-MC samples is the mean from 48 turbulent 10-minute simulations. To mimic the seed-to-seed uncertainty, each of the PCE predictions is also evaluated as the mean of 48 normally distributed random realizations, with mean and standard deviation prescribed by the PCE model for mean and standard deviation of the loads respectively. Following text was added: *Each of the quasi-MC samples is the mean from 48 turbulent 10-minute simulations. To mimic the seed-to-seed uncertainty, each of the PCE predictions is also evaluated as the mean of 48 normally distributed random realizations, with mean and standard deviation prescribed by the PCE model for mean and standard deviation of the blade flapwise DEL respectively.*	Added descriptive text (at left).
Page 19, line 7: consider the two in different chapters. model reduction is very interesting, but the sensitivity indices can be calculated with other surrogates as well. also, SI and ANOVA should be introduced before model reduction	Correct, sensitivity indices can be calculated with other surrogates as well. We have taken parts of this section out and left it as part of Appendix A. Nevertheless, we have left the model reduction (in a separate section) because we do use the Galerkin approach with model reduction where we aim at retaining 99.5% of the variance.	Moved parts of section to appendix

Page 19, line 8: orthogonality meaning that input variables are independent?	The polynomials in the polynomial basis are orthogonal which eliminates the cross-terms (covariances) when computing the contribution of each individual polynomial to the model variance.	–
Page 20, line 20: this part should be more general as it is applicable to any surrogate model	Correct.	All reference to PCE are replaced with "surrogate"
Page 20, line 21: have you compared the monte carlo based and the pce-inherent indices?	In order to have a valid comparison, the Monte Carlo based indices have to be evaluated on a data set with the same distribution as the PCE training set. We did the comparison using the points from the high-fidelity database as means to validate our Monte Carlo-based approach, and the results were satisfactorily close.	–
Page 20, equation 23: how many points were used?	Approximately 500 per dimension. This is now noted in the text.	Approximately 500 per dimension. This is now noted in the text.
Page 20, line 30: again, please use only one expression for surrogate models	–	Changed from "metamodel" to "model"
Page 21, line 1: indicate dimensionality of new variables	The dimensionality is $N \times M$.	Dimensionality $N \times M$ noted in the text.
Page 21, line 2: what kind is typical? linear, polynomial, ...?	If the trend function is replaced by a constant (i.e. the mean of the field) the resulting model is referred to as simple Kriging; a linear trend is denoted as ordinary Kriging, while with any other more advanced function the model is called universal Kriging. For brevity, we only note this in the Appendix.	Only note Kriging detail in appendix
Page 21, line 3: w? not in eq 24	It's a typo, it should be $Z(x)$	corrected typo
Page 21, line 6: overall variance?	overall variance noted	corrected
Page 21, line 7: w?	w and x are two different points in the domain. Clarification added.	Clarification added.
Page 21, equation 26: R now bold?	$\mathbf{R}$ is the correlation matrix with individual elements R_{ij}, this is defined below equation A12.	
Page 21, line 14: N? P?	text added to Appendix: "N is the number of samples and P is the total number of terms output from the basis functions — which may be different than the number of dimensions M as a basis function (e.g. a higher-order polynomial) can return more than one term per variable"	text added to Appendix

Page 22, line 19: why is this an advantage	The Kriging model has a smooth surface and also provides an **exact** prediction at the training points, meaning that at least in the near vicinity of the training points it should outperform a model which does not satisfy these conditions	–
Page 22, line 30: why is this explained in so little detail?	A similar load prediction procedure using the quadratic response surface method is described in details in Toft et al., we think the reference provides a sufficient amount of details on how the method works.	–
Page 23, line 12: is this a fair comparison with the other models?	It is true that the model training points are less than for other methods, but we wanted to illustrate the specific experimental design that can be used with this method. One can also use the high-fidelity database points binned according to wind speed and fit a quadratic response surface for data in each bin. We tested that and in our experience it did not improve the results	–
Page 23, line 21: why is this pseudo MC? if it refers to the origin of the sampling points, it should still be considered MC as there is no difference in the evaluation procedure	–	text changed to "full MC"
Page 23, line 21: not clear why importance sampling and nearest neighbor interpolation are considered differently here. also a classification of the presented methodology would be helpful (i.e. surrogate modeling applied? number of simulation? etc) also which simulations are using the same set of points?	In the updated manuscript, all surrogate model approaches are presented in the same section. A table comparing the methods (number of samples, computing time etc.) is also introduced.	In the updated manuscript, all surrogate model approaches are presented in the same section. A table comparing the methods (number of samples, computing time etc.) is also introduced.
Page 23, line 27: it is very complicated to digest all these special rules for different models & sites. i propose to strongly simplify what has been done or include clear overviews that show what has been done efficiently. in written form is not sufficient	Information about the number of MC samples used in site-specific simulations is included in a new table. Together with some improved explanations it is hopefully clear how the rules for different models and sites are applied.	New table for site-specific simulations added, along with improved explanations.
Page 24, line 8: why are two approaches presented? one should be clearly enough and would lower the confusion	Only reference to bootstrapping is retained in the revised version.	Only bootstrapping presented

Page 24, line 12: how was bootstrapping applied for mc and surrogate models? with/without replacement, how many simulations out of all simulations is the reference? based on sampling from surrogate models?	An explanation about the way bootstrapping is applied is included in the end of section 3.3	New text added: *In the present study, bootstrapping is applied by generating independent bootstrap samples each with size equal to the entire data set. Both the sample points and the turbulence seed numbers are shuffled, meaning that the resulting confidence intervals should account for both the statistical uncertainty due to finite number of samples, and the uncertainty due to seed-to-seed variation. Note that these two uncertainty types are the only ones accounted for in the confidence intervals.*
Page 25, figure 12 caption: a table highlighting main characteristics of simulations would be helpful. here, the information on how many simulations were used for MC and all other simulations	two new tables are provided - with site-specific distribution properties, number of simulations used, and another one with characteristics of the surrogate models.	new tables added
Page 25, figure 12 caption: 5% and 95%?	It is the 95% confidence interval, containing 95% of the probability, between the 2.5% and 97.5% quantiles. The 95% confidence interval is a standard definition and we would prefer to retain it in the manuscript.	–
Page 25, line 1: not done for evaluation of fig 12? the figure is meaningless if the models under comparison are not based on a similar number of samples, no?	Yes the comparisons are based on the same number of samples of course. But Figure 12 has a different scope so this is first mentioned for Figure 13.	–
Page 25, line 9: better show as barplots	–	Tables 3-7 have been replaced with one table (now Table 7) showing the mean results from all sites (i.e. the last two lines from each of tables 3-7 from the first version of the manuscript), and two figures showing the results for individual sites as bar plots.
Page 26, line 1: not clear how these samples are distributed	They are simply discrete wind speed values from 4 to 25m/s, and with deterministic turbulence intensity as prescribed by the IEC 61400-1 standard.	–
Page 26, line 1: IEC?	–	Corrected
Page 26, line 4: not clear why this would happen	It is because fewer points from the high-fidelity database will have high probabilities with respect to the site-specific distribution.	Note added to text.
Page 26, line 10: better NRMSE	–	Changed to NRMSE

Page 28, table 3: better to use plots then numeric output. as this is a comparison study the exact values are of limited importance results for different models should be presented in same plot, rather than different sites	–	Tables 3-7 have been replaced with one table (now Table 7) showing the mean results from all sites (i.e. the last two lines from each of tables 3-7 from the first version of the manuscript), and two figures showing the results for individual sites as bar plots.
Page 28, line 3: sobol indices only evaluated from PCE?	Sobol indices have been evaluated only from PCE, but using two different methods - one which directly uses the PCE coefficients, and another which utilizes Monte Carlo simulations with the model. The Monte Carlo based method is general and not limited to the PCE model. This is made clearer with the updated structure of the paper where more emphasis is put on the Sobol indices evaluation using Monte Carlo simulations.	updated structure of the paper
Page 28, line 4: shouldnt uniform distribution be assumed for calculation of sobol indices?	The Sobol indices are computed with respect to the quasi-MC sample point locations which are uniformly distributed in the interval [0,1)	–
Page 28, line 6: what does uniform & bounded stand for?	–	the phrase "uniform & bounded" was removed from the text
Page 28, line 7: total or single indices?	total indices, added to text	"total indices" added to text
Page 30, line 5: what is a measure for robustness here?	being sufficiently accurate in the entire domain, without creating outliers.	Text modified to: *sufficiently accurate over the majority of the sampling space*
Page 30, line 10: RMSE	–	Corrected
Page 33, figure 15: y-y plots would be more helpful for this comparison. the x-axis is without information	–	The plot in this figure was changed to a y-y plot as recommended.
Page 33, line 6: ANOVA may be performed with any surrogate, no?	Yes but in the case of the PCE this makes for a quick and efficient way of model reduction. This is clarified in the text now.	clarified in the text
Page 33, line 9: why deep?	We have some experience with making the same model with Neural Networks (Scrhøder, Dimitrov, Verelst and Sørensen, Torque 2018 conference proceedings). It takes at least 2 hidden layers to provide sufficient accuracy. Nevertheless, we've changed "deep" to "sufficiently large" to avoid misinterpretation.	changed "deep" to "sufficiently large"
Page 33, line 17: how for example?	It could be that the site conditions are uncertain or that the turbine is operated otherwise than intended. Noted in text.	Uncertainty possibilities noted in text
Page 35, line 5: summary and conclusions	–	Changed

Page 35, line 7: and monte carlo simulation, no?	MC simulation is just for reference, to compare the performance of other methods	–
Page 35, line 10: how many simulations were used?	There were many simulations used for different purposes (high-fidelity database, site-specific MC, a dedicated database to fit the quadratic RS). We think that listing and explaining all these in the conclusion will expand it unnecessarily. Instead we have added a sentence stating "... by training the surrogate models on a database with aeroelastic load simulations of the DTU 10MW reference wind turbine"	added explanatory sentence (also note earlier added table)
Page 35, line 12: wind shear and mtls	–	changed
Page 36, table 10: why L so much more importance here?	L affects the turbulence spectrum, which in turn affects the variation in rotor thrust force.	–

Reviewer 2

1) Focus on the most important topics. Perhaps, some topics of minor interest can be left out (or be used in a second paper). Examples are IS, LHS, CI based on the logN distribution, several figures, sensitivity analysis, and extreme loads. Firstly, this would help to shorten the paper to make it easier to read. Secondly, you could give some more (important) details on the other topics.	–	A significant part of the paper was removed or moved to an appendix. The CI based on the logN distribution was removed, also the mentioning of the LHS including the figure showing it, the theory of the surrogate model approaches was shortened and parts were moved to an Appendix.

2) The structure of the paper might be reconsidered. In the beginning, it is confusing that you mix up different topics (e.g.: In section 2, there are subsections concerning the database itself and concerning "reduction methods").	We agree with that comment. The structure of the paper has been modified, so that now all reduced-order model descriptions are in the same section. Some of the theory is moved to an appendix.	New paper structure:
		
		1 Introduction
		
		2 Definition of the surrogate load modelling procedure
		
		    2.1 Step-by-step description
		
		    2.2 Definition of variable space
		
		    2.3 Defining the ranges of input variables
		
		    2.4 Reference high fidelity load database
		
		    2.5 Database specification
		
		3 Post-processing and analysis
		
		    3.1 Time series postprocessing and cycle counting
		
		    3.2 Definition of lifetime damage-equivalent loads
		
		    3.3 Uncertainty estimation and confidence intervals (only bootstrapping to remain)
		
		4 Reduced-order models
		
		    4.1 Obtaining site-specific results using Importance Sampling (shortened)
		
		    4.2 Obtaining site-specific results using multi-dimensional interpolation (shortened)
		
		    4.3 Polynomial chaos expansion (shortened)
		
		    4.4 Universal Kriging with polynomial chaos basis functions (shortened)
		
		    4.5 Quadratic response surface (shortened)
		
		    4.6 Sensitivity indices and model reduction (shortened)
		
		5 Model training and performance
		
		    5.1 Convergence
		
		    5.2 One-to-one comparison and mean squared error
		
		    5.3 Variable sensitivities (shortened)
		
		6 Site-specific calculations

3) The explanations regarding the environmental conditions remain quite vague. For the database, the reader has to "search for" the distributions utilised. For the sites, they are not given and dependencies are not.	We agree that the explanations regarding the environmental conditions especially at the validation sites were insufficient, this is also pointed out by the other reviewer. We have now added explanatory text to Section 6.1, as well as a table (Table 6) listing the functional relationships which define the conditional distribution properties.	–
4) The implementation of importance sampling is questionable. IS should focus the sampling on important regions (those conditions where high fatigue damages occur). You sample according to the uniform (database) distributions. This might be the reason why IS is performing so badly.	Correct, the importance sampling density is not optimal. Nevertheless, we use a procedure where we try to pick the most important points, by evaluating $h(X)$ for all points in the database and taking only a fraction of them with the highest importance. An explanation for this was though missing in the paper. We have now added some clarifications to the text.	New text in Section 4.1: *"This is a non-standard application of the IS approach, because normally the IS sample distribution is chosen to maximize the probability density of the integrand. In the present case, this objective can be satisfied only approximately and only in cases where the number of IS samples, N_{IS}, is smaller than the total number of database samples, N. Under these conditions, the importance sampling weights $(f(\mathbf{x_i})/h(\mathbf{x}_i)$ from Eq.8 can be evaluated for all points in the database, but only the N_{IS} points with the highest weights are included in the further calculations. This is the approach adopted in the present paper."*
5) It would be beneficial, if you should revise the theoretical sections. These sections need more detailed explanations. As you compare different methods, you cannot expect the reader to be an expert in all of them. So, don't leave out to many intermediate steps. If you don't want to give more details, then you should leave out the whole mathematical derivation and give only the final equations (and refer to the corresponding literature).	Here we are facing a difficult choice. We are aware that adding explanations will make the work clearer, but at the same time the paper is already quite long and other important details need to be explained. Therefore a good balance is needed. Based on the reviewers' recommendations we have included additional explanations for some missing steps which are a unique part to this study (e.g. the procedures for deriving the environmental conditions joint distribution) but at the same time for theoretical methods available in literature we have reduced the text to some final equations, and placed the remaining explanations in an appendix.	–
6) Some equations seem to be inconsistent or have typos. Please, revise all equations carefully.	–	Equations have been revised

7) The comparison of the different methods lacks overview. Please, provide a Table or something similar summarising the number of samples used, the difference in CPU time, etc.	This was also pointed out by another reviewer. We have introduced a new table at the end of Section 6.2, which summarizes the number of training and evaluation samples, as well as the executions speed.	–
8) A discussion regarding the comparison would be interesting. Is it a fair comparison, if you don't take the 10000 calculations for the database into account? In my opinion it is questionable to compare 1000 MCS samples with PCE based on a database with 10000 samples. Especially since the database (probably) has to be build up for every new design, this is not really "fair". So, this approach "only" helps to analyse the same turbine at different sites. This should be clarified or it has to be explained why the comparison is "fair".	We have not included the MCS with the intention to compare it to a PCE or Kriging model. It is rather intended as a reference which all other methods should compare to. This is made clearer in the text, in Section 6.2. Also, in some places the surrogate model list is given as consisting of 6 models, which is misleading as we actually have 5 models and 1 reference. This is now corrected. With regards to the database, this is exactly its scope - to be able to use it for a single turbine type on different sites. This is already stated in section 2.2.	Text in section 6.2 has been changed.
Page 1, title: The title is not really matching the main topic of the paper. "Surrogate models" should appear somehow.	The title has been changed to reflect the use of surrogate models.	The new title reads *From wind to loads: wind turbine site-specific load estimation with surrogate models trained on high-fidelity load databases*
Page 1, line 4: Are IS and NN really surrogate methods?	Indeed, IS and NN are different than the machine-learning based regression models and can be considered as a sort of "table lookup" procedures. Nevertheless we think it is useful to have a single term that encompasses all approaches, and "surrogate models" and "load mapping functions" are the best candidates.	Some clarifications are added to the first paragraph of Section 6, to notify the reader that the IS and NN approaches differ from the remaining 3.
Page 1, line 9: If you don't name the other properties here, leave it out in the abstract.	–	The last sentence from the abstract was removed.
Page 1, line 17: Formatting error?	–	Corrected.
Page 1, line 22: Also mention examples for Kriging and IS, e.g. Dynamic reliability based design optimization of the tripod substructure of offshore wind turbines: Hezhen Yang, Yun Zhu, Qijin Lu, Jun Zhang Importance Sampling for Reliability Evaluation With Stochastic Simulation Models: Youngjun Choe, Eunshin Byon & Nan Chen	Thanks for the suggested references. The first one was included in the introduction, while the second one was listed in the section dedicated on IS, together with a recent paper by Graf et al. (2018).	added
Page 2, line 10: Is there a reference?	–	Two references were added (Dimitrov et al., 2017, Bak et al., 2013)
Page 2, line 14: Sounds strange: You are not talking about high-fidelity loads, but loads calculated using high-fidelity models	–	The name of the section is changed to "Definition of the surrogate load modelling procedure

Page 4, equation 1: This is not clear. Either leave it out or give more explainations: What type of dependent distributions do you use?	The dependent distributions for the high-fidelity load database are given in Table 1. The distributions are uniform and the bounds are conditionally dependent on other variables. The dependent distributions for the site-specific calculations are now given in the new Table 6. Please see also the response to Reviewer 1.	–
Page 4, line 12: Perhaps you can shorten this section by including the references in Table 1. If you want to keep it, explanations for the bounds of ϕ_h, ϕ_v, and ρ are missing	Explanation for the bounds of the last three variables was added to the manuscript. The bounds for these three variables are simply chosen arbitrarily to cover what we consider a usefully wide range.	–
Page 5, table 1: It would be nice, if this Table summarises the whole environmental conditions considered. Hence, include distributions (or state that you are using uniform distributions for the database itself (U is beta-distributed?)) and dependencies (Since, uniform distributions are used, only the bounds are dependent?)	The database uses uniform distributions with the exception of the wind speed - and as the reviewer correctly points out only the bounds are dependent. The dependencies are actually given in Table 1. We have now added a note to the text saying that for the database only the bounds are dependent. On the other hand, the site-specific load simulations use true conditional distributions - these are now defined in the new Table 6.	Added note to the text: the bounds are dependent only for the database; the site-specific load simulations use true conditional distributions, now defined in the new Table 6.
Page 5, table 1: Comma is missing	–	Corrected
Page 5, line 9: High-fidelity loads?	The name of this section was changed to "'sampling procedure'", see response to reviewer 1	–
Page 6, line 14: If you are not discussing LHS, leave it out	–	Indeed, we have now removed the discussion about LHS

Page 7, figure 2: Why is U beta-distributed? Include distributions in Table 1.	The distributions are now included in Table 1. U is beta-distributed in order to obtain more samples at low wind speeds where the bounds of other variables are wider and the sample space is more sparsely covered.	Following was added to text: *For the case of building a high-fidelity load database, all variables given in Table 1 except the wind speed are uniform, and only the distribution bounds are conditional on other variables as specified by the 2nd and 3rd columns of the table. The bounds of several variables are conditional on the wind speed, and as shown on Figure 2 they are wider at low wind speeds, meaning that more sample points are needed to cover the space evenly. This dictates that the choice of distribution for the wind speed should provide more samples at low wind speeds. In the present study we have selected a Beta distribution, but other choices as e.g. a truncated Weibull are also feasible.*
Page 7, line 3: Interesting apporach to use 8h of simulation per sampling point. Have you checked or any reference that this leads to better results than only 1h per sampling point and 8 times more sampling points (also including seed-to-seed variations, but more different conditions due to more sampling points)	We actually use 8 one-hour simulations, when stating 8h we simply mean the total duration of the simulations. A single 8-hour simulation would bring limitations to the turbulence generation procedure, where due to memory limitations only a turbulence box with given maximum number of points can be generated (16384 or 32768 points longitudinally). Making such a turbulence box correspond to 8h duration would mean very low temporal resolution of the generated wind field (in the order of 0.5 - 1 turbulence planes per second). For clarifying what we do, the text is changed to "'For each sample point, eight simulations, with 3800s duration each, are carried out. The first 200s of the simulations are discarded in order to eliminate simulation run-in time transients, and the output is 3600s (1h) of load time series from each simulation."'	updated text to explain

Page 7, line 7: Why don't you use 10min simulations, if you keep the conditions stationary anyway?	We wanted to capture some of the low-frequency fluctuations generated by the Mann model turbulence, especially at larger turbulence length scales. When we generate a 6x longer turbulence box, it includes more of these low-frequency variations, which in fact introduce some degree of non-stationarity when looking at 10-minute windows. So this results in some, in our opinion, more realistic seed-to-seed variations.	–
Page 7, line 9: What run-in time is used?	The run-in time is 200s. This is now explained in the text (see response to earlier comment).	now explained in the text
Pgae 8, line 3: Are the simulations 1h or 10min? This is confusing now.	Simulations are 1h long, subsequently split into 10min chunks to compute 10-min damage-equivalent loads. We have added a bulletpoint explaining that.	added a bulletpoint explanation
Page 8, line 8: This is somehow confusing: S_i are the load ranges. They are not estimated using the rainflow counting algorithm, but n_i is counted. Please, reformulate the expression.	Actually the rainflow counting algorithm by definition outputs a list of single load half-cycles where each half-cycle has a unique amplitude and a direction (positive or negative). For each half-cycle determined by the rainflow algorithm $n_i = 1$. The binning is only a postprocessing step and is in principle not necessary for evaluation of damage-equivalent loads, it is only done in the cases when the load spectrum needs to be visualised or shared in simplified form.	–
Page 9, line 1: Perhaps put this section in section 4 or leave out IS. Mixing the creation of the database with the investigated "reduction concepts" makes it hard to understand	–	This is now part of section 4.
Page 9, line 2: Use "section" not §	–	Corrected
Page 9, line 11: Notation is not consistent with section 2.5.2	–	Notation for variables $\mathbf{X}$ was made consistent with section 2.5.2.
Page 9, line 16: This is not really the idea of IS. For IS, you should choose $h(X_i)$ so that your sampling is concentrated on "important" regions (where high damages occur). These regions have to determined beforehand (e.g. using surrogate models). This is not done here! Therefore, the bad performance of IS is due to the chosen sampling function $h(X_i)$	This relates to one of the general comments, see earlier discussion.	–

Page 9, line 21: Again, this section might fit better in section 4 in order not to mix the database and the "surrogate" models	–	Moved to section 4
Page 11, equation 9: Φ^{-1} ?	Equation 9 was deleted as we don't use this method for CI estimation.	deleted Equation 9
Page 11, equation 9: $\mu + \Phi^{-1}(\alpha/2) * \sigma$ $\Phi^{-1}(\alpha/2)$ is already negative	Equation 9 was deleted as we don't use this method for CI estimation.	deleted Equation 9
Page 11, equation 9: Perhaps use "ln" instead of "log". "Log" is sometimes also used for $\log_{10}$. Or state that it is the natural log.	Equation 9 was deleted as we don't use this method for CI estimation.	deleted Equation 9
Page 11, line 28: Why do you explain both CI methods. In the end, you only use the bootstrapping approach. So leave the other one out.	–	Only bootstrapping was kept, the text about the other CI method was removed.
Page 12, line 10: At least for one site (e.g. site 0) you should list the distributions and dependencies you use	–	Distributions and dependencies are now listed for all sites in Table 6.
Page 13, figure 4: These Figures don't make clear where the locations are. So, either make it clear (e.g. a map of Denmark with all (site 0, 1, and 2) sites marked clearly) or leave these figures out.	All figures related to the site locations were left out, as the scope of the paper is not necessarily to analyse specific sites and their properties, and the paper is quite long anyway.	Removed all figures related to the site locations
Page 13, line 6: It might be nice to know the wind direction filtering you applied.	We agree; but again, the analysis of the particular sites is not directly in the scope of the present paper, we are interested most in the way the surrogate models perform for various conditions, so discussing the direction filtering would add complexity to the paper but not necessarily contribute to the conclusions.	–
Page 13, line 8: So, this is just one site. The "sites" 2-4 are just different wind directions. Perhaps, you could clarify this (e.g. site 2_west, site 2_north, site 2_east or something similar instead of 2-4)	–	We have changed the definition from "'sites'" to "'virtual sites'" and noted that virtual sites are created by direction filtering.
Page 14, figure 5: Leave it out	–	Figure deleted.
Page 15, figure 6: Leave it out	–	Figure deleted.
Page 15, figure 7: You don't use this Figure. Leave it out.	–	Figure deleted.
Page 16, figure 8: Perhaps you can use this Figure to visualise the directional filtering by plotting the sectors (mountains, flat region) in this Figure	–	The Figure has been removed entirely.
Page 16, figure 9: You don't use this Figure. Leave it out	–	Figure deleted.

Page 17, equation 12: (alpha>=0) is not needed, as alpha element of N alpha has to be >= 0	–	(alpha>=0) is removed.
Page 17, line 13: This section is really hard to understand, especially as you cannot expect the reader to be expert in all methods. Additional explanations are needed! Some examples (e.g. a list of the first Lagendre polynomials) would help	–	We have done several things to improve this section. Some of the more advanced explanations were placed in an Appendix; a list of the first Legendre polynomials as well as the recurrence formula was provided.
Page 17, equation 14: Here: N_p = (M+p) choose p = (M+p)!/(M!p!) would help to understand the selection based on eq (12). An example with, e.g. p=1, M=2, would clarify it: N_p = (2+1) choose 1 = 3 Psi_0 = P_(0,1)*P_(0,2) Psi_1 = P_(1,1)*P_(0,2) Psi_2 = P_(0,1)*P_(1,2)	–	We have added ($(M+p)$ choose p) to the equation formula. However, this is now outside the main paper and part of Appendix A instead - so we have skipped further explanations as we assume the reader can find that
Page 17, equation 15: Do we need alpha here? j is already the index for all N_p polynomials. So, using two indices might be confusing or is there a reason for it?	We need alpha as it indexes the different variable dimensions, i.e., each multivariate polynomial with index j is built as the product of M univariate polynomial terms, and alpha indexes these univariate polynomial terms. This is now mentioned in the manuscript.	Need for α mentioned in the manuscript
Page 17, equation 16: This is really confusing! this is not g(x), as it could be assumed by considering line 17. Here, we are determining the regression terms. Use another notation.	–	We have replaced the $\mathbf{X}$ on line 17 with $\boldsymbol{\xi}(\mathbf{X})$
Page 17, equation 17: ξ^i not xi^j?	–	Corrected
Page 17, equation 17: Again, do we need alpha here?	Please see response to our earlier comment	–
Page 18, equation 18: Hard to understand! It would help, if you state that the approxiamtion in eq (15) yields: y = Psi*S and eq (18) is the solulation of y = Psi*S	–	The suggested statement is added in Appendix A.
Page 18, line 3: g(x) or g(xi)?	–	It is g(xi), now corrected
Page 18, equation 19: You might leave out the whole section on LASSO. If not, make clear that is only used in a second step?	The LASSO is not used as a second step, but as an alternative approach for determining the polynomial terms by gradient-based optimization.	–
Page 18, equation 20: eps_NRMS	–	Corrected.
Page 19, line 1: NRMS?	–	Corrected
Page 19, line 5: approximately	–	Changed to "'approximately"'

Page 19, line 7: Perhaps leave out the whole sensitivity analysis. The paper is very long, it will become even longer with more (important) explanations	We prefer to keep the sensitivity analysis as it leads to some important conclusions regarding the influence of several environmental variables on loads. Nevertheless, we have modified the manuscript so that the sensitivity analysis is seen in a more general (and hopefully easier to understand) form rather than as part of the PCE theory section.	modified the manuscript so that the sensitivity analysis is seen in a more general way
Page 20, line 7: It is not really clear which PCE you use in the end for the results (5005 or 200 polynomials?)	–	We have now explained that we use non-truncated PCE for the results, while the truncation is applied as an example to a specific PCE model which was also used for variable sensitivity analysis.
Page 20, equation 22: perheps use j instead of alpha, as the index was (mainly) "j" in section 4.1	Good point, we have exchanged j and alpha in this paragraph, as we actually use both indices.	We exchanged α for j index
Page 20, equation 23: Here, it is not clear what you use (this becomes only clear while reading the results)	–	In the updated structure of the paper it is made clear in Section 5.3 that we use the MC-based Sobol indices for the site-specific distribution and PCE-based indices for the high-fidelity database.
Page 20, equation 24: Using your defined dimensions of beta and f(x), this should be $f^T(x) * \beta$?	Indeed, this is the right definition, we have modified the equations where necessary.	modified the equations where necessary
Page 21, line 3: In eq (24), it is Z(x). Be consistent	–	Corrected
Page 21, equation 25: Perhaps, x_i and x_j are clearer than w and x. A definition of w (or x_i and x_j) could be helpful	w is now defined as a point in the domain distinct from x, and w and x are jointly Gaussian distributed. We prefer to use w and x instead of x_i and x_j because later the indexes i and j are used for a different purpose.	w is now defined as a point in the domain distinct from x, and w and x are jointly Gaussian distributed.
Page 21, line 10: Before stating eq (26), the joint distribution of Y(x) and Y(x') would be nice. $(Y(x')Y(x))^T \ N[(f(x')\Psi)^T * \beta, \sigma^2 * ....]$	–	The joint distribution of Y(x) and Y(x') is now stated in Appendix A (equation A12). A large part of these formulations are though omitted from the main manuscript for simplicity, and the reader is referred to the Appendix.
Page 21, equation 26: Do we need σ_Y^2? It is not used.	The definitions of both μ_Y and σ_Y^2 have been removed from the main manuscript. They are retained in the Appendix - where σ_Y^2 is also given as it provides more completeness of the description.	definitions of both μ_Y and σ_Y^2 have been removed from the main manuscript, but retained in the Appendix
Page 21, line 14: This is not really consitent with f(x) in line 2	–	All equations in the section regarding Kriging are modified for consistency.

Page 21, line 14: Define N and P	–	N and P are defined below equation A12 (Appendix A).
Page 21, equation 27: This is not clear without further explanations. Perhaps state that beta, σ^2, and θ can be determined by minimising $-\log(L(y\|\beta,\sigma^2,\theta))$	Isn't that exactly what we are stating with the phrase *"A suitable approach is to find the values of β, σ^2 and θ which maximize the likelihood of y"* which is written just above the equation?	–
Page 21, equation 28: this is the solution of $d(-\log(L))/d(\beta)=0$	–	Clarification added to Appendix A.
Page 21, equation 29: this is the solution of $d(-\log(L))/d(\sigma^2)=0$	–	Clarification added
Page 22, equation 30: What is D_theta, why not theta?	–	D_θ is changed to θ
Page 22, line 23: Is the higher computing time of Kriging a real problem? Normally the creation of the database is the limiting factor (see overall comments as well)	Yes we think in this case the higher computing time becomes a problem as it is an order of magnitude longer than other methods (table 8). It may still be applicable for one-off computations, but poses difficulties for carrying out e.g. parametric studies or optimization.	–
Page 23, line 8: Do you know that this is possible for other parameters than the wind speed? Perhaps, it is beneficial to use several TI response surface as well (this might become complicated having many response surfaces, but you have to justify your decision)	We have added an explanation that using more response surfaces will make it complicated as it will require additional multi-dimensional interpolation.	Text added: *This approach may in principle be extended to include additional variables as e.g. turbulence, however doing so will reduce the practicality of the procedure as it will require multi-dimensional interpolation between large number of models and the uncertainty may increase.*
Page 23, line 11: Why are these variables (and not others) replaced by thier mean values. Sensitivity analyses?	We explain that these are variables with relatively low importance according to the sensitivity analysis	–
Page 23, line 12: Explain that this number is $22*(1+2k+2^k)$	–	Explanation included
Page 23, line 28: Is this a fair comparison? You use only 1000 MCS samples, but the meta-models are calibrated on 10000 samples. Hence, the meta-models (including the creation of the database) require a 10 times higher computing time.	The meta-models and their computing times are evaluated on exactly the same number of samples as the MC simulation. This is clarified with some additional explanations and is also visible in Table 8.	clarified with some additional explanations
Page 23, line 31: How many samples do you use?	It is the same sample used for the full site-specific MC simulations, this is now clarified.	clarified in text
Page 24, line 8: If you use eq (10), don't mention eq (9)	–	Equation (9) and the supporting text have been removed from the manuscript, as well as any text mentioning it.

Page 25, line 2: How many samples are used?	–	Number of samples is listed in Table 8 (reference added to text).
Page 26, figure 13: The high uncertainty of IS might be a result of the badly chosen h(X). Leave IS out or revise it.	These are the results from the best possible choice of $h(X)$ which can be drawn from the existing database and does not involve carrying out new simulations. We have added a clarification though that this is a non-standard use of IS, see response to general comment 4)	
Page 28, table 3: Do we need all these Tables? Perhaps, just use two Tables: first one like Table 3 (one method, all sites, all loads); second one with all methods, all sites, one load	Tables 3-7 have been replaced with one table (now Table 7) showing the mean results from all sites (i.e. the last two lines from each of tables 3-7 from the first version of the manuscript), and two figures showing the results for individual sites as bar plots.	
Page 28, line 1: Perhaps leave out this section. Sensitivities could be regarded in a seperate paper in more detail.	As discussed earlier (see response to comments for page 19) we would like to keep the sensitivity analysis, in a modified form so it is easier to understand.	
Page 28, line 8: You should briefly mention why you have different numbers of variables in Table 9 and 10.	The text now states: *The indices for the site-specific distribution corresponding to reference site 0 are computed using the Monte-Carlo based method described in Section 4.6 as direct PCE indices are not available for this sample distribution. The resulting total Sobol indices for the 6 variables available at site 0 are listed in Table 4.*	
Page 28, line 8: You use different methods in Table 9 and 10. This has to be stated and justified (e.g. for the site, PCE based sensitivity indices are not available) or use MCS based indices in both cases.	This is now stated and justified in the text, see response to the previous comment.	
Page 29, line 1: Maybe leave this out or briefly discuss it in section 6	We have left the ETM computation out of the paper	
Page 32, table 8: Normalised	Corrected	
Page 33, figure 15: The NRMS error would be more illustrative.	We have computed the NRMSE as a statistical measure for an entire evaluation set (and the normalization is with respect to number of samples), while with this figure we would like to show the one-to-one agreement so we can't use the NRMSE.	
Page 33, figure 15: three? Kriging?	Corrected	
Page 33, line 15: about	Changed to "about"	
Page 38, line 28: Wind Energy Science Discussion, under review	Corrected	
Page 39, line 12: This is accepted by now	Corrected	

From wind to loads: wind turbine site-specific load estimation ~~using databases~~ with surrogate models trained on high-fidelity load ~~simulations~~databases

Nikolay Dimitrov, Mark Kelly, Andrea Vignaroli, and Jacob Berg

DTU Wind Energy, Technical University of Denmark
Risø Campus, Frederiksborgvej 399, 4000 Roskilde, Denmark

Correspondence to: Nikolay Dimitrov (nkdi@dtu.dk)

Abstract. We define and demonstrate a procedure for quick assessment of site-specific lifetime fatigue loads ~~, using surrogate models calibrated~~ using simplified load mapping functions (surrogate models), trained by means of a database with high-fidelity load simulations. The performance of ~~six~~ five surrogate models is assessed by comparing site-specific lifetime fatigue load predictions at ten sites using an aeroelastic model of the DTU 10MW reference wind turbine. The surrogate methods ~~include~~ are polynomial-chaos expansion, quadratic response surface, universal Kriging, importance sampling, and nearest-neighbor interpolation. Practical bounds for the database and calibration are defined via nine environmental variables, and their relative effects on the fatigue loads are evaluated by means of Sobol sensitivity indices.

Of the surrogate-model methods, polynomial-chaos expansion ~~provided~~ provides an accurate and robust performance in prediction of the different site-specific loads. Although the Kriging approach showed slightly better accuracy, it also demanded more computational resources. ~~Taking into account other useful properties of the polynomial chaos expansion method within the performance comparisons, we consider it to generally be the most useful for quick assessment of site-specific loads.~~

1 Introduction

Before installing a wind turbine at a particular site, it needs to be ensured that the wind turbine structure is sufficiently robust to withstand the environmentally-induced loads during its entire lifetime. As the design of serially-produced wind turbines is typically based on a specific set of wind conditions, i.e. a site class defined in the IEC (2005) standard, any site where the conditions are more benign than the reference conditions is considered feasible. However, often one or more site-specific parameters will be outside this ~~envelope—and~~ envelope – and disqualify the site as infeasible, unless it is shown that the design load limits are not going to be violated under site-specific conditions. Such a demonstration requires carrying out simulations over a full design load basis, which adds a significant burden to the site assessment process.

1

Various methods and procedures have been attempted for simplified load assessment ~~such as~~ for wind energy applications. Kashef and Winterstein (1999), Manuel et al. (2001) use probabilistic expansions based on statistical moments~~, (Kashef and Winterstein, 1~~ ~~,~~. Simple multivariate regression models of first order ~~(Mouzakis et al., 1999) and second order (Toft et al., 2016), and expansions using orthogonal polynomial basis (Murcia et al., 2018).~~ are employed by Mouzakis et al. (1999), Stewart (2014), while in Toft et al. (2016) a second-order response surface is used. Another response surface approach using artificial neural networks is described in Müller et al. (2017). Polynomial chaos expansion (PCE) is employed by Ganesh and Gupta (2013) for blade load prediction, albeit on a very simple structural representation. Teixeira et al. (2017) use a Kriging surrogate model to map the load variations with respect to offshore environmental conditions. Other relevant studies use some of the methodologies which represent specific analysis steps shown in the present work. These include Hübler et al. (2017) where variance-based sensitivity analysis is employed, Yang et al. (2015) where Kriging is used to enable efficient implementation of reliability-based design optimization, and Murcia et al. (2018) where polynomial chaos expansions are used to carry out uncertainty propagation. In the latter, the model training sample is generated using a Monte Carlo simulation with a quasi-random sequence, a technique which is also employed in Müller and Cheng (2018) and Graf et al. (2016). An alternative to the surrogate modelling approach discussed in this paper could be the load set reduction, as described in e.g. Häfele et al. (2018); Zwick and Muskulus (2016) which also reduces the number of simulations required. This approach however still requires carrying out high-fidelity simulations which leads to using more time for simulation set-up, computations and post-processing, while with a surrogate model the lifetime equivalent load computation takes typically less than a minute on a regular personal computer. The studies most in line with the scope of the present paper are those by Müller et al. (2017), Teixeira et al. (2017) and Toft et al. (2016). The former two employ advanced surrogate modelling techniques (artificial neural networks and Kriging respectively), however the experimental designs are relatively small and with limited range of variation for some of the variables, and the discussion does not focus on the practical problem of computing lifetime-equivalent site-specific loads. The computation of site-specific lifetime-equivalent design loads is the main focus in Toft et al. (2016), however with a limited number of variables, and using a low-order quadratic response surface. The vast majority of the studies employ a single surrogate modelling approach, meaning that it has not been possible to directly compare the performance of different approaches.

In the present work, we analyze, refine and expand the existing simplified load assessment methods, and provide a structured approach for practical implementation of a ~~high-fidelity load database~~ surrogate modelling approach for site feasibility assessment. The study aims at fulfilling the following ~~three~~ four specific goals:

- define a simplified load assessment procedure which can take into account all the relevant external parameters required for full characterization of the wind fields used in load simulations;

- define feasible ranges of variation of the wind-related parameters, dependent on wind turbine rotor size; ~~and~~

- demonstrate how different surrogate modelling approaches can be successfully employed in the problem, and compare their performance; and

- obtain estimates of the statistical uncertainty and parameter sensitivities.

The scope of the present study is loads generated under normal power production, which encompasses design load cases (DLC) 1.2 and 1.3 from the IEC 61400-1 standard (IEC, 2005). These load cases are the main contributors to the fatigue limit state (DLC1.2) and often the blade extreme design loads (DLC1.3) (Dimitrov et al., 2017; Bak et al., 2013). The methodology used can easily be applied to other load cases governed by wind conditions with a probabilistic description. Loads generated during fault conditions (e.g. grid drops) or under deterministic wind conditions (e.g. operational gusts without turbulence) will in general not be (wind climate) site-specific. The loads analysis is based on the DTU 10MW reference wind turbine (Bak et al., 2013) simulated using the Hawc2 software (Larsen and Hansen, 2012).

2 ~~High-fidelity loads database~~Definition of the surrogate load modelling procedure

2.1 Schematic description

Figure 1 shows a schematic overview of the procedure for site-specific load assessment using simplified load mapping functions (here referred to in general as surrogate models). The main advantage of this procedure is that the computationally expensive high-fidelity simulations are only carried out once, during the model training process (top of Figure 1). In the model deployment process (bottom of Figure 1), only the coefficients of the trained surrogate models are used, and a site-specific load evaluation typically takes less than a minute on a standard personal computer.
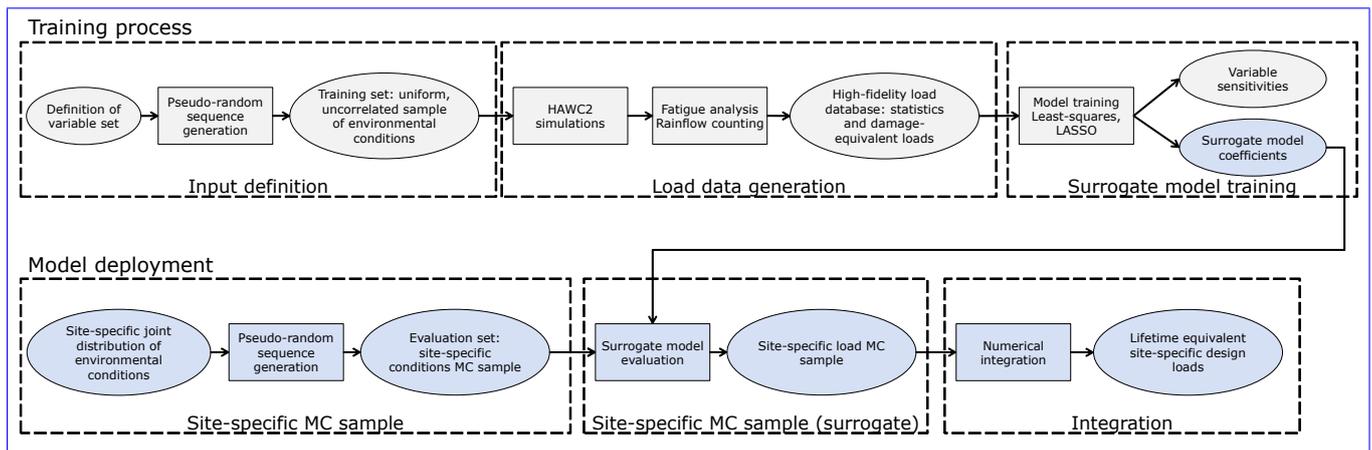
Figure 1. Schematic overview of the site-specific load analysis procedure

2.2 Definition of variable space

The turbulent wind field serving as input to aeroelastic load simulations can be fully characterized statistically by the following variables:

○ mean wind field across the rotor plane as described by the

3

- average wind speed at hub height~~, U~~,

- vertical wind shear exponent~~, α~~,

- wind veer (change of mean flow direction with height, $\Delta\varphi$);

○ turbulence described via

5
- variance of wind fluctuations, σ_u^2,

- turbulence probability density function (e.g. Gaussian),

- turbulence spectrum defined by the Mann (1994) model with parameters

· turbulence length scale L,

· anisotropy factor Γ,

10
· turbulence dissipation parameter $\alpha\varepsilon^{2/3}$;

○ air density ρ;

○ mean wind inflow direction relative to the turbine in terms of

- vertical inflow (tilt) angle $\bar{\varphi}_v$ and

- horizontal inflow (yaw) angle $\bar{\varphi}_h$.

15   All ~~of the parameters above~~ the quantities referred to above are considered in terms of 10-minute average values. All variables, except the ~~parameters~~ variables defining mean inflow direction, are probabilistic and site-dependent in nature. The mean inflow direction ~~parameters~~ variables represent a combination of deterministic factors (i.e. terrain inclination or yaw direction bias in the turbine) and random fluctuations due to e.g. large-scale turbulence structures or variations in atmospheric stability. Mean wind speed, turbulence and wind shear are well known to affect loads and are considered in the ~~IEC61400-1~~ IEC 61400-1

20   standard. In Kelly et al. (2014) a conditional relation describing the joint probability of wind speed, turbulence and wind shear was defined. The effect of implementing this wind shear distribution in load simulations was assessed in Dimitrov et al. (2015), showing that wind shear has importance especially for blade deflection. The Mann model parameters L and Γ were also shown to have a noticeable influence on wind turbine loads (Dimitrov et al., 2017). By definition, for ~~given~~ a given combination of L and Γ the $\alpha\varepsilon^{2/3}$ parameter from the Mann model is directly proportional to $\sigma_u^2 L^{-2/3}$ (Mann, 1994; Kelly, 2018), and

25   can therefore be omitted from the analysis. The probability density function (pdf) typically used to synthesize time series of velocity components from the Mann-model spectra is Gaussian. For a slightly smaller turbine, the NREL 5MW turbine, the assumption of Gaussian turbulence has been shown to not impact the fatigue loads (Berg et al., 2016). The final list of inflow-related parameters thus reads (see Table 1 for details)

$$\{U, \sigma_u, \alpha, L, \Delta\varphi, \Gamma, \bar{\varphi}_h, \bar{\varphi}_v, \rho\}.$$

The loads experienced by a wind turbine are a function of the wind-derived factors described above, and of the structural properties and control system of the wind turbine. Therefore, a load characterization database taking only wind-related factors into account is going to be turbine-specific.

The variables describing the wind ~~inflow~~ field often have a significant correlation between them, and any site-specific load or power assessment has to take this into account using an appropriate description of the joint distribution of input variables. At the same time, most probabilistic models require inputs in terms of a set of independent and identically distributed (i.i.d) variables. The mapping from the space of i.i.d variables to joint distribution of physical variables requires applying an isoprobabilistic transformation as e.g. the Nataf transform (Liu and Der Kiureghian, 1986), and the Rosenblatt transformation (Rosenblatt, 1952). In the present case, it is most convenient to apply the Rosenblatt transformation, ~~which~~ because it allows more complex conditional dependencies than the Nataf transformation which implies linear correlation. The Rosenblatt transformation maps a vector of n dependent variables $\mathbf{X}$ into a vector of independent components $\mathbf{Y}$ based on conditional relations:

$$\mathbf{X} \to \mathbf{Y} = \begin{pmatrix} F_1(X_1) \\ \vdots \\ F_{k|1,\ldots,k-1}(X_k|X_1,\ldots,X_{k-1}) \\ \vdots \\ F_{n|1,\ldots,n-1}(X_n|X_1,\ldots,X_{n-1}) \end{pmatrix}. \tag{1}$$

Further mapping of $\mathbf{Y}$ to a standard ~~Normal~~ normal space vector $\mathbf{U}$ is sometimes applied, i.e.

$$\mathbf{Y} \to \mathbf{U} = \begin{pmatrix} \Phi^{-1}(Y_1) \\ \ldots \\ \Phi^{-1}(Y_n) \end{pmatrix}. \tag{2}$$

For the currently considered set of variables, the Rosenblatt transformation can be applied in the order defined in Table 1 - i.e., the wind speed is considered independent of other variables, the turbulence is dependent on the wind speed, the wind shear is conditional on both wind speed and turbulence, etc. For any variable in the sequence, it is not necessary that it is dependent on all higher-order variables (it may only be conditional on a few of them or even none), but it is required that it is independent from lower-order variables.

2.3 Defining the ranges of input variables

The choice for ranges of variation of the input variables needs to ensure a balance between two objectives: a) covering as wide a range of potential sites as possible, while b) ensuring that the load simulations produce valid results. To ensure validity of load simulations, the major assumptions behind the generation of the wind field and computation of aerodynamic forces should not be violated, and the instantaneous wind field should have physically meaningful values.

For the case of building a high-fidelity load database, all variables given in Table 1 except the wind speed are uniform, and only the distribution bounds are conditional on other variables as specified by the 2nd and 3rd columns of the table. The bounds

of several variables are conditional on the wind speed, and as shown on Figure 2 they are wider at low wind speeds, meaning that more sample points are needed to cover the space evenly. This dictates that the choice of distribution for the wind speed should provide more samples at low wind speeds. In the present study we have selected a Beta distribution, but other choices as e.g. a truncated Weibull are also feasible.

Table 1. Bounds of variation of the variables considered. All values are defined as statistics over 10-minute reference period.

heightVariable	Lower bounds	Upper bounds	Distr
U	$U \geq 4\,\text{m/s}$	$U \leq 25\,\text{m/s}$	B
σ_u	$\sigma_u \geq 0.025 \cdot U\,(\text{m/s})$	$\sigma_u \leq 0.18\left(6.8 + 0.75U + 3\left(\frac{10}{U}\right)^2\right)(\text{m/s})$	Un
α	$\alpha \geq \alpha_{ref,LB} - 0.23\left(\frac{U_{max}}{U}\right)\left(1 - \left(0.4\log\frac{R}{z}\right)^2\right)$	$\alpha \leq \alpha_{ref,UB} + 0.4\left(\frac{R}{z}\right)\left(\frac{U_{max}}{U}\right)$	Un
L	~~$L \geq \max\{7.5\text{m}, (15\text{m})\cdot|\alpha|^{-2/3}\}$~~ $L \geq \max\{7.5\text{m}, (15\text{m})\cdot|\alpha|^{-2/3}\}$	$L \leq 275\,\text{m}$	Un
Γ	$\Gamma \geq 1$	$\Gamma \leq 5$	Un
$\Delta\varphi_h$	$\Delta\varphi_h \geq -0.1D\left(\frac{5}{U}\right)$	$\Delta\varphi_h \leq \min\{60°\sin|\phi|, \quad 1.0D\left(\frac{5}{U}\right)^2\}$	Un
$\bar{\varphi}_h$	$\bar{\varphi}_h \geq -10°$	$\bar{\varphi}_h \leq 10°$	Un
$\bar{\varphi}_v$	$\bar{\varphi}_v \geq -10°$	$\bar{\varphi}_v \leq 10°$	Un
ρ	$\rho \geq 1.1\,\text{kg/m}^3$	$\rho \leq 1.35\,\text{kg/m}^3$	Un

Where
- R is the rotor radius, D the rotor diameter;
- $\alpha_{ref,LB} = 0.15, \alpha_{ref,UB} = 0.22$ are reference wind shear exponents at 15m/s wind speed;
- $U_{max} = 25$m/s is the upper bound of the wind speed;
- ϕ is the reference latitude (here chosen as $50°$).

5   The turbulence intensity, $I_u = \sigma_u/U$, upper limit can be written as the IEC-prescribed form (ed. 3, sub-class A) with $I_{ref,A}$=18%, plus a constant (representing the larger expected range of TI, to span different sites) and a term that encompasses low-windspeed sites and regimes which have higher turbulent intensities. This form is basically equivalent to $\sigma_{u,\text{IEC}} + I_{ref,A}U_{\text{cut-in}}[1 + (U_{\text{cut-out}}/U)]$ with $\{U_{\text{cut-in}}, U_{\text{cut-out}}\}=\{\cancel{3}4,25\}$m/s. The bounds for turbulence intensity as function of mean wind speed are shown on Figure 2. The limits on shear exponent were chosen following the derivations and findings of Kelly et al.

10 (2014) for $P(\alpha|U)$, expanding on the established $\sigma_\alpha(U)$ form to allow for a reasonably wide and inclusive range of expected cases, and also accounting for rotor size per height above ground. This includes an upper bound which allows for enhanced shear due e.g. to lower-level jets and terrain-induced shear; the lower bound also includes the R/z dependence, but does not expand the space to the point that it includes jet-induced negative shear (these are generally found only in the top portion of the rotor). The condition $L > \max\{7.5\text{m}, (15\text{m})|\alpha|^{-2/3}\}$ arises from consideration of the relationship between L, α, σ_u, and

15 ε; small shear tends to correlate with larger motions (as in convective well-mixed conditions), as $L \simeq zI_u/\alpha$ (Kelly, 2018) . The minimum scale (7.5 m) and proportionality constant (15 m) are set to allow a wide range of conditions (though most

sites will likely have a scaling factor larger than 15 m). The maximum Mann-model length scale is chosen based on the limits of where the model can be fitted to measured spectra; this is dictated also by the limits of stationarity in the atmospheric boundary layer (and applicability of Taylor's hypothesis). The range of Γ is also dictated by the minimum expected over non-complex terrain within reasonable use of the turbulence model (smaller Γ might occur for spectra fitted at low heights over hills, but such spectra should be modelled in a different way, as in e.g. Mann (2000)). The range of veer is limited in a way analogous to shear exponent, i.e. it has a basic $1/U$ dependence; this range also depends upon the rotor size, just as $(dU/dz)|_{\text{rotor}} = \alpha D/U$ (Kelly and van der Laan, 2018). The limits for $\Delta\varphi_h$ above peak follow from the limits on α, while for unstable conditions ($\Delta\varphi_h < \Delta\varphi_{h,\text{peak}}$, e.g. all $\Delta\varphi_h < 0$) then the veer limit follows a semi-empirical form based on observed extremes of $\partial\varphi_h/\partial z$. For the remaining variables, $\bar{\varphi}_h$, $\bar{\varphi}_v$, and ρ, the bounds are chosen arbitrarily such that they are wide enough to encompass the values typically used in a design load basis.

2.4 ~~Reference high-fidelity load database~~Sampling procedure

Building a large database with high-fidelity load simulations covering the entire variable space is a central task in the present study as such a database can serve several purposes:

1) be directly used as a site assessment tool by probability-weighting the relative contribution of each point to the design loads;

2) serve as an input for calibrating simplified models~~such as orthogonal-polynomial based expansions;~~

3) ~~be used as reference for the performance of load models calibrated by other means~~, i.e., surrogate models and response surfaces.

Characterizing the load behaviour of a wind turbine over a range of input conditions requires an experimental design covering the range of variation of all variables with sufficient resolution. In the case of having more than 3-4 dimensions, a full factorial design with multiple levels quickly becomes impractical due to the exponential increase in the number of design points as function of number of dimensions. Therefore, in the present study we resort to a Monte Carlo (MC) simulation as the main approach for covering the joint distribution of wind conditions. For assuring better and faster convergence, we use the low-discrepancy Halton sequence in a quasi-Monte Carlo approach (Caflisch, 1998). ~~Figure ?? shows an experimental design based on Halton sequence, compared to crude Monte Carlo and Latin Hypercube designs (Mckay et al., 2000).~~ While a crude Monte Carlo integration has a convergence rate proportional to the square root of the number of samples N, i.e., the mean error $\bar{\varepsilon} \propto N^{-0.5}$, the convergence rate for a ~~pseudo-Monte~~ quasi-Monte Carlo with a low-discrepancy sequence results in $\bar{\varepsilon} \propto N^{-\lambda}$, $0.5 \leq \lambda \leq 1$. For low number of dimensions and smooth functions, the ~~pseudo-Monte~~ quasi-Monte Carlo sequences show a significantly improved performance over the Monte Carlo, e.g. $\lambda \to 1$, however for multiple dimensions and discontinuous functions the advantage over crude Monte Carlo is reduced (Morokoff and Caflisch, 1995). Nevertheless, even for the full 9-dimensional problem discussed here, it is expected that $\lambda \approx 0.6$ (Morokoff and Caflisch, 1995), which still means about an order of magnitude advantage, e.g., 10^4 ~~pseudo-Monte~~ quasi-Monte Carlo samples should result in about
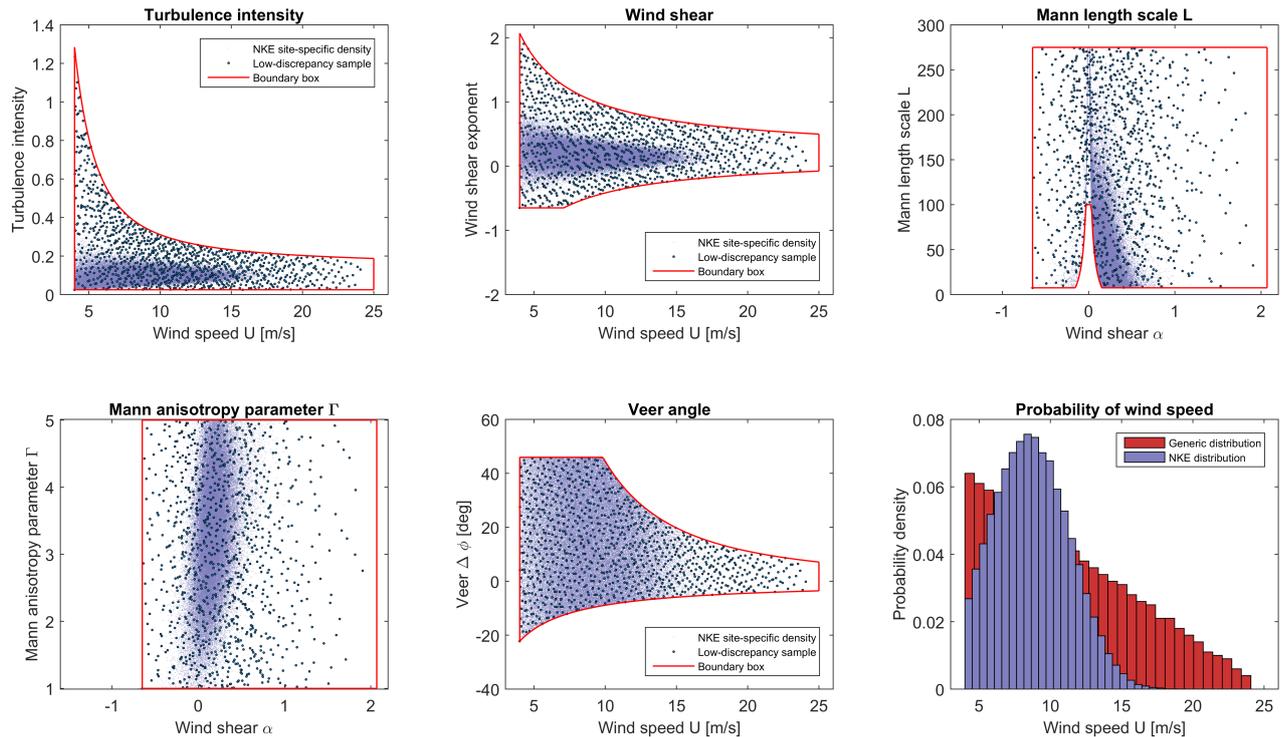
Figure 2. Sample distributions obtained using 1024 low-discrepancy points within a 6-dimensional variable space $\{U, I_u, \alpha, \Delta\phi_h, L, \Gamma\}$. Here U is Beta-distributed, while the other variables are uniformly distributed within their ranges. Solid lines show the sampling space bounds which are curved due to conditional dependencies. Blue shading shows the site-specific variable distribution for the NKE reference site (site 0, c.f. Table 5/Section ??6.1).

the same error as 10^5 crude Monte Carlo samples. A disadvantage of the quasi-random sequences is that their properties typically deteriorate in high-dimensional problems, where periodicity and correlation between points in different dimensions may appear (Morokoff and Caflisch, 1995). However, such behaviour typically occurs when more than 20-25 dimensions are used. In the present problem the dimensionality is limited by the computational requirements of the surrogate models and the

5   aeroelastic simulations used to train them. Therefore the behaviour of quasi-random sequences in high dimensions does not have implications for the present study. The Halton sequence is applied by taking consequentially all points in the quasi-random series without omission and without repetitions, starting from the second point. The first point in the sequence is discarded as it contains zeros (i.e., the lower bounds of the interval $[0,1]$) in all dimensions, which corresponds to zero joint probability for the input variables $\mathbf{X}$.

10   ~~Comparison of several simulation-based experimental design approaches. Examples show random (Monte Carlo and Latin Hypercube) or pseudorandom (Halton sequence) samples of size 100 drawn from a uniform distribution within a unit hypercube.~~

2.5 Database specification

A large-scale generic load database is generated ~~using~~ in order to serve as a training data set for the load mapping functions. The point sampling is done using a Halton low-discrepancy ~~sample points~~ sequence within the 9-dimensional variable space defined in section ~~??~~ 2.4 (Figure 2 shows the bounds for the first 6 variables). The database setup is the following:

- Up to 10^4 ~~pseudo-random~~ quasi-random MC sample points in the interval $[0, 1)$ are generated, following a low-discrepancy sequence for obtaining evenly distributed points within the parametric space.

- The physical values of the stochastic variables for all quasi-MC samples are obtained by applying a Rosenblatt transformation using the conditional distribution bounds given in Table 1 and using uniform distribution density, except for the wind speed for which a Beta distribution is used.

- For each sample point, eight simulations, with ~~1h~~ 3800s duration each, are carried out. The first 200s of the simulations are discarded in order to eliminate simulation run-in time transients, and the output is 3600s (1h) of load time series from each simulation.

- The ~~simulation parameters~~ Mann model simulation parameters (L, Γ, $\alpha\epsilon^{2/3}$) which determine the turbulence intensity are tuned to match the required 10-minute turbulence statistics (1 h statistics are slightly different due to longer sampling time).

- Each 1h time-series ~~can be~~ is split into six 10-minute series, which on average will have the required statistics. This leads to a total of 48 10-minute time series for each quasi-MC sample point.

- Simulation conditions are kept stationary over each 1 h simulation period.

- The DTU 10MW reference wind turbine model (Bak et al., 2013), with the basic DTU Wind Energy controller (Hansen and Henriksen, 2013), is used in the ~~HAWC2~~ Hawc2 aeroelastic software (Larsen and Hansen, 2012).

2.6 ~~Postprocessing and analysis~~

By choosing to run 1h simulations followed by splitting up of the time series instead of directly simulating 10-minute periods, we want to capture some of the low-frequency fluctuations generated by the Mann model turbulence, especially at larger turbulence length scales. When we generate a longer turbulence box, it includes more of these low-frequency variations, which in fact introduce some degree of non-stationarity when looking at 10-minute windows.

2.5.1 ~~Postprocessing~~

3 Post-processing and analysis

3.1 Time series postprocessing and cycle counting

The main quantities of interest from the load simulation output are the short-term (10-minute) fatigue damage-equivalent loads (DEL), and the 10-minute extremes (minimum or maximum, depending on the load type). For each load simulation, four statistics (mean, standard deviation, minimum and maximum values) are calculated for each load channel. For several selected load channels, the 1 Hz DEL for a reference period T_{ref} are estimated using the expression

$$S_{\text{eq}} = \left[\sum \frac{n_i S_i^m}{N_{\text{ref}}} \right]^{1/m} \tag{3}$$

where $N_{\text{ref}} = f \cdot T_{\text{ref}}$ is a reference number of cycles (~~$N_{\text{ref}} = 3600$~~ $N_{\text{ref}} = 600$ for obtaining 1 Hz-equivalent DEL over a ~~1~~10 ~~h~~ min period), S_i are load range cycles estimated using a rainflow counting algorithm (Rychlik, 1987), and n_i are the number of cycles observed in a given range. For a specific material with fatigue properties characterized by an S-N curve of the form $K = N \cdot S^m$ (where K is the material-specific Wöhler constant), the fatigue damage D accumulated over one reference period equals

$$D(T_{\text{ref}}) = \frac{N_{\text{ref}}}{K} S_{\text{eq}}^m. \tag{4}$$

3.1.1 ~~Definition of lifetime damage-equivalent loads~~

3.2 Definition of lifetime damage-equivalent loads

Obtaining site-specific lifetime fatigue loads from a discrete set of simulations requires integrating the short-term damage contributions over the long-term joint distribution of input conditions. The lifetime damage-equivalent fatigue load is defined as

$$S_{\text{eq,lifetime}} = \left[\int_{\mathbf{X} \in R^9} [S_{\text{eq}}(\mathbf{X})]^m f(\mathbf{X}) d\mathbf{X} \right]^{1/m} \tag{5}$$

where $f(\mathbf{X})$ is the joint distribution of the multidimensional vector of input variables $\mathbf{X}$. With the above definition, $S_{\text{eq,lifetime}}$ is a function of the expected value of the short-term equivalent loads conditional on the distribution of environmental variables. The integration in eq. (5) is typically performed numerically over a finite number of realizations drawn from the joint distribution of the input variables, e.g. by setting up a look-up table or carrying out a Monte Carlo simulation. Thus the continuous problem is transformed into a discrete one:

$$S_{\text{eq,lifetime}} = \left[\sum_{i=1}^{N} ~~[S_{\text{eq}}(\mathbf{X}_i)]^m p(\mathbf{X}_i)~~ [S_{\text{eq}}(\mathbf{x}_i)]^m p(\mathbf{x}_i) \right]^{1/m} , \tag{6}$$

where $\mathbf{X}_i, i = 1 \ldots N \mathbf{x}_i, i = 1 \ldots N$, is the ith realization of $\mathbf{X}$ out of N total realizations, and $p(\mathbf{X}_i)$ $p(\mathbf{x}_i)$ is the relative, discretized probability of $\mathbf{X}_i \mathbf{x}_i$, which is derived by weighting the joint pdf values of $\mathbf{X}$ so that they satisfy the condition $\sum_{i=1}^{N} p(\mathbf{X}_i) = 1 \sum_{i=1}^{N} p(\mathbf{x}_i) = 1$. For a standard Monte Carlo simulation, each realization is considered to be equally likely, and $p(\mathbf{X}_i) = 1/N p(\mathbf{x}_i) = 1/N$.

3.2.1 ~~Obtaining site-specific results using Importance Sampling~~

3.3 Uncertainty estimation and confidence intervals

With the present problem of evaluating the uncertainty in aeroelastic simulations, for any specific combination of environmental conditions, $\mathbf{x}_i$, there will be uncertainty in the resulting damage-equivalent loads, $S_{eq}(\mathbf{x}_i)$. Part of this uncertainty is statistical by nature and is caused by realization-to-realization variations in the turbulent wind fields used as input to the load simulations. This uncertainty is normally taken into account by carrying out load simulations with multiple realizations (seeds) of turbulence inputs. Confidence intervals reflecting such an uncertainty can be determined in a straightforward way using the bootstrapping technique (Efron, 1979). Its main advantage is robustness and no necessity for assuming a statistical distribution of the uncertain variable. With this approach, each function realization is given an integer index, e.g., from 1 to N for N function realizations. Then, a "bootstrap" sample is created by generating random integers from 1 to N, and, for each random integer, assigning the original sample point with the corresponding index, as part of the new bootstrap sample. Since the generation of random integers allows number repetitions, the bootstrap sample will in most cases differ from the original sample. To obtain a measure of the uncertainty in the original sample, a large number of bootstrap samples are drawn, and the resultant quantity of interest (e.g. the lifetime fatigue load) is computed for each of them. Then, the empirical distribution of the set of outcomes is used to define the confidence intervals. If M bootstrap samples have been drawn, and R is the set of outcomes ranked by value in ascending order, then the (confidence interval) bounds for a confidence level c_ℓ are

$$\left\{ CI^-_{S_{eq,\text{lifetime}}}(c_\ell), \, CI^+_{S_{eq,\text{lifetime}}}(c_\ell) \right\} = \left\{ R_{[c_\ell M/2]}, \, R_{[(1-c_\ell/2)M]} \right\} \tag{7}$$

where the square brackets $[x]$ indicate the integer part of x, and $R_{[x]}$ means the value in R with rank equal to $[x]$. In the present study, bootstrapping is applied by generating independent bootstrap samples each with size equal to the entire data set. Both the sample points and the turbulence seed numbers are shuffled, meaning that the resulting confidence intervals should account for both the statistical uncertainty due to finite number of samples, and the uncertainty due to seed-to-seed variation. Note that these two uncertainty types are the only ones assumed, for the confidence intervals; reducing the CI by creating a large number of model realizations does not eliminate other model uncertainties, nor does it remove uncertainties in the input variables.

4 Load mapping functions

In this section we present five different approaches which can be used to map loads from a high-fidelity database to integrated site-specific design loads:

   1) Importance sampling,

   2) Nearest-neighbor interpolation,

   3) Polynomial chaos expansion,

   4) Universal Kriging, and

5   5) Quadratic response surface.

The first two methodologies carry out a direct numerical integration over the high-fidelity database presented in Section 2.5, while the latter three are machine learning models which are trained using the same database. Despite the different nature of the functions, they serve the same purpose and for brevity we will refer to all of them as "surrogate models".

4.1 Importance Sampling

10 Figure 2 ~~shows~~ showed the distributions of the first 6 input variables from our high-fidelity database (~~§~~Section 2.5), along with the site-specific distributions for reference site 0 (c.f. Table 5 for site list).

One of the simplest and most straightforward (but not necessarily most precise) ways of carrying out the integrations needed to obtain predicted statistics is to use Importance Sampling ('IS'), where probability weights are applied on each of the database sample points ~~(see e.g. Ditlevsen and Madsen, 1996). The~~ (Ditlevsen and Madsen, 1996). The IS method and

15 various modifications of it are commonly used in wind energy-related applications (e.g. Choe et al., 2015; Graf et al., 2018). In the classical definition of IS, the integration (importance sampling) function for determining the expected value of a function ~~$g(\mathbf{x})$~~ $g(\mathbf{X})$ is given by

$$\underset{\sim}{\mathbb{E}[g(\mathbf{x})]}\mathbb{E}[g(\mathbf{X})] = \frac{1}{N}\sum_{i=1}^{N}g(\underset{\sim}{\mathbf{x}}\mathbf{X})\frac{f(\mathbf{x}_i)}{h(\mathbf{x}_i)}\frac{f(\mathbf{X}_i)}{h(\mathbf{X}_i)}, \tag{8}$$

where in our application

20 -

- $i = 1 \ldots N$ is the sample point number;

- ~~$\mathbf{x}_i = [x_{1,i}, x_{2,i}, \ldots, x_{9,i}]$~~

    • $\mathbf{X}_i = [x_{1,i}, x_{2,i}, \ldots, x_{9,i}]$ is a 9-component vector array specifying the values of the 9 environmental variables considered at sample point i;

25 - ~~$f(\mathbf{x}_i) = f(x_{1,i}) \cdot f(x_{2,i}|x_{1,i}) \cdot \ldots \cdot f(x_{9,i}|x_{8,i}, \ldots, x_{1,i})$~~

    • $f(\mathbf{X}_i) = f(x_{1,i}) \cdot f(x_{2,i}|x_{1,i}) \cdot \ldots \cdot f(x_{9,i}|x_{8,i}, \ldots, x_{1,i})$ is the joint pdf of sample point i according to the **site-specific** probability distribution; and

~~- $h(\mathbf{x}_i) = h(x_{1,i}) \cdot h(x_{2,i}|h_{1,i}) \cdot \ldots \cdot h(x_{9,i}|x_{8,i},\ldots,x_{1,i})$~~

- $h(\mathbf{X}_i) = h(x_{1,i}) \cdot h(x_{2,i}|h_{1,i}) \cdot \ldots \cdot h(x_{9,i}|x_{8,i},\ldots,x_{1,i})$ is the joint pdf of sample point i according to the **generic** probability distribution used to generate the database for the 9 variables.

Based on the above, it is clear that only points in the database which also have a high probability of occurrence in the site-specific distribution will have a significant contribution to the lifetime load estimate. This could be considered as a non-standard application of the IS approach, because typically the IS sample distribution is chosen to maximize the probability density of the integrand. In the present case, this objective can be satisfied only approximately, and only in cases where the number of IS samples is smaller than the total number of database samples ($N_{\text{IS}} < N$). Under these conditions, the importance sampling weights ($f(\mathbf{X}_i)/h(\mathbf{X}_i)$ from Eq. 8) can be evaluated for all points in the database. However, the approach adopted in the present paper is to include only the N_{IS} points with the highest weights (as shown in Section 6.1). ~~Therefore, the IS procedure has relatively slow convergence compared to e.g. a pseudo-MC simulation. Figure 5 shows an example of the convergence of an IS integration for reference site 0, based on using a high-fidelity database with 10^4 points . Convergence of an importance sampling (IS) calculation of the blade root moment from the hi-fi database, towards site-specific lifetime fatigue loads for reference site (site 0).~~

4.1.1 ~~Obtaining site-specific results using multi-dimensional interpolation~~

4.2 Multi-dimensional interpolation

Estimating an expected function value with a true multi-dimensional interpolation from the high-fidelity database would require finding a set of neighboring points which form a convex polygon. For problem dimensions higher than 3, this is quite challenging due to the non-structured sample distribution. However, it is much easier to find a more crude approximation by simply finding the database point closest to the function evaluation point in a nearest-neighbor approach. This is similar to the table look-up technique often used with structured grids; the denser the distribution of the sample points is, the closer will the results be to an actual Monte Carlo simulation. Finding the nearest neighbor to a function evaluation point requires determining the distances between this point and the rest of the points in the sample space. This is done most consistently in a normalized space, i.e. where the input variables have equal scaling. The cdf (cumulative distribution function) of the variables is an example of such a space, as all cdf's have the same range of $(0,1)$. Thus, the normalized distance between a new evaluation point and an existing sample is computed as the vector norm of the (e.g. 9-dimensional vector) differences between the marginal cdf for the two points:

$$|x| = \sqrt{\mathbf{D}^T \mathbf{D}} \tag{9}$$

where $\mathbf{D} = \mathbf{Y} - \hat{\mathbf{Y}}$ is the difference between the current evaluation point $\mathbf{Y}$ and the existing sample points in the reference database, $\hat{\mathbf{Y}}$. The vector $\mathbf{Y}^T = [F_1(X_1), F_2(X_2|X_1),\ldots,F_n(X_n|X_1,\ldots,X_{n-1})]$ consists of the marginal cdf functions of the input variables $\mathbf{X}$ as obtained using a Rosenblatt transformation.

Since some of the input variables may have significantly bigger influence on the result than other variables, it may be useful to weight the cdf of different variables according to their importance (e.g. by making the weights proportional to the variable sensitivity indices; see Section 4.1).

4.2.1 ~~Uncertainty estimation and confidence intervals~~

5 ~~With~~

4.3 Polynomial chaos expansion

Polynomial Chaos Expansion (PCE) is a popular method for approximating a stochastic function of multiple random variables using an orthogonal polynomial basis. For the present problem ~~of evaluating the uncertainty in aeroelastic simulations, for any specific combination of environmental conditions, $\mathbf{X}_i$, there will be uncertainty in the resulting damage-equivalent loads,~~

10 ~~$S_{eq}(\mathbf{X}_i)$. Part of this uncertainty is statistical by nature and is caused by realization-to-realization variations in the turbulent wind fields used as input to the load simulations. This uncertainty is normally taken into account by carrying out load simulations with multiple realizations (seeds) of turbulence inputs. Provided that enough load realizations have been generated, the seed-to-seed uncertainty can be characterized by the sample statistics (mean and standard deviation) and with an assumption about the statistical distribution. For the damage-equivalent loads which are non-negative by definition, a log-normal distribution~~

15 ~~is a suitable choice; then the confidence intervals for $(S_{eq,lifetime})^m$ for confidence level α can be found by :~~

$$CI^+_{(S_{eq,lifetime})^m}(\alpha) = \exp\left(\mu_{LN} + \Phi(1 - \alpha/2) \cdot \sigma_{LN}\right)$$

$$CI^-_{(S_{eq,lifetime})^m}(\alpha) = \exp\left(\mu_{LN} - \Phi(\alpha/2) \cdot \sigma_{LN}\right)$$

, using a Wiener-Askey Generalized PCE (Xiu and Karniadakis, 2002) employing Legendre polynomials is considered most suitable for any (scaled) variable $\xi \in [-1, 1]$. Because Legendre polynomials $P_n(\xi)$ are orthogonal with respect to a uniform

20 probability measure, the PCE can conveniently be applied to the cumulative distribution functions of the variables $\mathbf{X}$ which are defined in the interval $[0, 1]$. Then

$$\xi_i = 2F(X_i) - 1, \tag{10}$$

where

$$\mu_{LN} = \log\left(\frac{\mathrm{E}[(S_{eq,lifetime})^m]}{\sqrt{1 + \frac{\sigma^2_{(S_{eq,lifetime})^m}}{\mathrm{E}[(S_{eq,lifetime})^m]^2}}}\right) \quad ; \quad \sigma_{LN} = \sqrt{\log\left(1 + \frac{\sigma^2_{(S_{eq,lifetime})^m}}{\mathrm{E}[(S_{eq,lifetime})^m]^2}\right)}$$

25 ~~are the parameters of the log-normal distribution,~~ $F(X_i)$ is the cumulative distribution function of a variable $X_i \in \mathbf{X}, i = 1, \ldots, M$. The Legendre polynomial coefficients can be generated using the recurrence relation

$$(n+1)P_{n+1}(\xi) = (2n+1)\xi P_n(\xi) - nP_{n-1}(\xi) \tag{11}$$

14

where the first two entries, $P_0(\xi) = 1$ and ~~Φ denotes the standard Normal cumulative distribution function. It should be noted that the confidence intervals defined above are given in terms of~~ $P_1(\xi) = \xi$, serve for initialization. The aim of using PCE is to represent a scalar quantity $S = g(\mathbf{X})$ in terms of a truncated sequence $\tilde{S}(\mathbf{X}) + \varepsilon$, where ε is a zero-mean residual term. With this definition, the multivariate generalized PCE of dimension M and maximum degree p is given by

$$\tilde{S}(\boldsymbol{\xi}) = \sum_{j=0}^{N_p-1} S_j \Psi_{\boldsymbol{\gamma},j}(\boldsymbol{\xi}); \tag{12}$$

here $\Psi_{\boldsymbol{\gamma}}$ are multivariate orthogonal polynomials composed of the product of univariate polynomials having (nonnegative integer) orders defined by the vector $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_M]$, with the total of orders being constrained by the degree: $\sum_{i=1}^{M} \gamma_i \leq p$. The unknown coefficients $S_j \in \mathbf{S} = [S_1,\ldots,S_{Np}]$ need to be determined, and $\boldsymbol{\xi} = [\xi_1,\ldots,\xi_M]$ are functions of $\mathbf{X}$ as defined in eq. (10). Training the PCE model amounts to determining the vector of coefficients, $\mathbf{S}$. For a more detailed explanation of the ~~lifetime DEL raised to~~ training process, as well as the basic PCE theory, the ~~power of the Wøhler slope m, meaning that they actually reflect the range of variation of the lifetime fatigue damage, as visible from eq. . Using the sample information, confidence intervals can also be determined in a straightforward way using the bootstrapping technique (Efron, 1979). Its main advantage is robustness and no necessity for assuming a statistical distribution of the uncertain variable . With this approach, each function realization is given an integer index, e. g., from 1 to~~ reader is referred to Appendix A (and further to Xiu and Karniadakis, 200?

.

4.4 Universal Kriging with polynomial chaos basis functions

Kriging (Sacks et al., 1989; Santher et al., 2003) is a stochastic interpolation technique which assumes the interpolated variable follows a Gaussian process. A Kriging model is described (Sacks et al., 1989) by

$$Y(\mathbf{X}) = \mathbf{f}(\mathbf{X})^T \boldsymbol{\beta} + Z(\mathbf{X}), \tag{13}$$

where for N ~~for N function realizations. Then, a~~ evaluation samples and an M-dimensional problem, $\mathbf{X}$ represents an $M \times N$ matrix of input variables and $Y(\mathbf{X})$ is the output vector. The term $\mathbf{f}(\mathbf{X})^T \boldsymbol{\beta}$ is the mean value of the Gaussian process (a.k.a. the "~~bootstrap~~trend"~~sample is created by generating random integers from 1 to N, and , for each random integer, assigning the original sample point with the corresponding index, as part of the new bootstrap sample. Since the generation of random integers allows number repetitions, the bootstrap sample will in most cases differ from the original sample. To obtain a measure of the uncertainty in the original sample, a large number of bootstrap samples are drawn, and the resultant quantity of interest ()~~ represented as a set of basis functions $\mathbf{f}(\mathbf{X}) = [f_1(\mathbf{X}),\ldots,f_P(\mathbf{X})]$ and regression coefficients $\boldsymbol{\beta} = [\beta_1,\ldots,\beta_P]$, whereas $Z(\mathbf{X})$ is a zero-mean stationary Gaussian process. The (joint) probability distribution of the Gaussian process is characterized by its covariance; for two distinct 'points' $\mathbf{X}$ and $\mathbf{W}$ in the sample domain the covariance is defined as

$$V(\mathbf{W},\mathbf{X}) = \sigma^2 R(\mathbf{W},\mathbf{X},\boldsymbol{\theta}), \tag{14}$$

where σ^2 is the overall process variance which is assumed to be constant, and $R(\mathbf{W},\mathbf{X},\boldsymbol{\theta})$ is the correlation between $Z(\mathbf{X})$ and $Z(\mathbf{W})$. The hyperparameters $\boldsymbol{\theta}$ define the correlation behavior, in terms of e.g. ~~the lifetime fatigue load)is computed for~~

each of them. Then, the empirical distribution of the set of outcomes is used to define the confidence intervals. If M bootstrap samples have been drawn, and R is the set of outcomes ranked by value in ascending order, then the bounds for confidence level α equal

$$CI^+_{S_{eq,lifetime}}(\alpha) = R_{[(1-\alpha/2)M]}$$

5 $$CI^-_{S_{eq,lifetime}}(\alpha) = R_{[\alpha M/2]}$$

where the square brackets $[x]$ indicate the integer part of x, and $R_{[x]}$ means the value in R with rank equal to $[x]$. correlation length scale(s). Since the mean and variance of the Gaussian field can be expressed as functions of θ (this is shown in details in Appendix A), the calibration of the Kriging model amounts to determining the trend coefficients and obtaining an optimal solution for θ.

10 Note that the confidence intervals estimated with the above procedure describe only the statistical uncertainty due to The functional form of the mean field $\mathbf{f}(\mathbf{X})^T\boldsymbol{\beta}$ is identical to the generalized PCE defined in eq. (A8), meaning that the PCE is a possible candidate model for the finite number of random samples, and due to seed-to-seed variation. Narrowing them down by e.g. creating a large number of model realizations does not eliminate other model uncertainties as well as uncertainties in the input variables. mean in a Kriging interpolation. We adopt this approach and define the Kriging mean as a PCE with

15 properties as described in section 4.3. A suitable approach for tuning the Gaussian field statistics is to find the values of $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$ which maximize the likelihood of the training set variables $\mathbf{Y}$, i.e. minimize the model error in a least-squares sense (Lataniotis et al., 2015). This is described in Appendix A.

5 Reference sites

The site-specific load calculation methods presented in this study are validated against a set of reference site-specific load
20 calculations on a number of different sites which cover a wide fraction of the variable domain included within the high-fidelity database. In order to show a realistic example of situations where a site-specific load estimation is necessary, the majority of the sites chosen are characterized with conditions which slightly exceed the standard conditions specified by a certain type-certification class. Exceptions are site 0 which has the most measured variables available and is therefore chosen as a main reference site, and the "sites" representing standard IEC class conditions. The IEC classes are included as test sites as they
25 are described by only one independent variable (mean wind speed). They are useful test conditions as it may be challenging to correctly predict loads as function of only one variable using a model based on up to 9 random variables. The list of test sites is given in Table 5. Reference sites used for validation of the site-specific load estimation methods. Site No. Location Terrain Specific condition Variables included 0 Denmark Flat agricultural - $U,\sigma_u,\alpha,L,\Gamma,\Delta\varphi$ 1 Denmark Flat agricultural IIIC exceedance U,σ_u,α 2 N. Denmark Forested IIIB exceedance U,σ_u,α 3 N. Denmark Forested IA exceedance U,σ_u,α
30 4 N. Denmark Forested IIA exceedance U,σ_u,α 5 USA, Colorado Mountain foothills Low-wind U,σ_u,α 6 USA, Colorado Mountain foothills Low-wind U,σ_u,α IEC IA, NTM - - Standard reference class U IEC IIB, NTM - - Standard reference class U IEC IIIC, NTM - - Standard reference class U IEC IIB, ETM - - Standard reference class U

~~For each site, the joint distributions of all variables are defined in terms of conditional dependencies, and generating simulations of site-specific conditions is carried out using the Rosenblatt transformation, . The conditional dependencies are described in terms of functional relationships between the governing variable and the distribution parameters of the dependent variable, e.g. the mean and standard deviation of~~ The main practical difference between regression- or expansion-type models

5 such as regular PCE and the Kriging approach is how the training sample is used in the model: in pure regression-based approaches the training sample is used to only calibrate the regression coefficients, while in Kriging (and in other interpolation techniques) the training sample is retained and used in every new model evaluation. As a result the Kriging model may have an advantage in accuracy, since the model error tends to zero in ~~the turbulence are modelled as linearly dependent on the wind speed as recommended by the IEC 61400-1 standard, while the mean wind shear is dependent on the mean wind~~

10 ~~speed and on the turbulence, as defined by (Kelly et al., 2014). With this procedure, Pseudo-Monte Carlo samples of the environmental conditions at each site are generated from the respective joint distribution, and fed as input to load simulations. The resulting reference lifetime equivalent loads are then defined as the sample means from the Monte Carlo simulations, while the uncertainty in the lifetime loads is estimated using bootstrapping~~vicinity of the training points; however, this comes at the expense of an increase in the computational demands for new model evaluations. For a Kriging model, a gain in accuracy over

15 the model represented by the trend function will only materialize in problems where there is a noticeable correlation between the residual values at different training points. In a situation where the model error is independent from point to point (as e.g. in the case when the error is only due to seed-to-seed variations in turbulence) the inferred correlation length will tend to zero and the Kriging estimator will be represented by the trend function alone.

4.1 ~~Site locations~~Quadratic response surface

20 ~~Site 0 (also also referred to herein as the reference site) is located at the Nørrekær Enge wind farm in Northern Denmark (Borraccino et al., 2, over flat, open agricultural terrain (Fig. ??) . Nørrekær Enge site and mast. Permissions by DTU Wind Energy. Site 1 is a flat-terrain, near-coastal site at the the National Centre for Wind Turbines at Høvsøre, Denmark (Peña et al., 2016, and shown here in Fig. ??. Høvsøre site and mast. Permissions by DTU Wind Energy. Sites 2 to 4 are based on the wind conditions measured at the Østerild Wind Turbine Test Field which is located in a large forest plantation in North-western Denmark (Hansen et al., 2014)~~

25 ~~(Fig. ??). The fully instrumented mast at Østerild is the blue dot furthest South/Down on the map. Permissions by DTU Wind Energy. Left: Cup anemometer (left) and wind vane (right) at 244 m height, Østerild test site. Permissions by DTU Wind Energy. Due to the forested surroundings of the site, the flow conditions are more complex than those in Nørrekær Enge and Høvsøre. By applying different filtering according to wind direction, three imaginary site climates are generated and considered as sites 2–4.~~

30 ~~Sites 5~~ A quadratic-polynomial response surface (RS) method based on Central Composite Design (CCD) is a reduced-order model which, among other applications, has been used for wind turbine load prediction (Toft et al., 2016). The procedure involves fitting a quadratic polynomial regression ('response surface') to a normalized space of i.i.d. variables, which are derived from the physical variables using an isoprobabilistic transformation—such as the Rosenblatt transformation given in eq. (1) and ~~6 are located at NREL's National Wind Technology Center (NWTC) , near the base of the Rocky Mountain foothills~~

17

just south of Boulder, Colorado (see Fig. ??; c.f. Clifton et al., 2013). Similar to Østerild, directional filtering is applied to ~~(2)~~. The design points used for calibrating the response surface in k dimensions form a combination of a central point, axial points a distance of $\sqrt{k}$ in each dimension, and a 2^k 'factorial design' set where there are two levels (points) per variable dimension located at unit distance from the origin; this is shown in Figure 3 for the case of $k = 2$ variables (dimensions). Due to the structured design grid required, it is not possible to use this approach with the sample points from the high-fidelity database described in section 2; therefore we implement the procedure using an additional set of simulations. The low order of the response surface also prohibits full characterization of the highly nonlinear turbine response as function of mean wind speed using a single response surface. Therefore, multiple response surfaces are calibrated for wind speeds from 4 to 25 m s^{-1} in 1 m s^{-1} steps. This approach may in principle be extended to include additional variables such as turbulence (σ_u), however doing so will reduce the practicality of the ~~NTWC data to split it into two virtual sites—accounting for the different conditions and wind climates from the two ranges of directions considered. Aerial view of NWTC. Permissions by NREL. Map of NWTC. Permissions by NREL.~~

5 ~~Reduced order models~~

~~In this section we present three different reduced order models: 1) Polynomial chaos expansion, 2) Universal Kriging, and 3) Quadratic response surface . All three methodologies are to be calibrated to the database presented in Section 2.5 in order to predict site-specific loads.~~

4.1 ~~Polynomial chaos expansion~~

~~Polynomial Chaos Expansion (PCE)is a popular method for approximating a stochastic function of multiple random variables using an orthogonal polynomial basis. In the classical definition of PCE (Ghanem and Spanos, 1991) the input random variables X are defined on $(-\infty, \infty)$, with Hermite polynomials typically used as~~ procedure as it will require multi-dimensional interpolation between large number of models and the uncertainty may increase. However, due to the ~~polynomial basis.~~[1] ~~Choosing a polynomial basis which is orthogonal to a non-Gaussian probability measure turns the PCE problem into the so-called Wiener-Askey or Generalized chaos, (Xiu and Karniadakis, 2002). For the present problem , a Generalized PCE using Legendre polynomials is considered most suitable as the Legendre polynomials $P_n(\xi)$ are orthogonal with respect to a uniform probability measure in the interval $\xi = [-1, 1]$, which means that the PCE can conveniently be applied on the cumulative distribution functions~~ exponential increase of the number of design points with increasing problem dimension, it is not practical to fit response surface covering all 9 variables considered. Instead, we choose to replace three of the variables ~~X~~

---

[1] ~~In the classical definition of the PC decomposition used in e.g. spectral stochastic finite element methods (Ghanem and Spanos, 1991), the input random variables are Normally distributed (Gaussian), which means that the Hermite polynomials are a suitable Hilbertian basis—since the Hermite polynomials are orthogonal with respect to the Gaussian probability measure. In this case, the properties of the Hermite polynomials dictate that the random variables X are defined on $(-\infty, \infty)$.~~

which are defined in the interval $[0,1]$ so that

$$\xi_i = 2F(X_i) - 1,$$

where $F(X_i)$ is the cumulative distribution function of a variable $X_i \in \mathbf{X}$, $i = 1,\ldots,M$. With this definition, the PCE represents a model applied to a set of transformed variables which due to the applied transformation are independent and identically

5   distributed ('i. i.d.'). Note that eq. and the evaluation of the cumulative distribution in general does not ensure independence - so an appropriate transformation should also account for the dependence between variables. In the present case, it is convenient to apply the Rosenblatt transformation as defined in with relatively low importance (yaw, tilt, and air density) with their mean values. The result is a 6-dimensional problem consisting of 22 different 5-dimensional response surfaces, which require $22 \cdot (1 + 2 \cdot 5 + 2^5) = 946$ design points in total. Analogous to the high-fidelity database, 8h of simulations are carried out

10   for characterizing each design point. The polynomial coefficients of the response surface are then defined using least-squares regression with the same closed-form solution defined by eq. . For the current implementation of PCE, only eq. is required since the expansion is based on the Legendre polynomials, however the transformation to standard Normal space in eq. is used for other procedures, e.g. the response surface model discussed later.

Using the notation defined by Sudret (2008), we consider the family of univariate Legendre polynomials $P_n(\xi)$. A multivariate,

15   generalized PCE with M dimensions and maximum polynomial degree (A8). For any sample point p is defined as the product of univariate Legendre polynomials where the maximum degree is less than or equal to p. The univariate polynomial family for dimension i can be in the central composite design, the corresponding row in the design matrix is defined as

$$P_{\alpha_i}(\xi), \qquad \text{where} \quad i = 1,\ldots,M, \quad \alpha_i \in \mathbb{N}, \quad \sum_{i=1}^{M} \alpha_i \leq p(\alpha_i \geq 0).$$

20   $$\mathbf{\Psi}_p = \left[\{1\}, \quad \{U_1,\ldots U_n\}, \quad \{U_1^2,\ldots U_n^2\}, \quad \{U_i \cdot U_j,\, i = 1\ldots n,\, j = 1\ldots(i-1)\}\right] \qquad (15)$$

The multivariate polynomial of dimension M is then defined as

$$\Psi_{\boldsymbol{\alpha}} = \prod_{i=1}^{M} P_{\alpha_i}(\xi_i)$$

The total number of polynomials of this type is

$$N_p = \frac{(M+p)!}{M!\,p!}$$

25   The aim of using PCE is to represent a scalar quantity $S = g(\mathbf{X})$ in terms of a truncated sequence $\tilde{S}(\mathbf{X}) + \varepsilon$ where ε is a zero-mean residual term. With this definition, the multivariate generalized PCE of dimension M and maximum degree p is given by

$$\tilde{S}(\boldsymbol{\xi}) = \sum_{j=0}^{N_p - 1} S_j \Psi_{\boldsymbol{\alpha},j}(\boldsymbol{\xi})$$

19
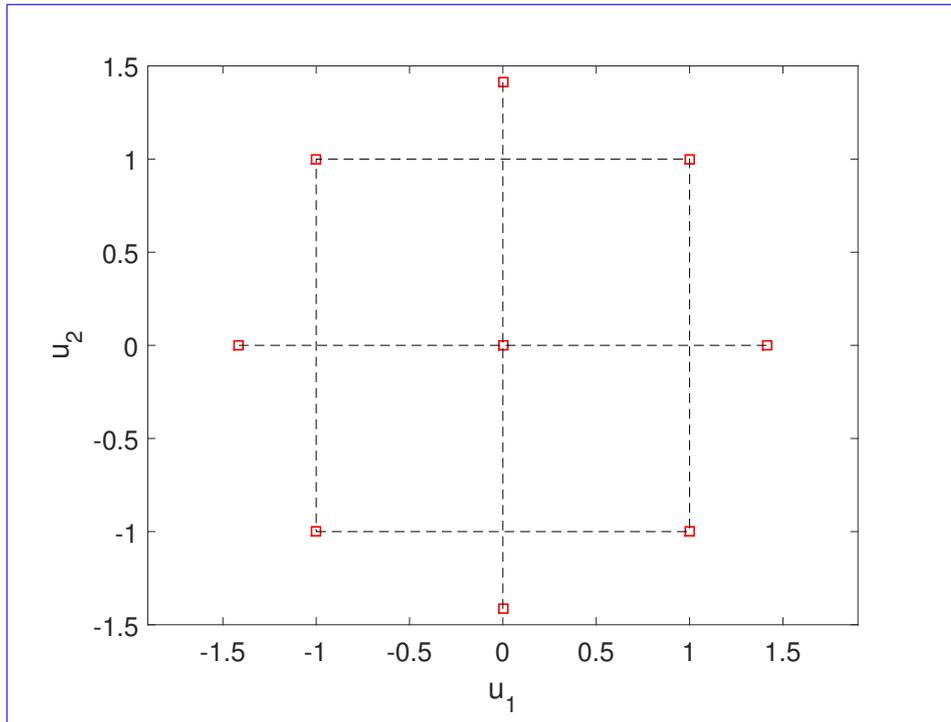
Figure 3. Example of a rotatable Central Composite Design (CCD) in a 2-dimensional standard normal space $[u_1, u_2]$. The CCD consists of a central point, a 2^k 'factorial design' with 2 levels and $k = 2$ dimensions, and axial points at distance $u = \sqrt{2}$, meaning that all the outer points lie on a circle.

~~where $S_j \in \mathbf{S} = [S_1, \ldots, S_{Np}]$ are unknown coefficients which need to be determined, and $\boldsymbol{\xi} = [\xi_1, \ldots \xi_M]$ are functions of~~ $\mathbf{U}$ are standard normal variables derived from the physical variables $\mathbf{X}$ ~~as defined in . The most straightforward way of determining $\mathbf{S}$ is minimizing the variance of the residual ε using a least-squares regression approach:~~ by an isoprobabilistic transformation.

4.1 Sensitivity indices

We use the global Sobol indices, (Sobol, 2001), for evaluating the sensitivity of the response with respect to input variables. Having trained a surrogate model, the total Sobol indices can be computed efficiently by carrying out a Monte Carlo simulation on the surrogate. For number of dimensions equal to M (e.g. $M = 9$ in the present study) and for N (quasi) Monte-Carlo samples the required experimental design represents an $N \times 2M$ matrix. This is divided into two $N \times M$ matrices, $\mathbf{A}$ and $\mathbf{B}$. Then, for each dimension i, $i = 1 \ldots M$, a third matrix $\mathbf{AB}_i$ is created by taking the ith column of $\mathbf{AB}_i$ equal to the ith column from $\mathbf{B}$, and all other columns taken from $\mathbf{A}$. The load surrogate is then evaluated for all three matrices, resulting in three model estimates: $f(\mathbf{A})$, $f(\mathbf{B})$, and $f(\mathbf{AB}_i)$. By repeating this for $i = 1 \ldots M$, simulation-based Sobol' sensitivity indices

can be estimated as

$$\mathbf{S}SU_i = \min \frac{1}{N_e}\sum_{i=1}^{N_e} g \frac{1}{N}\sum_{j=1}^{N} f(\mathbf{B})_j \left(f(\mathbf{AB}_i)_j - \sum_{j=0}^{N_p-1} S_j \Psi_{\boldsymbol{\alpha},j}(\boldsymbol{\xi}^{(i)})^2 f(\mathbf{B})_j \right) \tag{16}$$

where N_p is the number of polynomial coefficients in the PCE and N_e is the number of sampling points in the experimental design. For this purpose, a design experiment has to be set up and the so-called design matrix $\boldsymbol{\Psi}$ needs to be constructed:

$$\boldsymbol{\Psi}_{ij} = \Psi_{\boldsymbol{\alpha},j}(\xi^{(j)}); \qquad i = 1,\ldots,N_e, \qquad j = 1,\ldots,N_p.$$

Under the condition that the residuals are (approximately) Normally-distributed, the solution to equation is given by

$$\mathbf{S} = (\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi})^{-1}\cdot\boldsymbol{\Psi}^{\mathrm{T}}\cdot\mathbf{y},$$

with $\mathbf{y} = g(\mathbf{x}^{(i)})$ being a vector with the outcomes of the functional realizations obtained from the design experiment, where $i = 1\ldots N_e$ $j = 1\ldots N$ is the row index in the design matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{AB}_i$ (Saltelli et al., 2008). For the problem discussed in the present study, it was sufficient to use approximately 500 MC samples per variable dimension in order to compute the total Sobol indices.

The solution of eq. requires that the so-called information matrix ($\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}$) is well-conditioned, which normally requires that the number of collocation points N_e is significantly larger than the number of expansion coefficients N_p. Subsequently

4.2 Model reduction

For any polynomial-based regression model which includes dependence between variables, the problem grows steeply in size when the number of dimensions, M and, and the maximum polynomial order, p, increase. In such situations, it may be desirable to limit the number of active coefficients by carrying out a Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1996), which regularizes the regression by penalizing the sum of the absolute value of regression coefficients. For a PCE model, the objective function using a LASSO regularization is

$$\mathbf{S} = \min\left\{ \left| \left| \frac{1}{2N_e}\sum_{i=1}^{N_e}\left[g(\boldsymbol{\xi}^{(i)}) - \sum_{j=0}^{N_p-1} S_j \Psi_{\boldsymbol{\alpha},j}(\boldsymbol{\xi}^{(i)}) \sum^{N_p-1} S_j \Psi_{\boldsymbol{\gamma},j}(\boldsymbol{\xi}^{(i)}) \right]^2 + \lambda \sum_{j=0}^{N_p-1} |S_j| \right| \right| \right\} \tag{17}$$

where λ is a positive regularization parameter; larger values of λ increase the penalty and reduce the absolute sum of the regression coefficients, while $\lambda = 0$ is equivalent to ordinary least-squares regression.

4.2.1 Convergence of PCE

In the present study, the LASSO regularization is used on the PCE-based models to reduce the number of coefficients.

One useful corollary of the orthogonality in the PCE polynomial basis is that the contribution of each individual term to the total variance of the expansion (i.e. the individual Sobol indices) can be easily computed based on the coefficient values (see

21

Appendix A). This property can be used for eliminating polynomials which do not contribute significantly to the variance of the output, thus achieving a sparse, more computationally efficient reduced model. By combining the variance truncation and the LASSO regression technique in eq. (17), a model reduction of an order of magnitude or more can be achieved. For example, for the 9-dimensional PCE of order 6 discussed in Section 5.3, using LASSO regularization parameter $\lambda = 1$ and retaining the polynomials which have a total variance contribution of 99.5%, resulted in a reduction of the number of polynomials from 5005 to about 200.

5   Model training and performance

5.1   Model convergence

We assess the convergence of PCE by calculating the normalized root-mean-square ~~(NRMS) error~~ error (NRMSE) between a set of observed quantities (i.e. damage-equivalent loads from simulations) $\mathbf{y} = g(\mathbf{X}^{(i)}), i = 1 \ldots N$, and the PCE predictions, $\tilde{\mathbf{y}} = \tilde{S}(\mathbf{X}^{(i)}), i = 1 \ldots N$, over the same set of N sample points $\mathbf{X}^{(i)}$ ~~:~~ from the training sample defined in Section 2:

$$\varepsilon_{N,\mathrm{RMS}}\mathrm{NRMS} = \frac{1}{\mathrm{E}[\mathbf{y}]} \sqrt{\frac{\sum_{i=1}^{N}(\tilde{y}_i - y_i)^2}{N}} \tag{18}$$

where $\mathrm{E}[\mathbf{y}]$ is the expected value of the observed variable. Figure 4 shows the ~~NRMS error for a~~ NRMSE for a non-truncated PCE of order 6 and with 6 dimensions, as function of the number of samples used to train the PCE, and the hours of load simulations (i.e. number of seeds) used for each sample point. The ~~NRMS error~~ NRMSE shown on Figure 4 is calculated based on a set of 500 ~~pseudo-MC~~ quasi-MC points sampled from the joint pdf of reference site 0, and represents the difference in blade root flapwise DEL observed in each of the 500 points vs. the DEL predicted by a PC expansion trained on a selection of points from the high-fidelity database described in Section 2. Each of the quasi-MC samples is the mean from 48 turbulent 10-minute simulations. To mimic the seed-to-seed uncertainty, each of the PCE predictions is also evaluated as the mean of 48 normally distributed random realizations, with mean and standard deviation prescribed by the PCE model for mean and standard deviation of the blade flapwise DEL respectively.

Figure 4 illustrates a general tendency that using a few thousand training samples leads to convergence of the values of the PC coefficients, and the remaining uncertainty is due to seed-to-seed variations and due to the order of the PCE being lower than what is required for providing an exact solution at each sample point. Using longer simulations per sample point does not lead to further reduction in the statistical uncertainty due to seed-to-seed variations - with 4000 training samples, the ~~RMSE error~~ NRMSE for 1h simulation per sample is almost identical to the error with 8h simulation per sample. The explanation for this observation is that the seed-to-seed variation introduces an uncertainty not only between different simulations within the same sampling point, but also between different sampling points. This uncertainty materializes as an additional variance which is not explained by the smooth PCE surface. Further increase in the number of training points or simulation length will only reduce this statistical uncertainty, but will not contribute significantly to changes in the model predictions as the flexibility of the model is limited by the maximum polynomial order. Therefore, the ~~RMSE error~~ model performance achieved under these
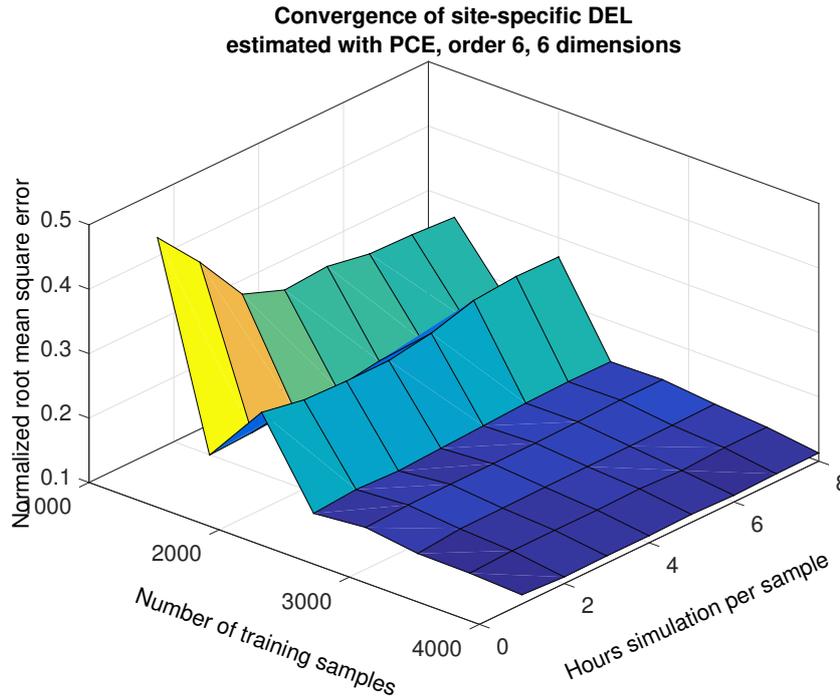
Figure 4. Convergence of a PCE of dimension 6 and order 6, as function of number of collocation points and hours of simulation per collocation point. The z-axis represents the ~~NRMS error~~ NRMSE obtained from the difference between 500 site-specific ~~pseudo-MC~~ quasi-MC samples of blade root flapwise DEL for reference site 0, and the corresponding predictions from PCE.

conditions can be considered near to the ~~smallest~~ best possible for the given PCE order and number of dimensions~~, and further increase in the number of training points or simulation length will not introduce noticeable improvement~~. However, it should be noted that the number of training points required for such convergence will differ according to the order and dimension of the PCE, and higher order and more dimensions will require more than the ~~∼ 3000~~ approximately 3000 points which seem sufficient for a PCE of order 6 and dimension 6, as shown on Figure 4.

5.1.1 ~~Sensitivity indices and model reduction~~

~~One useful corollary of the orthogonality in~~ The IS procedure has relatively slow convergence compared to e.g. a quasi-MC simulation. Figure 5 shows an example of the convergence of an IS integration for reference site 0, based on computing the target (site-specific) distribution weights for all 10^4 points in a reference high-fidelity database. The confidence intervals are obtained by bootstrapping.

5.2 One-to-one comparison and mean squared error
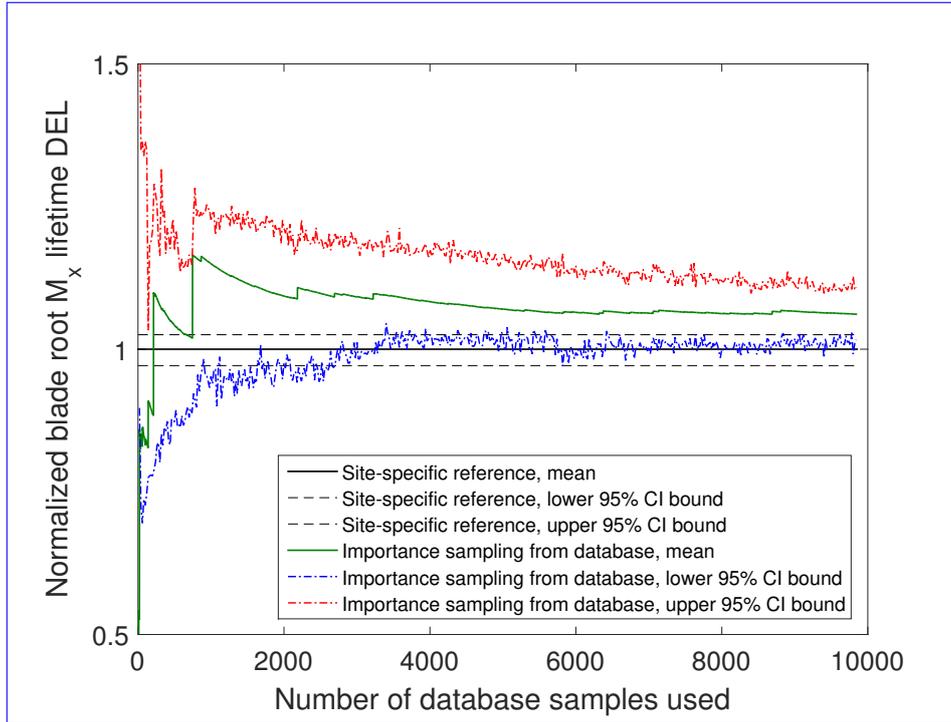
Figure 5. Convergence of an importance sampling (IS) calculation of the blade root moment from the hi-fi database, towards site-specific lifetime fatigue loads for reference site 0 (table 5).

Since the prediction of lifetime fatigue loads is the ~~PCE polynomial basis is that the total variance of the expansion can be expressed as the sum of the contributions from individual terms (Sudret, 2008):~~

$$\mathrm{Var}\left[\tilde{S}(\boldsymbol{\xi})\right] = \mathrm{Var}\left[\sum_{j=0}^{N_p-1} S_j \Psi_{\boldsymbol{\alpha},j}(\boldsymbol{\xi})\right] = \sum_{j=1}^{N_p-1} S_j^2 \mathrm{E}\left[\Psi_{\boldsymbol{\alpha},j}^2(\boldsymbol{\xi})\right]$$

~~Each of the terms in the sum~~ main purpose of the present study, the performance of the load prediction methods with respect to estimating the lifetime DEL is the main criterion for evaluation. However, the lifetime DEL as an integrated quantity will efficiently identify model bias but may not reveal the magnitude of some uncertainties which result in zero-mean error. As an additional means of comparison we calculate the normalized root-mean square error (NRMSE), defined in eq.~~represents the contribution of~~ (18) resulting from a point-by-point comparison of load predictions from a reduced-order model against actual reference values. The reference values are the results from the site-specific aeroelastic load simulations for reference site 0. At each sample point, the reference value y_i represents the mean DEL from all turbulence seeds simulated with these conditions. The values of the NRMSE for site 0 for Kriging, RS, and PCE-based load predictions are listed in Table 2. In addition, Figure 6 presents a one-to-one comparison where for a set of 200 sample points the load estimates from the site-specific MC simulations are compared against the corresponding surrogate model predictions in terms of $y-y$ plots.

Table 2. Normalised root mean square error characterizing the difference between aeroelastic simulations and reduced-order models. Load channel abbreviations are the following: TB: tower base; TT: tower top; MS: main shaft; BR: blade root. Loading directions consist of M_x: fore-aft (flapwise) bending, M_y: side-side (edgewise) bending, and M_z: torsion.

NORMALIZED RMS ERROR - SITE 0								
	Load channels							
Prediction model	TB M_x	TB M_y	TT M_x	TT M_y	TT M_z	MS M_z	BR M_x	BR M_y
Quadratic RS	0.0452	0.1404	0.1981	0.2612	0.0644	0.2280	0.1504	0.0098
PC expansion	0.0362	0.0955	0.1019	0.2089	0.0362	0.1530	0.0620	0.0084
Kriging	0.0334	0.0706	0.0837	0.1761	0.0368	0.1072	0.0519	0.0083
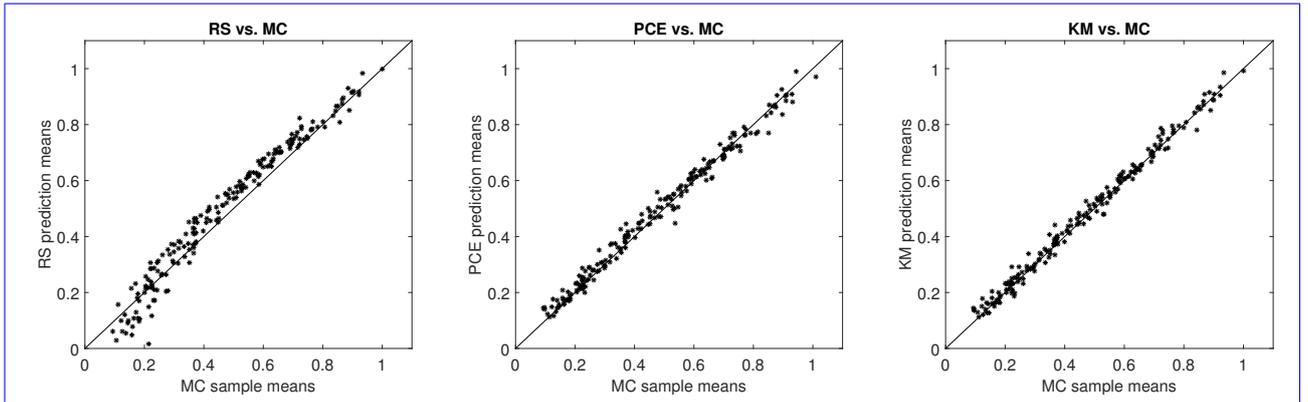
Figure 6. $y-y$ plots comparing the blade root flapwise 1Hz damage-equivalent load predictions for three load surrogate models - quadratic Response Surface, Polynomial Chaos expansion, and Kriging model, compared against site-specific Monte Carlo simulation. The x-axis represents the loads obtained using site-specific Monte Carlo simulations for reference site 0, and the y-axis represents the mean 1Hz-equivalent load estimated for the same sample points using a surrogate model. All values are normalized with the maximum equivalent load attained from the site-specific Monte Carlo simulation.

The RMS error analysis reveals a slightly different picture. In contrast to the lifetime DEL where the Kriging, PCE and RS methods showed very similar results, the ~~variables contained in the respective multivariate polynomials $\Psi_{\alpha,j}$ where $j = 0 \ldots N_p - 1$. This property can be used for eliminating polynomials which do not contribute significantly to the variance of the output, thus achieving a sparse, more computationally efficient reduced model. By combining the variance truncation and the LASSO regression technique in eq. , a model reduction of an order of magnitude or more can be achieved. For example, for a 9-dimensional PCE of order 6, using LASSO regularization parameter $\lambda = 1$ and retaining the polynomials which have a total variance contribution of 99.5%, resulted in a reduction of the number of polynomials from 5005 to about 200~~RMS error of the quadratic RS is for some channels about twice the RMS error of the other two approaches.

5

25

Denoting by $\mathscr{F}_{i_1,\ldots,i_s}$ the set of all polynomials dependent on a specific combination of input variables $(i_1,\ldots,i_s)$ (and only on them), the sum of variance contributions over $\mathscr{F}_{i_1,\ldots,i_s}$ normalized with the total variance represents the PCE-based Sobol' index with respect to variable set $\mathscr{F}_{i_1,\ldots,i_s}$ (Sudret, 2008):

$$SU_{i_1,\ldots,i_s} = \left(\sum_{\boldsymbol{\alpha} \in \mathscr{F}_{i_1,\ldots,i_s}} S_{\boldsymbol{\alpha}}^2 \mathrm{E}\left[\Psi_{\boldsymbol{\alpha}}^2(\boldsymbol{\xi}) \right] \right) \cdot \left(\mathrm{Var}\left[\tilde{S}(\boldsymbol{\xi}) \right] \right)^{-1}.$$

5 Based on eq. it is also straightforward to obtain the total Sobol indices for a given variable j by summing all $SU_{i_1,\ldots,i_s}$ where $j \in (i_1,\ldots,i_s)$. Note that since each variable appears in multiple cross-terms in the expansion, the contributions of some polynomial coefficients are included multiple times in the total Sobol' indices

5.3 Variable sensitivities

As described earlier in Section 4.1, we consider variable sensitivities (i.e. the influence of input variables on the variance of
10 the outcome) in terms of Sobol indices. By definition the Sobol indices will be dependent on the variance of input variables, meaning that for one and the same model, the Sobol indices will be varying under different (site-specific) input variable distributions. Taking this into account, we evaluate the Sobol indices for the two types of joint variable distributions used in this study - 1) a site-specific distribution, and 2) the joint distribution used to generate the database with high-fidelity load simulations. The total Sobol indices for the high-fidelity load database variable range are computed directly from the PCE fitted
15 to the sum of the total indices will typically exceed 1. database by evaluating the variance contributions from the expansion coefficients (see Appendix A) and are listed in Table 3. The indices for the site-specific distribution corresponding to reference site 0 are computed using the Monte-Carlo based method described in Section 4.1 as direct PCE indices are not available for this sample distribution. The resulting total Sobol indices for the 6 variables available at site 0 are listed in Table 4. The two tables show similar results - the mean wind speed and the turbulence are the most important factors affecting both fatigue and
20 extreme loads. Other two variables which show a smaller but still noticeable influence are the wind shear α, and the Mann model turbulence length scale L. The effect of wind shear is pronounced mainly for blade root loads which can be explained by the rotation of the blades which, if subjected to wind shear, will experience cyclic changes in wind velocity. The effect of Mann model Γ, veer, yaw, tilt, and air density within the chosen variable ranges seems to be minimal, especially for fatigue loads.

25 The Sobol indices estimated using the above procedure represent the relative contribution to the model variance from variables following the joint input distribution used to calibrate the PCE. In the present case, this distribution would span the uniform variable space of the

6 Site-specific calculations

6.1 Reference sites

Table 3. PCE-based Sobol sensitivity indices for the high-fidelity load database variable ranges.

FATIGUE LOAD SENSITIVITY INDICES									
	Variables								
Load channel	U	σ_u	α	L	Γ	$\Delta\varphi_h$	$\bar{\varrho}_h$	$\bar{\varrho}_v$	ρ
Tower base fore-aft moment M_x	0.42	0.65	0.01	0.03	0.02	0.01	0.00	0.00	0.01
Tower base side-side moment M_y	0.62	0.42	0.05	0.04	0.04	0.02	0.02	0.02	0.02
Blade root flapwise moment M_x	0.20	0.64	0.19	0.02	0.01	0.00	0.01	0.00	0.02
Blade root edgewise moment M_y	0.22	0.54	0.25	0.05	0.03	0.01	0.01	0.03	0.01
Tower top yaw moment M_z	0.14	0.85	0.00	0.02	0.01	0.00	0.00	0.00	0.01
Main shaft torsion M_z	0.48	0.53	0.01	0.06	0.01	0.01	0.01	0.01	0.01

Table 4. Site-specific Sobol sensitivity indices derived for Site 0 using Monte Carlo simulation from a PCE.

FATIGUE LOAD SENSITIVITY INDICES					
	Variables				
Load channel	U	σ_u	α	L	Γ
Tower base fore-aft moment M_x	0.08	1.32	0.07	0.18	0.09
Tower base side-side moment M_y	0.90	0.09	0.07	0.23	0.13
Blade root flapwise moment M_x	0.42	0.38	0.05	0.01	0.01
Blade root edgewise moment M_y	0.43	0.18	0.26	0.22	0.11
Tower top yaw moment M_z	0.23	0.53	0.01	0.08	0.01
Main shaft torsion M_z	0.47	0.36	0.06	0.03	0.07

The low-fidelity site-specific load calculation methods presented in this study are validated against a set of reference site-specific load calculations on a number of different virtual sites, based on real-world measurement data which cover most of the variable domain included within the high-fidelity ~~database defined in Section 2, and~~ database. In order to show a realistic example of situations where a site-specific load estimation is necessary, the ~~indices will correspond to~~ majority of the ~~load variation~~

5   ~~within the entire variable ranges as defined in Table 1. It may be more relevant to compute site-specific Sobol indices . This can be carried out efficiently by a Monte Carlo simulation on the PCE. For number of dimensions equal to M and for N (pseudo) Monte-Carlo samples the required experimental design represents an $N \times 2M$ matrix. This is divided into two $N \times M$ matrices, $\mathbf{A}$ and $\mathbf{B}$. Then, for each dimension i~~ virtual sites chosen are characterized with conditions that slightly exceed the standard conditions specified by a certain type-certification class. Exceptions are site 0, which has the most measured variables available and is therefore chosen as a primary reference site, ~~$i = 1 \ldots M$, a third matrix $\mathbf{AB}_i$ is created by taking the ith column~~

10   ~~of $\mathbf{AB}_i$ equal to the ith column from $\mathbf{B}$, and all other columns taken from $\mathbf{A}$. The load response function, i.e.~~ and the ~~PCE, is then evaluated for all three matrices, resulting in three model estimates: $f(\mathbf{A})$, $f(\mathbf{B})$, and $f(\mathbf{AB}_i)$. By repeating this for~~

27

$i = 1 \ldots M$, simulation-based Sobol' sensitivity indices can be estimated as

$$SU_i = \frac{1}{N}\sum_{j=1}^{N} f(\mathbf{B})_j \left(f(\mathbf{AB}_i)_j - f(\mathbf{B})_j \right)$$

where $j = 1 \ldots N$ is the row index in the design matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{AB}$, (Saltelli et al., 2008). virtual "sites" representing standard IEC class conditions. The IEC classes are included as test sites as they are described by only one independent variable (mean wind speed). They are useful test conditions as it may be challenging to correctly predict loads as function of only one variable using a model based on up to 9 random variables. The list of test sites is given in Table 5.

Table 5. Reference virtual sites used for validation of the site-specific load estimation methods.

Site No.	Location	Terrain	Specific condition	Variables included	MC sample size
0	Denmark	Flat agricultural	-	$U, \sigma_u, \alpha, L, \Gamma, \Delta\varphi$	492
1	Denmark	Flat agricultural	IIIC exceedance	U, σ_u, α	823
2	Denmark	Forested	IIIB exceedance	U, σ_u, α	884
3	Denmark	Forested	IA exceedance	U, σ_u, α	949
4	Denmark	Forested	IIA exceedance	U, σ_u, α	871
5	Colorado, USA	Mountain foothills	Low-wind	U, σ_u, α	657
6	Colorado, USA	Mountain foothills	Low-wind	U, σ_u, α	853
IEC IA, NTM	-	-	Standard reference class	U	22
IEC IIB, NTM	-	-	Standard reference class	U	22
IEC IIIC, NTM	-	-	Standard reference class	U	22
IEC IIB, ETM	-	-	Standard reference class	U	22

6.2 Universal Kriging with polynomial chaos basis functions

Kriging (Sacks et al., 1989; Santher et al., 2003) is a stochastic interpolation technique which assumes the interpolated variable follows a Gaussian process. A Kriging metamodel is described (Sacks et al., 1989) by

$$Y(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{f}(\mathbf{x}) + Z(\mathbf{x}),$$

where x represents the input variables, and $Y(\mathbf{x})$ is the output. The term $\boldsymbol{\beta}^T\mathbf{f}(\mathbf{x})$ is Site 0 (also also referred to herein as the reference site) is located at the Nørrekær Enge wind farm in Northern Denmark (Borraccino et al. , over flat, open agricultural terrain. Site 1 is a flat-terrain, near-coastal site at the the mean value of the Gaussian process (a.k.a. National Centre for Wind Turbines at Høvsøre, Denmark (Peña et al., 2016). Sites 2 to 4 are based on the wind conditions measured at the Østerild Wind Turbine Test Field which is located in a large forest plantation in North-western Denmark (Hansen et al., 2014). Due to the "trend") represented as a set of basis functions $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_P(\mathbf{x})]$ and

~~regression coefficients $\boldsymbol{\beta} = [\beta_1,\ldots,\beta_P]$; $Z(\mathbf{x},\mathbf{w})$ is a stationary, zero-mean Gaussian process. The probability distribution~~ of forested surroundings of the ~~Gaussian process is characterized by its covariance~~

$$V(\mathbf{w},\mathbf{x}) = \sigma^2 R(\mathbf{w},\mathbf{x},\boldsymbol{\theta})$$

~~where σ^2 is the process variance which is assumed to be constant,~~ site, the flow conditions are more complex than those in Nørrekær Enge and Høvsøre. By applying different filtering according to wind direction, three virtual site climates are generated and considered as sites 2–4.

Sites 5 and ~~$R(\mathbf{w},\mathbf{x},\boldsymbol{\theta})$ is the correlation between $Z(\mathbf{x})$ and $Z(\mathbf{w})$. The hyperparameters $\boldsymbol{\theta}$ define the correlation behavior, in terms of e.g. a correlation length. Given a set of points $\mathbf{x} = [x_1, x_2, \ldots x_N]$ where the true function values $\mathbf{y} = Y(\mathbf{x})$ are known, the aim is to obtain a model prediction at a new point, x'. Based on Gaussian theory, the known values $Y(\mathbf{x})$~~ 6 are located at NREL's National Wind Technology Center (NWTC), near the base of the Rocky Mountain foothills just south of Boulder, Colorado (Clifton et al., 2013). Similar to Østerild, directional filtering is applied to the NTWC data to split it into two virtual sites—accounting for the different conditions and wind climates from the two ranges of directions considered.

For each site, the joint distributions of all variables are defined in terms of conditional dependencies, and generating simulations of site-specific conditions is carried out using the Rosenblatt transformation, eq. (1). The conditional dependencies are described in terms of functional relationships between the governing variable and the distribution parameters of the dependent variable, e.g. the mean and standard deviation of the turbulence are modelled as linearly dependent on the wind speed as recommended by the IEC 61400-1 (ed. 3, 2005) standard. The wind shear exponent is dependent on the mean wind speed and on the turbulence, and the ~~Kriging predictor $\hat{Y}(x')$ will be jointly Gaussian distributed, and $\hat{Y}(x')$ will have the following mean and variance (Santher et al., 2003):~~

$$\mu_{\hat{Y}}(x') = \mathbf{f}(x')^T\boldsymbol{\beta} + r(x')^T\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\beta})$$
$$\sigma^2_{\hat{Y}}(x') = \sigma^2(1 - r(x')^T\mathbf{R}^{-1}r(x') + u(x')^T[\boldsymbol{\Psi}^T\mathbf{R}^{-1}\boldsymbol{\Psi})^{-1}u(x')].$$

distribution parameters of α are defined by the relationship (Kelly et al., 2014; Dimitrov et al., 2015)

$$\mu_{\alpha|I_c,u} = \alpha_{\text{ref}} + \frac{I_{\text{c,ref}} - I_c(U)}{I_c(U)\cdot c_\alpha} \tag{19}$$

$$\sigma_\alpha = 1/U$$

where μ_α and σ_α are the mean and standard deviation of α, respectively; $I_c(U) = (\sigma_u/U | F(\sigma_u) = Q)$ is a characteristic turbulence intensity based on a turbulence quantile Q, as a function of the wind speed U. Here ~~$\boldsymbol{\Psi}$ is the design matrix collecting the terms constituting the basis functions $\mathbf{f}(\mathbf{x})$,~~

$$\Psi_{ij} = f_j(x_i) \text{ for } i = 1\ldots N \text{ and } j = 1\ldots P;$$

~~$r(x')$~~$I_{\text{c,ref}} = I_c(U = 15\,\mathrm{m\,s^{-1}})$ is the ~~vector of cross-correlations between the prediction point x' and the known points $\mathbf{x}$; $\mathbf{R}$ is the correlation matrix of the known points,~~

$$R_{ij} = R(x_i, x_j, \theta) \text{ for } i,j = 1,\ldots,N; \text{ and}$$

$u(x') = \mathbf{\Psi}^T \mathbf{R}^{-1} r(x') - f(x')$. Using the predictor functions above requires determining the regression coefficients ($\boldsymbol{\beta}$), the field variance (σ^2), reference characteristic turbulence intensity at $U = 15\,\mathrm{m\,s^{-1}}$ and the correlation hyperparameters ($\boldsymbol{\theta}$). A suitable approach is to find the values of $\boldsymbol{\beta}$, σ^2 $\alpha_{\mathrm{ref}} = \alpha|(I_c = I_{c,\mathrm{ref}}, U = 15\,\mathrm{m\,s^{-1}})$ is a reference wind shear exponent, with α_c being an empirically determined constant. Since the turbulence quantities $I_c(u)$ and $I_{c,\mathrm{ref}}$ are defined by the conditional

5 relationship between wind speed and turbulence, the only site-specific parameters required for characterizing the wind shear are α_{ref} and $\boldsymbol{\theta}$ which maximize the likelihood of $\mathbf{y}$, i. e. minimize the model error in a least-squares sense (Lataniotis et al., 2015) ÷

$$\mathcal{L}(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = \frac{\det(\mathbf{R})^{-1/2}}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{\Psi}\boldsymbol{\beta})^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{\Psi}\boldsymbol{\beta})\right].$$

Here the hyperparameters, $\boldsymbol{\theta}$, appear within the correlation matrix $\mathbf{R}$. Having set up the design matrix $\mathbf{\Psi}$, c_α. For each of the

10 expansion coefficients $\boldsymbol{\beta}$ can be determined with the least-squares approach,

$$\boldsymbol{\beta} = \beta(\boldsymbol{\theta}) = (\mathbf{\Psi}^T \mathbf{R}^{-1} \mathbf{\Psi})^{-1} \mathbf{\Psi}^T \mathbf{R}^{-1} \mathbf{y}.$$

The field variance is obtained as

$$\sigma^2 = \sigma^2(\boldsymbol{\theta}) = \frac{1}{N}(\mathbf{y} - \mathbf{\Psi}\boldsymbol{\beta})^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{\Psi}\boldsymbol{\beta}).$$

From and it follows that $\boldsymbol{\beta}$ and σ^2 can be expressed as functions of $\boldsymbol{\theta}$. Therefore, calibrating the Kriging model amounts to

15 finding the values of $\boldsymbol{\theta}$ which maximize the likelihood. By combining eqns. A14–A16 this leads to the optimization problem

$$\boldsymbol{\theta} = \arg\min_{D_\theta}\left(\frac{1}{2}\log\left(\det(\mathbf{R})\right) + \frac{N}{2}\log\left(2\pi\sigma^2\right) + \frac{N}{2}\right).$$

For a problem with M dimensions, we assume that the correlation between sample points can be modelled using an anisotropic separable correlation function ((Sacks et al., 1989; Lataniotis et al., 2015), which assumes a different correlation parameter for each dimension. The total correlation is expressed as the product of the individual one-dimensional correlation

20 functions,

$$R(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}) = \prod_{i=1}^{M} R(x_i, x_i', \theta_i).$$

The one-dimensional correlation functions are assumed to follow an exponential relation to the distance $h = (x_i - x_i')$ between points,

$$R(h, \theta) = \exp\left(-\frac{|h|}{\theta}\right).$$

25 The functional form of the mean field $\mathbf{f}(\mathbf{x})^T \boldsymbol{\beta}$ is identical to the generalized PCE defined in eq. , meaning that the PCE is a possible candidate model for the mean in a Kriging interpolation. We adopt this approach and define the Kriging mean as a PCE with properties as described in section 4.3.

~~The main practical difference between regression- or expansion-type models such as regular PCE and the Kriging approach is in the way the training sample is used in the model: in~~ physical sites, the wind speed, turbulence and wind shear distribution parameters are defined based on anemometer measurements at the sites. The results are listed in Table 6. In addition, high-frequency 3D sonic measurements were available at site 0 for the ~~pure regression-based approaches the training sample is used to only calibrate the regression coefficients, while in Kriging as in other interpolation techniques the training sample is retained and used in every new model evaluation.As a result the Kriging model may have an advantage in accuracy since the model error tends to zero in the vicinity of the training points; however this comes at the expense of an increase in the computational demands for new model evaluations.The extra computational burden is mainly the time necessary to assemble $\mathbf{r}(\mathbf{x}')$, the matrix of cross-correlations between the prediction points and the training sample, and the time to multiply $\mathbf{r}(\mathbf{x}')$ with other equation terms. Thus, while for a PCE the model evaluation time $t(N)$ for a sample of size N would follow $t(N) = \mathcal{O}(N)$, for a Kriging model $t(N) = \mathcal{O}(N^2)$. For a Kriging model, a gain in accuracy over the model represented by the trend function will only materialize in problems where there is a noticeable correlation between the residual values at different training points. In a situation where the model error is independent from point to point (as e. g. in the case when the error is only due to seed-to-seed variations in turbulence) the inferred correlation length will tend to zero and the Kriging estimator will be represented by the trend function alone.~~ entire measurement period, which allowed for estimating Mann turbulence model parameters using the approach described in (Dimitrov et al., 2017).

6.2 ~~Quadratic response surface~~

~~A quadratic-polynomial response surface (RS) method based on Central Composite Design (CCD) is a reduced-order model which, among other applications, is also used for wind turbine load prediction (Toft et al., 2016). The procedure involves fitting a quadratic polynomial regression (a response surface) to a normalized space of i.i.d variables which are derived from the physical variables using an isoprobabilistic transformation as the Rosenblatt transformation given in eqs. and . The design points used for calibrating the response surface form a combination of a central point, axial points with unit shifts in one dimension at a time, and a 2^k factorial design where k equals the number of dimensions and there are two levels per variable dimension (Figure 3) . Due to the structured design grid required, it is not possible to use this approach with the sample points from the high-fidelity database described in section 2. Therefore, we implement the procedure using an additional set of simulations. Due to the low-order of the response surface, it is also not possible to fully characterize the turbine response as function of mean wind speed using a single response surface. Therefore, multiple response surfaces are calibrated for wind speeds from 4 to 25 m/s in~~

With this procedure, 1000 quasi-Monte Carlo samples of the environmental conditions at each site are generated from the respective joint distribution. All realizations where the wind speed is between the DTU 10MW wind turbine cut-in and cut-out wind speed are fed as input to load simulations. The actual number of load simulations for each site are given in Table 6. Similarly to the load database simulations, 8 simulations with 1h duration are carried out ~~for characterizing each design~~ at each site-specific MC sample point. The ~~polynomial coefficients of the response surface~~ resulting reference lifetime equivalent loads are then defined ~~using least-squares regression with the same closed-form solution defined by~~ by applying eq. ~~. For any~~

Table 6. Parameters defining the conditional distribution relationships used in computing joint distributions of the environmental conditions for the test sites/conditions.

	Site
	0
1 ~~m/s steps. However, due to the exponential increase of the number of design points with increa~~	
	2
	3
	4
	5
	6
	7 (IEC IA, NTM)
	8 (IEC IIB, NTM)
9 ~~variables considered. Instead, we choose to replace three of the variables (yaw, tilt, and air density) with their mean values. The result is a 6-dimensio~~	

~~sample point p in the central composite design, the corresponding row in the design matrix is defined as~~

$$\mathbf{\Psi}_p = \left[\{1\}, \quad \{U_1, \ldots U_n\}, \quad \{U_1^2, \ldots U_n^2\}, \quad \{U_i \cdot U_j, \, i=1 \ldots n, \, j=1 \ldots (i-1)\} \right]$$

~~where $\mathbf{U}$ are standard Normal variables derived from the physical variables $\mathbf{X}$ by an isoprobabilistic transformation. Example of a rotatable Central Composite Design (CCD) in a 2-dimensional standard Normal space $[u_1, u_2]$. The CCD consists of a~~
5 ~~central point, a 2^k factorial design with 2 levels and $k=2$ dimensions, and axial points at distance $u=\sqrt{2}$, meaning that all the outer points lie on a circle.~~ (6) on the Monte Carlo simulations using equal weights $p(\mathbf{X}_i) = 1/N$, while the uncertainty in the lifetime loads is estimated using bootstrapping on the entire MC sample.

7 ~~Results from site-specific calculations~~

6.1 Lifetime fatigue loads

10 The lifetime damage-equivalent (DEL) loads are computed for all reference sites in Table 5, using the ~~six methodologies~~ five load surrogate models defined above: 1) ~~full pseudo-Monte Carlo simulation; 2)~~ quadratic response surface; ~~3~~2) polynomial chaos expansion, ~~4~~3) importance sampling, ~~5~~4) nearest-neighbor interpolation; and ~~6~~5) Kriging with the mean defined by polynomial chaos basis functions. Methods ~~3–6~~ 2–5 are based on data from the high-fidelity load database described in Section 2. In addition to the surrogate model computations, a full MC reference simulation is carried out for each validation
15 site. The load predictions with the Monte Carlo approach are obtained by carrying out ~~HAWC2 aeroelastic simulations with the~~ Hawc2 aeroelastic simulations on the same DTU 10MW reference wind turbine ~~for each validation site~~ model used for training the load surrogate models. A total of ~~approx. 1000 pseudo-MC~~ $N_{\mathrm{MC}} = 1000$ quasi-MC samples are drawn from the joint distribution of environmental input variables characterizing each site, and 8h of aeroelastic simulations are carried out

for each ~~pseudo-MC sample~~ of the quasi-MC samples where the wind speed is between cut-in and cut-out. An exception are the IEC-based sites, where the standard IEC procedure is followed and loads are evaluated for 22 wind speeds from cut-in to cut-out in 1m/s steps. For each site, the full Monte Carlo simulation is then used as a reference to test the performance of the other five methods. The load predictions from the PCE, Kriging, the quadratic RS and the nearest-neighbor interpolation procedures all use a ~~pseudo-MC~~ quasi-MC simulation of the respective model with ~~inputs drawn from the site-specific distributions of environmental variables~~the same sample set of environmental inputs used in the reference MC simulations. The load predictions with Importance Sampling are based on the probability-weighted contributions from ~~all~~ the samples in a high-fidelity load database. For each site-specific distribution, the database samples are ordered according to their weights, and only a number of points, N_{IS}, with the highest weights are used in the estimation. For the sake of easier comparison between different methods, it is chosen that $N_{IS} = N_{MC}$. Based on computations from PCE with different number of dimensions and different maximum order, it was observed that expansions of order 4 or 5 may not be sufficiently accurate for some response channels. This is illustrated in Figure 7 where the prediction of main shaft torsion loads using order 4 and order 6 PC expansion are compared against other methods, and the order 4 calculation shows a significant bias. Therefore, the PC expansion used for reporting the results in this section is based on the same 6-dimensional variable input used with the quadratic response surface, and has a maximum order of 6. For evaluating confidence intervals from the reduced-order models (Kriging, PCE and quadratic response surface), two reduced-order models of each type are calibrated - one for the mean values, and the other for the standard deviations. This allows ~~computing confidence intervals directly by eq. , or~~ generating a number of realizations for each sampled combination of input variables, and subsequently computing confidence intervals by bootstrapping (eq. (7)). In ~~the present, we use the latter approach, because it allows bootstrapping simultaneously~~ this way, the bootstrapping is done simultaneously for a random sample of the input variables and the random seed-to-seed variations within each sample. As a result, the obtained confidence interval reflects the combination of seed-to-seed uncertainty and the uncertainty due to finite number of samples from the distribution of the input variables. This approach also allows consistency with the Importance Sampling and Nearest-Neighbor interpolation techniques, where ~~bootstrapping is also~~ the same bootstrapping approach is used. Since the lifetime fatigue load is in essence an integrated quantity subject to the law of large numbers, the uncertainty in computations based on a random sample of size N will be proportional to $1/\sqrt{N}$. Comparing uncertainties and confidence intervals as defined in Section 3.3 will therefore only be meaningful when approximately the same number of samples is used for all calculation methods. This approach is used for generating Figures 8 and 9, where the performance of all site-specific load estimation methods is compared for reference site 0, for 8 load channels in total, with number of samples as listed in Table 8. Figure 8 shows results for tower base and tower top fore-aft and side-side bending moments, and Figure 9 displays the tower top yaw moment, the main shaft torsion, and blade root flapwise and edgewise bending moments.

The results for Site 0 show that for all methods the prediction of blade root and tower top loads is more accurate than the prediction of tower base loads. Also, overall the predictions from the reduced models — the quadratic RS and the PCE, as well as from the Kriging model, are more robust than the IS and nearest-neighbor (NN) interpolation techniques. Similar performance is observed for most other validation sites. The ~~full~~ summarized site-specific results for all surrogate-based load estimation methods are shown in ~~Tables ??, ??, ??, ??, and ??, for the PCE, Kriging, quadratic RS, Importance Sampling, and~~
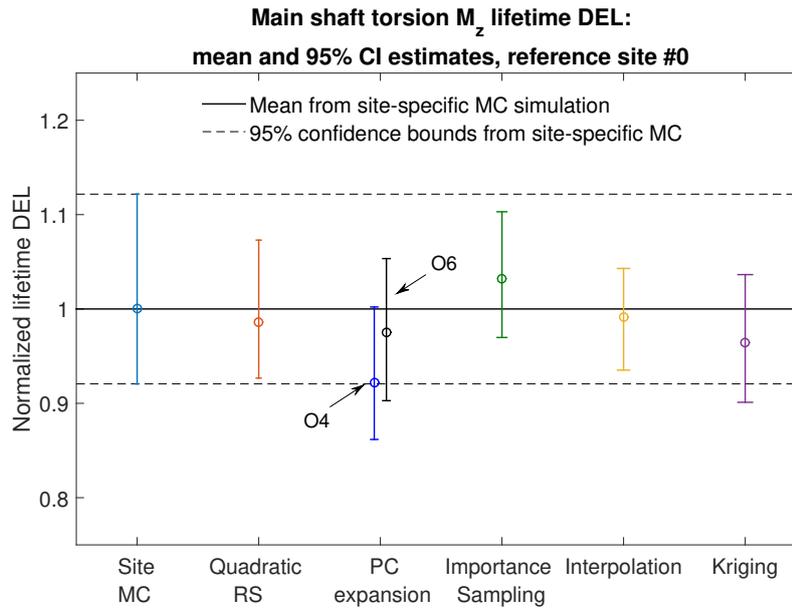
Figure 7. Comparison of predictions of the lifetime damage-equivalent loads for six different estimation approaches. All values are normalized with respect to the mean estimate from a site-specific Monte Carlo simulation, and the error bars represent the bounds of the 95% confidence intervals. Results from two PCEs are shown: the blue bar corresponds to the output of a 4th order PCE, while the black bar corresponds to a 6th order PCE.

~~NN-interpolation techniques respectively. In all tables, the values represent the mean estimates and~~ Table 7. In order to compute these values, the load estimates for each site and load channel are normalized to the results obtained with the direct site-specific Monte Carlo simulations. The values given in Table 7 are averaged over all reference sites. The results for individual sites and load channels are depicted as bar plots in Figures 10 and 11, for tower load and rotor load channels respectively. The largest

5 observed errors amount to $\approx 9\%$ with Kriging, $\approx 10\%$ for the PCE, $\approx 10\%$ for the quadratic RS, $\approx 24\%$ for IS, and ~ 15–17% for NN-interpolation. Noticeably, the low wind speed, high turbulence site 5 seems as the most difficult for prediction—for most load prediction methods this is the site where the largest error is found. All models except the Kriging also show relatively large errors for the IEC class-based sites. That can be attributed to significantly smaller number of samples used for the ~~ICE-based~~ IEC-based sites (22 samples where only the wind speed is varied in 1m/s steps from 4 to 25m/s). As mentioned above, the

10 statistical uncertainty in the estimation of the lifetime DEL will diminish with increasing number of samples. In addition to this effect, as discussed in section 4.1, the uncertainty of the IS model can increase when the site-specific distribution has fewer dimensions than the model because fewer points from the high-fidelity database will have high probabilities with respect to the site-specific distribution. It can be expected that this effect is strongest for the IEC class-based sites, as for them only a single variable — the wind speed — is considered stochastic.
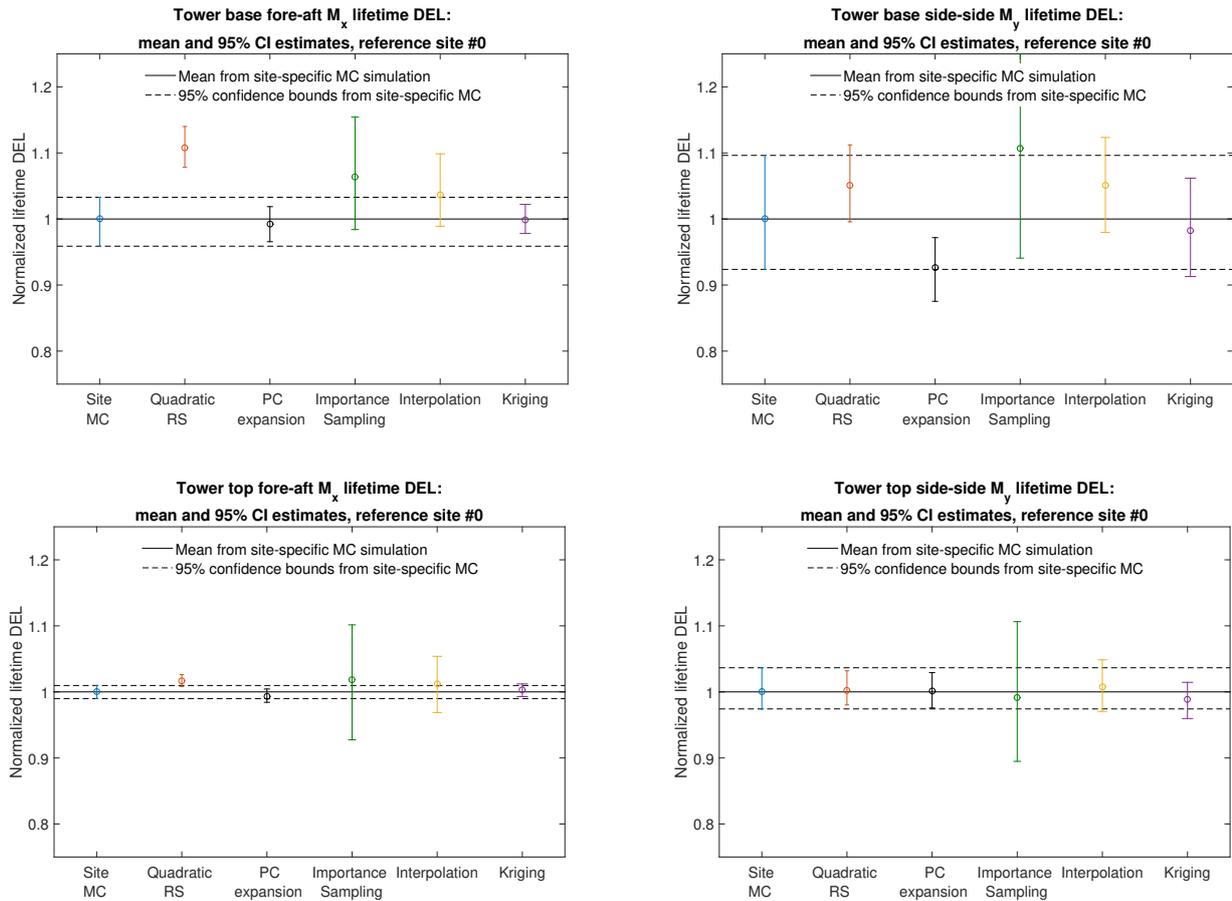
Figure 8. Comparison of predictions of the lifetime damage-equivalent loads for six different estimation approaches. All values are normalized with respect to the mean estimate from a site-specific Monte Carlo simulation.

6.2 ~~One-to-one comparison and mean squared error~~

~~Since the prediction of lifetime fatigue loads is the main purpose of the present study, the performance of the load prediction methods with respect to estimating the lifetime DEL is the main criterion for evaluation. However, the lifetime DEL as an integrated quantity will efficiently identify model bias but may not reveal the magnitude of some uncertainties which result in zero-mean error. As an additional means of comparison we calculate the normalized root-mean square (NRMS) error, defined in eq. resulting from a point-by-point comparison of load predictions from a reduced-order model against actual reference values. The reference values are the results from the site-specific aeroelastic load simulations for reference site 0. At each sample point, the reference value y_i represents the mean DEL from all turbulence seeds simulated with these conditions. The values of the NRMS error for site 0 for Kriging, RS, and PCE-based load predictions are listed in Table 2. Figure 6 shows a one-to-one comparison for a short sequence of sample points.~~

Yaw moment M$_z$ lifetime DEL:
mean and 95% CI estimates, reference site #0

Normalized lifetime DEL

—— Mean from site-specific MC simulation
--- 95% confidence bounds from site-specific MC

Site MC | Quadratic RS | PC expansion | Importance Sampling | Interpolation | Kriging

Main shaft torsion M$_z$ lifetime DEL:
mean and 95% CI estimates, reference site #0

Normalized lifetime DEL

—— Mean from site-specific MC simulation
--- 95% confidence bounds from site-specific MC

Site MC | Quadratic RS | PC expansion | Importance Sampling | Interpolation | Kriging

Blade root flapwise M$_x$ lifetime DEL:
mean and 95% CI estimates, reference site #0

Normalized lifetime DEL

—— Mean from site-specific MC simulation
--- 95% confidence bounds from site-specific MC

Site MC | Quadratic RS | PC expansion | Importance Sampling | Interpolation | Kriging

Blade root edgewise M$_y$ lifetime DEL:
mean and 95% CI estimates, reference site #0

Normalized lifetime DEL

—— Mean from site-specific MC simulation
--- 95% confidence bounds from site-specific MC

Site MC | Quadratic RS | PC expansion | Importance Sampling | Interpolation | Kriging

Figure 9. Comparison of predictions of the lifetime damage-equivalent loads for six different estimation approaches. All values are normalized with respect to the mean estimate from a site-specific Monte Carlo simulation.

Root mean square error characterizing the difference between aeroelastic simulations and reduced-order models. Load channel abbreviations are the following: TB: tower base; TT: tower top; MS: main shaft; BR: blade root. Loading directions consist of M_x: fore-aft (flapwise) bending, M_y: side-side (edgewise) bending, and M_z: torsion. Prediction model TB M_x TB M_y TT M_x TT M_y TT M_z MS M_z BR M_x BR M_y Quadratic RS 0.0452 0.1404 0.1981 0.2612 0.0644 0.2280 0.1504 0.0098 PC expansion 0.0362 0.0955 0.1019 0.2089 0.0362 0.1530 0.0620 0.0084 Kriging 0.0334 0.0706 0.0837 0.1761 0.0368 0.1072 0.0519 0.0083

The RMS error analysis reveals a slightly different picture. In contrast to the lifetime DEL where the Kriging, PCE and RS methods showed very similar results, the RMS error of the quadratic RS is for some channels about twice the RMS error of the other two approaches.

6.2 Variable sensitivities
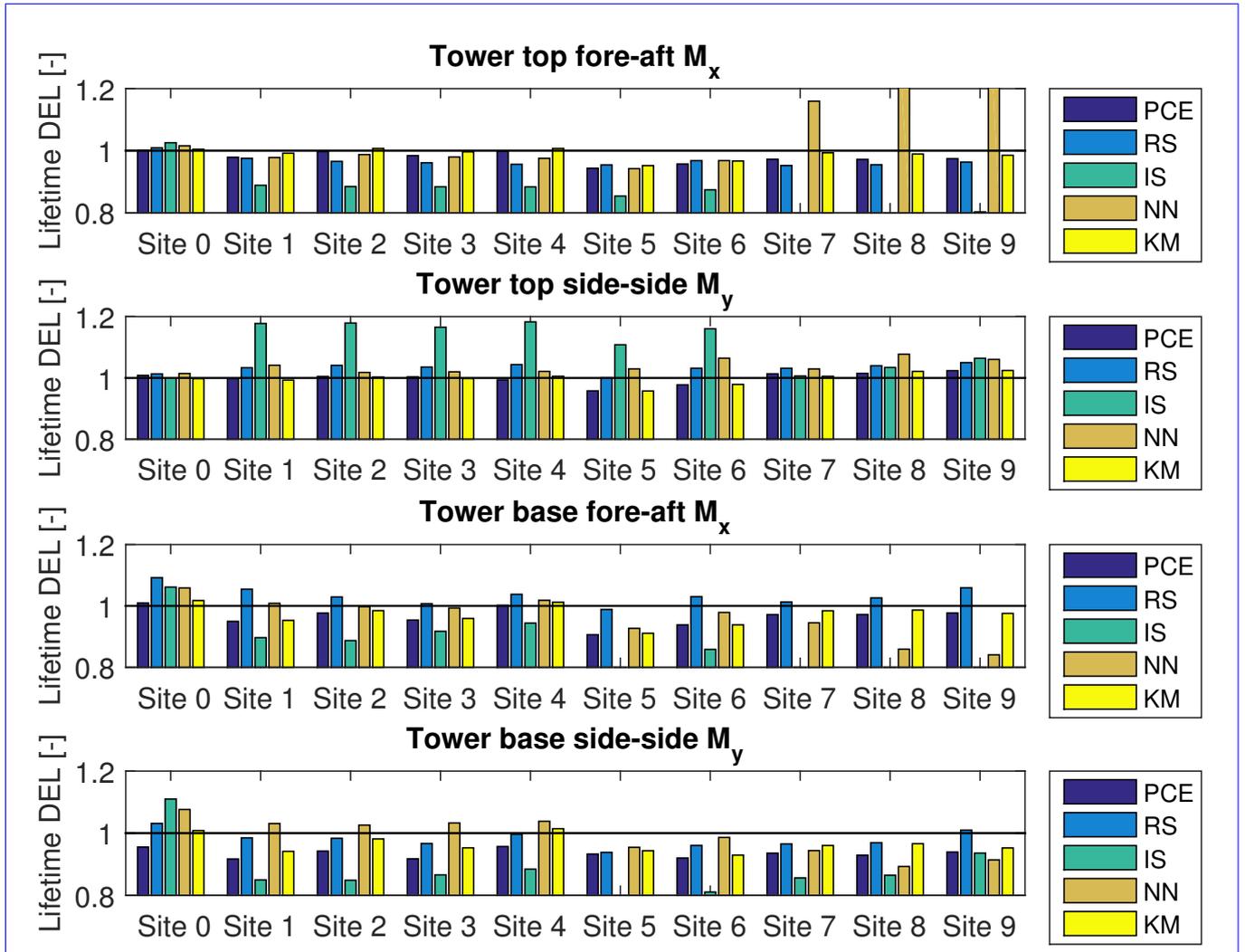
Figure 10. Predictions of lifetime damage-equivalent tower loads for five different estimation approaches and four load channels for the different sites (0–6) and IEC conditions (virtual sites 7–9). All values are normalized with respect to the mean estimate from a site-specific Monte Carlo simulation. The abbreviations refer to: PCE: Polynomial Chaos Expansion; RS: quadratic Response Surface; IS: Importance Sampling; NN: Nearest-neighbor Interpolation; KM: Universal Kriging Model.
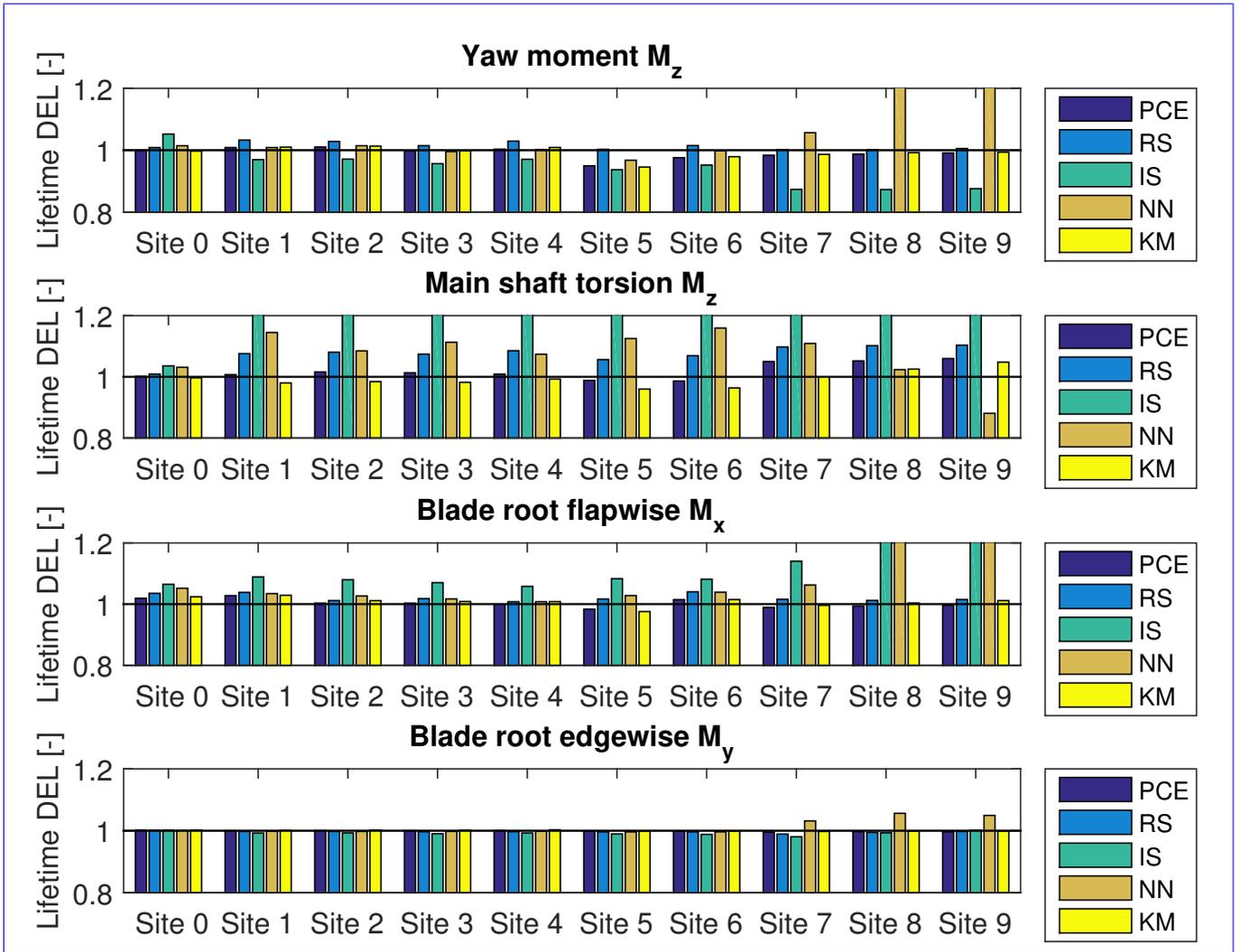
Figure 11. ~~Sample-by-sample comparison~~ Predictions of ~~the 1Hz~~ lifetime damage-equivalent ~~load predictionsloads (yaw,~~ ~~for two load prediction methods – quadratic Response Surfaceshaft~~ torsion, blade-root) for five different estimation approaches and ~~Polynomial Chaos expansion, compared against~~ four load channels. All values are normalized with respect to the mean estimate from a site-specific Monte Carlo simulation. ~~Each point on the x-axis represents a sample point from the site-specific distribution of input variables for reference site 0, and the y-axis represents the estimated mean 1Hz-equivalent load~~ The abbreviations refer to: PCE: Polynomial Chaos Expansion; RS: quadratic Response Surface; IS: Importance Sampling; NN: Nearest-neighbor Interpolation; KM: Universal Kriging Model.

Table 7. ~~Lifetime damage-equivalent fatigue~~ Lifetime-equivalent load predictions ~~obtained using 6-dimensional Polynomial Chaos Expansion of order 6, relative~~ normalized with respect to ~~predictions from site-specific~~ Monte Carlo simulations and averaged over 10 reference sites. Load channel abbreviations are the following: TB: tower base; TT: tower top; MS: main shaft; BR: blade root. Loading directions consist of M_x: fore-aft (flapwise) bending~~,~~; M_y: side-side (edgewise) bending~~,~~; and M_z: torsion.

~~Site 0~~

~~Site 1 0.950 0.917 0.979 0.998 1.009 1.007 1.028 1.000 Site 2 0.977 0.942 0.997 1.005~~

~~Lifetime damage-equivalent fatigue load predictions obtained using the Universal Kriging approach with a 6-di~~

~~Site number TB M_x TB M_y TT M_x TT M_y TT M_z MS M_z BR M_x BR M_y Site 0 1.017 1.008 1.004 0.998 0.997 0.997 1.024 1.002 Site 1 0.953 0.94~~

~~Lifetime damage-equivalent fatigue load predictions obtained using 6-dimensional Quadratic R~~

~~Site number TB M_x TB M_y TT M_x TT M_y TT M_z MS M_z BR M_x BR M_y Site 0 1.092 1.031 1.009 1.013 1.009 1.009 1.035 1.001 Site 1 1.055 0.98~~

~~Lifetime damage-equivalent fatigue load predictions obtained using Importance Sampling of points fr~~

~~Site number TB M_x TB M_y TT M_x TT M_y TT M_z MS M_z BR M_x BR M_y Site 0 1.059 1.098 1.015 1.004 1.039 1.042 1.050 1.002 Site 1 0.893 0.84~~

~~Lifetime damage-equivalent fatigue load predictions obtained using nearest-neighbor interpolation in a 6-~~

~~Site number TB M_x TB M_y TT M_x TT M_y TT M_z MS M_z BR M_x BR M_y Site 0 1.077 1.093 1.018 1.029 1.036 1.052 1.056 1.000 Site 1 1.002 1.0~~

~~As described earlier in Section 4.1, we consider variable sensitivities (i. e. the influence of input variables on the variance of the outcome) in terms of Sobol indices. By definition the Sobol indices will be dependent on the variance of input variables, meaning that for one and the same model , the Sobol indices will be varying under different (site-specific) input variable distributions. Taking this into account, we evaluate the Sobol indices for the two types of joint variable distributions used in~~ 5 ~~this study - 1) a site-specific distribution, and 2) the uniform, bounded joint distribution used to generate the database with high-fidelity load simulations. The Sobol indices for~~

Another important aspect to consider when comparing the performance of the surrogate models is the ~~high-fidelity load database variable range are listed in Table 3, while the indices for the~~ model execution speed, and whether there is a tradeoff between speed and accuracy. A comparison of the model evaluation times for the site-specific ~~distribution corresponding to~~ 10 ~~reference site~~ lifetime load computation for site 0 ~~are listed in Table 4. The two tables show similar results - the mean wind speed and the turbulence are the most important factors affecting both fatigue and extreme loads. Other two variables which show a smaller but still noticeable influence are the wind shear α, and the Mann model turbulence length scale L. The effect of wind shear is pronounced mainly for blade root loads which can be explained by the rotation of the blades which, if subjected~~

to wind shear, will experience cyclic changes in wind velocity. The effect of Mann model Γ, veer, yaw, tilt, and air density within the chosen variable ranges seems to be minimal, especially for fatigue loads. is given in Table 8. Noticeably the Kriging model requires significantly longer execution time than other approaches, which is mainly due to the requirement of populating a cross-correlation matrix.

Table 8. PCE-based Sobol sensitivity indices Model execution times for the high-fidelity lifetime damage-equivalent fatigue load database variable ranges.computations for site 0

Load channel U σ_u α L Γ $\Delta\varphi_h$ $\bar{\varphi}_h$ $\bar{\varphi}_v$ ρ Tower base fore-aft moment M_x 0.52 0.51 0.04 0.04 0.04 0.03 0.03 0.03 0.05 Tower base side-side moment M_y

Site-specific Sobol sensitivity indices derived for Site 0 using Monte Carlo simulation from a PCE. Load channel U σ_u α L Γ Tower base fore-aft moment M_x 0.08 1.32 0.07 0.18 0.09 Tower base side-side moment M_y 0.90 0.09 0.07 0.23 0.13 Blade root flapwise moment M_x 0.42 0.38 0.05 0.01 0.01 Blade root edgewise moment M_y 0.43 0.18 0.26 0.22 0.11 Tower top yaw moment M_z 0.23 0.53 0.01 0.08 0.01 Main shaft torsion M_z 0.47 0.36 0.06 0.03 0.07 Load channel U σ_u α L Γ Tower base fore-aft moment M_x 0.68 0.44 0.03 0.02 0.01 Tower base side-side moment M_y 0.52 0.40 0.07 0.22 0.17 Blade root flapwise moment M_x 0.76 0.21 0.02 0.02 0.01 Blade root edgewise moment M_y 0.62 0.05 0.10 0.05 0.05 Tower top yaw moment M_z 0.15 0.45 0.08 0.08 0.02 Main shaft torsion M_z 0.81 0.25 0.04 0.04 0.00

6.2 Mean extreme loads

Fatigue loads are the main quantity of interest in the present study, however the methods discussed can also be applied to other types of statistics as long as the load simulations carry the necessary information and cover the input variable space. For example, the conditions specified in the extreme-turbulence load case scenario DLC1.3 in IEC61400-1 class C and B are within the envelope of the high-fidelity database. DLC1.3 requires taking the mean of the extremes from all seeds simulated at a specific condition - meaning that the requirements for the reduced-order models are very similar to the case with fatigue loads. As a demonstration for the feasibility of these calculations, Figure **??** shows prediction of the mean of extremes for the DLC1.3 load case for IEC class IB, carried out by calibrating a PCE and a quadratic RS according to the same procedures as already defined for the fatigue DEL. Comparison of predictions of the mean of extremes obtained for the extreme turbulence load case DLC1.3 for IEC turbulence class B conditions. MC simulation (blue lines with crosses) denotes the reference values obtained with full-fidelity aeroelastic simulations.

7 Discussion and conclusions

7.1 Discussion

The previous sections of this paper described ~~the~~ a procedure for estimating site-specific lifetime damage-equivalent loads, using several simplified model techniques ~~which were applied for ten sites~~applied to ten different sites and conditions. Based

5 on the site-specific lifetime DEL comparisons, ~~three models~~ for quick site-specific load estimation the three models based on machine learning showed to be ~~viable (robust and sufficiently accurate) choices for quick site-specific load estimation: the PC~~ most viable (sufficiently accurate over the majority of the sampling space): polynomial chaos expansion, Kriging, and the quadratic ~~RS~~response surface. When estimating lifetime DEL, these methods showed approximately equal levels of uncertainty. However, in the one-to-one comparisons the quadratic RS model showed larger error, especially for sample points

10 corresponding to more extreme combinations of environmental conditions. This is due to the lower order and the relatively small number of calibration points of the quadratic RS, which means that the model accuracy decreases in the sampling space away from the calibration points, especially if there is any extrapolation. This inaccuracy is reflected in the ~~RMS error~~ NRMSE-error from one-to-one comparisons, but is less obvious in the lifetime fatigue load computations which average out errors with zero mean. The universal Kriging model demonstrated the ~~overall lowest uncertainty~~smallest overall uncertainty,

15 both in sample-to-sample comparisons and in lifetime DEL computations. This is to be expected, since the Kriging employs ~~an already~~ a well-performing model (the PCE) ~~,~~ and combines it with an interpolation scheme ~~which~~ that subsequently reduces the uncertainty even further. However, in most cases the observed improvement over a pure PCE is not significant. This indicates that the sources of the remaining uncertainty are outside the models—e.g. the seed-to-seed turbulence variations: the models being calibrated with turbulence realizations different from the ones used to compute the reference site-specific loads. As a

20 result the trend function ~~has the main~~ (the β term in Eq. 13) is the primary contribution to the Kriging estimator, and the influence of the Gaussian-field interpolation is minimal. A drawback of the Kriging model with respect to the other techniques is the larger computational demands due to the need of computing correlation matrices and the use of the training sample for each new evaluation.

For all site-specific load assessment methods discussed, the estimations are trustworthy only within the bounds of the variable

25 space used for model ~~calibration - extrapolation~~ calibration—extrapolation is either not possible, or may lead to unpredictable results. It is therefore important to ensure that the site-specific distributions used for load assessment are not outside the bounds of validity of the load estimation model.

The variable bounds presented in this paper are based on a certain degree of consideration of atmospheric physics employed in the relationships between wind speed, turbulence, wind shear, wind veer and turbulence length scale. The primary scope is to

30 encompass the ranges of conditions relevant for fatigue load analysis, and the currently suggested variable bounds include all normal-turbulence (NTM) classes. However, ~~in some situations~~ for some other calculations it may be more practical to choose other bound definitions~~-~~: e.g.~~for consistency with the IEC 61400-1 definitions of the~~ for the extreme turbulence models ~~,~~ ~~where~~ prescribed by the IEC 61400-1, the currently suggested bounds do not include ETM class A.

For the more advanced methods like PCE and Kriging, there is a practical limitation of the number of training points to be used ~~on~~ in a single-computer setup. For a PCE the practical limit is mainly subject to memory availability when assembling and inverting the information matrix, and for a PCE of order 6 and with 9 dimensions this limit is on the order of $1\text{–}2\cdot10^4$ points on a typical desktop computer (as of 2018). For Kriging, the computing time also plays a role: although a similar number of training points could be stored in memory as for the PCE, the computational time is much longer, and the practical limit of training points for most applications is less than for the PCE. However, as it was shown in Sections 4.3 and 6, a training sample of 10^4 points or even half of that should be sufficient for most applications in site-specific load prediction.

Considering the overall merits of the load prediction methods analyzed, the PCE provided an accurate and robust performance. The Kriging approach showed slightly better accuracy but at the expense of increased computational demands. Taking this together with the other useful properties of the PCE~~, e.g. the orthogonality allowing for creating sparse models or doing~~, such as orthogonality facilitating creation of sparse models through variance-based sensitivity analysis, we consider the PCE as the most useful method overall.

In addition to the ~~surrogate models~~ load-mapping approaches presented in this paper, Artificial Neural Networks (ANNs) are interesting alternative candidates. ANN (see Goodfellow et al., 2016) are machine learning models ~~which are gaining large~~ that have gained popularity due to their flexibility and history of successful ~~applications in many~~ application to many different problems. It is very likely that a ~~deep~~ sufficiently large Neural Network model can provide similar output quality and performance ~~than~~ as the methods described in the present study. This is therefore a matter worth of future research. However, the PCE-based models may sometimes have a practical advantage over ANNs, due to the "white-box" ~~features such known variance contributions and~~ features—such as being able to track separate contributions to variance (and uncertainty), as well as the possibility of obtaining analytical derivatives, which is important for applications to optimization problems.

The results from the site validations showed that for the majority of sites and load channels, the simplified load assessment techniques can predict the site-specific lifetime fatigue loads to within ~~$\approx 5\%$~~ about 5% accuracy. However, it should be noted that this accuracy is relative to full-fidelity load simulations, and not necessarily to the actual site conditions, where additional uncertainties (e.g. uncertainty in the site conditions or the turbine operating strategy) can lead to even larger errors. The procedures demonstrated in this study are thus very suitable for carrying out quick site feasibility assessments~~, which can~~; the latter can help to decide in a timely fashion ~~aid the decision on~~ whether to discard a given site as unfeasible, or ~~e.g.~~ to make additional high-fidelity computations or more measurements of site conditions. The same procedure, but with additional variables (e.g. 3 variables for wake-induced effects as in (Galinos et al., 2016)) may also be useful as objective function or constraint in farm optimization problems.

7.2 ~~Conclusions~~ Summary and conclusions

In the present work we defined and demonstrated a procedure for quick assessment of site-specific lifetime fatigue loads using load surrogate models calibrated by means of a database with high-fidelity load simulations. The performance of polynomial chaos expansion, quadratic response surface, universal Kriging, importance sampling, and nearest-neighbor interpolation in predicting site-specific lifetime fatigue loads was assessed by training the surrogate models on a database with aeroelastic

load simulations of the DTU 10MW reference wind turbine. Practical bounds of variation were defined for nine environmental variables and their effect on the lifetime fatigue loads was studied. The study led to the following main conclusions:

- The variable sensitivity analysis showed that mean wind speed and turbulence (standard deviation of wind speed fluctuations) are the factors having the highest influence on fatigue ~~and extreme~~ loads. The wind shear and the Mann turbulence length scale ~~and the wind shear exponent~~ were also found to have an ~~influence, and for the wind shear the effect is~~ appreciable influence, with the effect of wind shear being more pronounced for rotating components such as blades. Within the studied ranges of variation, the Mann turbulence parameter Γ, wind veer, yaw angle, tilt angle, and air density, were found to have small or negligible effect on the loads.

- The best performing models had errors of less than 5% for most sites and load channels, which is in the same order of magnitude as the variations due to realization-to-realization uncertainty.

- A universal Kriging model ~~with a~~ employing polynomial chaos expansion ~~used~~ as a trend function achieved the most accurate predictions, but also required the longest computing times.

- A polynomial chaos expansion with Legendre basis polynomials was concluded to be the approach with best overall performance.

- The procedures demonstrated in this study are well suited for carrying out quick site feasibility assessments ~~with a particular~~ conditional on a specific wind turbine model.

References

Bak, C., Zahle, F., Bitsche, R., Kim, T., Yde, A., Henriksen, L. C., Natarajan, A., and Hansen, M.: Description of the DTU 10MW reference wind turbine, Tech. Rep. I-0092, Technical University of Denmark, Department of Wind Energy, 2013.

Berg, J., Natarajan, A., Mann, J., and Patton, E.: Gaussian vs non-Gaussian turbulence: impact on wind turbine loads, Wind Energy, 19, 1975–1989, 2016.

Borraccino, A., Schlipf, D., Haizmann, F., and Wagner, R.: Wind field reconstruction from nacelle-mounted lidar short-range measurements, Wind Energy Science, 2, 269–283, https://doi.org/10.5194/wes-2-269-2017, https://www.wind-energ-sci.net/2/269/2017/, 2017.

Caflisch, R. E.: Monte Carlo and Quasi-Monte Carlo methods, Acta Numerica, pp. 1–49, 1998.

Choe, Y., Byon, E., and Chen, N.: Importance Sampling for Reliability Evaluation With Stochastic Simulation Models, Technometrics, 57, 351–361, https://doi.org/10.1080/00401706.2014.1001523, 2015.

Clifton, A., Schreck, S., Scott, G., Kelley, N., and Lundquist, J.: Turbine Inflow Characterization at the National Wind Technology Center, J. of Solar Energy Engineering, 135, 031 017, https://doi.org/doi:10.1115/1.4024068, 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.: Effect of Normal and Extreme turbulence spectral parameters on wind turbine loads, Renewable Energy, 101, 1180–1193, 2017.

Ditlevsen, O. and Madsen, H. O.: Structural Reliability Methods, Wiley, 1996.

Efron, B.: Bootstrap methods: another look at the jackknife, Annals of Statistics, 7, 1–26, 1979.

Galinos, C., Dimitrov, N., Larsen, T. J., Natarajan, A., and Hansen, K. S.: Mapping Wind Farm Loads and Power Production - A Case Study on Horns Rev 1, Journal of Physics: Conference Series, 753, 032 010, 2016.

Ganesh, N. and Gupta, S.: Estimating the Rain-Flow Fatigue Damage in Wind Turbine Blades Using Polynomial Chaos, in: Proceedings of the International Symposium on Engineering under Uncertainty: Safety Assessment and Management (ISEUSAM - 2012), https://doi.org/10.1007/978-81-322-0757-3, 2013.

Ghanem, R. G. and Spanos, P. D.: Stochastic finite elements - a spectral approach, Springer, Berlin, 1991.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, http://www.deeplearningbook.org, 2016.

Graf, P., Dykes, K., Damiani, R., Jonkman, J., and Veers, P.: Adaptive stratified importance sampling: hybridization of extrapolation and importance sampling Monte Carlo methods for estimation of wind turbine extreme loads, Wind Energy Science, 3, 475–487, https://doi.org/10.5194/wes-3-475-2018, 2018.

Graf, P. A., Stewart, G., Lackner, M., Dykes, K., and Veers, P.: High-throughput computation and the applicability of Monte Carlo integration in fatigue load estimation of floating offshore wind turbines, Wind Energy, 19, 861–872, https://doi.org/10.1002/we.1870, http://onlinelibrary.wiley.com/doi/10.1002/we.1608/fullhttp://doi.wiley.com/10.1002/we.1870, 2016.

Häfele, J., Hübler, C., Gebhardt, C. G., and Rolfes, R.: A comprehensive fatigue load set reduction study for offshore wind turbines with jacket substructures, Renewable Energy, 118, 99–112, https://doi.org/10.1016/j.renene.2017.10.097, 2018.

Hansen, B. O., Courtney, M., and Mortensen, N. G.: Wind Resource Assessment – Østerild National Test Centre for Large Wind Turbines, Tech. Rep. E-0052, Technical University of Denmark, Department of Wind Energy, 2014.

Hansen, M. H. and Henriksen, L. C.: Basic DTU Wind Energy Controller, Tech. Rep. E-0028, Technical University of Denmark, Department of Wind Energy, 2013.

Hübler, C., Gebhardt, C. G., and Rolfes, R.: Hierarchical four-step global sensitivity analysis of offshore wind turbines based on aeroelastic time domain simulations, Renewable Energy, 111, 878–891, https://doi.org/10.1016/j.renene.2017.05.013, http://dx.doi.org/10.1016/j.renene.2017.05.013, 2017.

IEC: International Standard IEC61400-1: Wind Turbines - Part 1: Design Guidelines, 2005.

5   Kashef, T. and Winterstein, S. R.: Relating turbulence to wind turbine blade loads: parametric study with multiple regression analysis, Solar Energy Engineering, 121, 156–161, 1999.

Kelly, M.: From standard wind measurements to spectral characterization: turbulence length scale and distribution, Wind Energy Science, 3, 533–543, https://doi.org/10.5194/wes-3-533-2018, https://www.wind-energ-sci.net/3/533/2018/, 2018.

Kelly, M. and van der Laan, P.: Veer and shear relations: beyond Ekman, towards statistical characterization, Quarterly Journal of the Royal
10   Meteorological Society, p. in preparation, 2018.

Kelly, M., Larsen, G., Natarajan, A., and Dimitrov, N.: Probabilistic meteorological characterization for turbine loads, Journal of Physics: Conference Series (online), 524, 012 076, 2014.

Larsen, T. J. and Hansen, A. M.: How to HAWC2, the user's manual, Tech. Rep. R-1597, Technical University of Denmark, Department of Wind Energy, 2012.

15   Lataniotis, C., Marelli, S., and Sudret, B.: UQLab user manual – Kriging (Gaussian process modelling), Tech. rep., Chair of Risk, Safety & Uncertainty Quantification, ETH Zurich, report UQLab-V0.9-105, 2015.

Liu, P. L. and Der Kiureghian, A.: Multivariate distribution models with prescribed marginals and covariances, Probabilistic Engineering Mechanics, 1, 105–112, 1986.

Mann, J.: The spatial structure of neutral atmospheric surface-layer turbulence, Journal of Fluid Mechanics, 273, 141–168, 1994.

20   Mann, J.: The spectral velocity tensor in moderately complex terrain, J. Wind Eng., 88, 153–169, 2000.

Manuel, L., Veers, P. S., and Winterstein, S. R.: Parametric models for estimating wind turbine fatigue loads for design, Solar Energy Engineering, 123, 346–355, 2001.

Mckay, M. D., Beckman, R. J., and Conover, W. J.: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code, Technometrics, 42, 55–61, 2000.

25   Morokoff, W. J. and Caflisch, R. E.: Quasi-Monte Carlo Integration, Computational Physics, 122, 218–230, 1995.

Mouzakis, F., Morfiadakis, E., and Dellaportas, P.: Fatigue loading parameter identification of a wind turbine operating in complex terrain, Wind Engineering and Industrial Aerodynamics, 82, 69–88, 1999.

Müller, K. and Cheng, P. W.: Application of a Monte Carlo procedure for probabilistic fatigue design of floating offshore wind turbines, Wind Energy Science, 3, 149–162, https://doi.org/10.5194/wes-3-149-2018, 2018.

30   Müller, K., Dazer, M., and Cheng, P. W.: Damage Assessment of Floating Offshore Wind Turbines Using Response Surface Modeling, Energy Procedia, 137, 119–133, https://doi.org/10.1016/j.egypro.2017.10.339, https://doi.org/10.1016/j.egypro.2017.10.339, 2017.

Murcia, J. P., Réthoré, P.-E., Dimitrov, N., Natarajan, A., Sørensen, J. D., Graf, P., and Kim, T.: Uncertainty propagation through an aeroelastic wind turbine model using polynomial surrogates, Renewable Energy, 119, 910–922, https://doi.org/https://doi.org/10.1016/j.renene.2017.07.070, 2018.

35   Peña, A., Floors, R., Sathe, A., Gryning, S.-E., Wagner, R., Courtney, M. S., Larsén, X. G., Hahmann, A. N., and Hasager, C. B.: Ten Years of Boundary-Layer and Wind-Power Meteorology at Høvsøre, Denmark, Boundary-Layer Meteorology, 158, 1–26, https://doi.org/10.1007/s10546-015-0079-8, https://doi.org/10.1007/s10546-015-0079-8, 2016.

Rosenblatt, M.: Remarks on a multivariate transformation, Annals of Mathematical Statistics, 23, 470–472, 1952.

Rychlik, I.: A New Definition of the Rainflow Cycle Counting Method, International Journal of Fatigue, 9, 119–121, 1987.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P.: Design and analysis of computer experiments, Statistical Science, 4, 409–423, 1989.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S.: Global Sensitivity Analysis: The Primer, Wiley, 2008.

5 Santher, T. J., Williams, B. J., and Notz, W. I.: The design and analysis of computer experiments, Springer-Verlag, 2003.

Sobol, I.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Mathematics and Computers in Simulation, 55, 271–280, 2001.

Stewart, G. M.: Design Load Analysis of Two Floating Offshore Wind Turbine Concepts, Ph.D. thesis, University of Massachusetts - Amherst, http://scholarworks.umass.edu/dissertations{_}2, 2014.

10 Sudret, B.: Global sensitivity analysis using polynomial chaos expansions, Reliability Engineering and Systems Safety, 93, 964–979, 2008.

Teixeira, R., O'Connor, A., Nogal, M., Krishnan, N., and Nichols, J.: Analysis of the design of experiments of offshore wind turbine fatigue reliability design with Kriging surfaces, Procedia Structural Integrity, 5, 951–958, https://doi.org/10.1016/j.prostr.2017.07.132, http://linkinghub.elsevier.com/retrieve/pii/S2452321617302445, 2017.

Tibshirani, R.: Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B, 58, 267–288, 1996.

15 Toft, H. S., Svenningsen, L., Moser, W., Sørensen, J. D., and Thørgensen, M. L.: Assessment of Wind Turbine Structural Integrity using Response Surface Methodology, Engineering Structures, 106, 471–483, 2016.

Xiu, D. and Karniadakis, G. E.: The Wiener-Askey polynomial chaos for stochastic differential equations, Journal of Scientific Computing, 191, 4927–4948, 2002.

Yang, H., Zhu, Y., Lu, Q., and Zhang, J.: Dynamic reliability based design optimization of the tripod sub-structure of offshore wind turbines,

20 Renewable Energy, 78, 16–25, https://doi.org/10.1016/j.renene.2014.12.061, 2015.

Zwick, D. and Muskulus, M.: Simplified fatigue load assessment in offshore wind turbine structural analysis, Wind Energy, 19, 265–278, https://doi.org/10.1002/we.1831, http://onlinelibrary.wiley.com/doi/10.1002/we.1608/fullhttp://doi.wiley.com/10.1002/we.1831, 2016.

Appendix A: Reduced-order models: background and theory

A1 Polynomial chaos expansion

25 Polynomial Chaos Expansion (PCE) is a popular method for approximating stochastic functions of multiple random variables, using an orthogonal polynomial basis. In the classical definition of PCE (Ghanem and Spanos, 1991) the input random variables $\mathbf{X}$ are defined on $(-\infty, \infty)$, with Hermite polynomials typically used as the polynomial basis.[1] Choosing a polynomial basis which is orthogonal to a non-Gaussian probability measure turns the PCE problem into the so-called Wiener-Askey or Generalized chaos, (Xiu and Karniadakis, 2002). For the present problem, a Generalized PCE using Legendre polynomials

30 is considered most suitable as the Legendre polynomials $P_n(\xi)$ are orthogonal with respect to a uniform probability measure in the interval $\xi = [-1, 1]$, which means that the PCE can conveniently be applied on the cumulative distribution functions of

---

[1] In the classical definition of the PC decomposition used in e.g. spectral stochastic finite element methods (Ghanem and Spanos, 1991), the input random variables are normally distributed (Gaussian), which means that the Hermite polynomials are a suitable Hilbertian basis—since the Hermite polynomials are orthogonal with respect to the Gaussian probability measure. In this case, the properties of the Hermite polynomials dictate that the random variables $\mathbf{X}$ are defined on $(-\infty, \infty)$.

the variables $\mathbf{X}$ which are defined in the interval $[0,1]$ so that

$$\xi_i = 2F(X_i) - 1, \tag{A1}$$

where $F(X_i)$ is the cumulative distribution function of a variable $X_i \in \mathbf{X}$, $i = 1, \ldots, M$. With this definition, the PCE represents a model applied to a set of transformed variables which due to the applied transformation are independent and identically distributed ('i.i.d.'). Note that eq. (10) and the evaluation of the cumulative distribution in general does not account for dependence between variables - this has to be addressed by applying an appropriate transformation. In the present case where the joint probability distribution of input variables is defined in terms of conditional dependencies, it is convenient to apply the Rosenblatt transformation as defined in eq. (1). For the current implementation of PCE, only eq. (1) is required since the expansion is based on the Legendre polynomials, however the transformation to standard normal space in eq. (2) is used for other procedures, e.g. the quadratic response surface model discussed later.

Using the notation defined by Sudret (2008), we consider the family of univariate Legendre polynomials $P_n(\xi)$. A multivariate, generalized PCE with M dimensions and maximum polynomial degree p is defined as the product of univariate Legendre polynomials where the maximum degree is less than or equal to p. The univariate polynomial family for dimension i can be defined as

$$P_{\alpha_i}(\xi), \quad \text{where} \quad i = 1, \ldots, M, \quad \alpha_i \in \mathbb{N}, \quad \sum_{i=1}^{M} \alpha_i \leq p. \tag{A2}$$

The multivariate polynomial of dimension M is then defined as

$$\Psi_{\boldsymbol{\alpha}} = \prod_{i=1}^{M} P_{\alpha_i}(\xi_i) \tag{A3}$$

With the above, each multivariate polynomial is built as the product of M univariate polynomial terms, and $\boldsymbol{\alpha}$ vector stores the orders for each univariate polynomial term. The total number of polynomials of this type is (Sudret, 2008):

$$N_p = \begin{pmatrix} (M+p) \\ p \end{pmatrix} \frac{(M+p)!}{M!p!} \tag{A4}$$

The aim of using PCE is to represent a scalar quantity $S = g(\boldsymbol{\xi}(\mathbf{X}))$ in terms of a truncated sequence $\tilde{S}(\boldsymbol{\xi}(\mathbf{X})) + \varepsilon$ where ε is a zero-mean residual term. With this definition, the multivariate generalized PCE of dimension M and maximum degree p is given by

$$\tilde{S}(\boldsymbol{\xi}) = \sum_{j=0}^{N_p - 1} S_j \Psi_{\boldsymbol{\alpha}, j}(\boldsymbol{\xi}) \tag{A5}$$

where $S_j \in \mathbf{S} = [S_1, \ldots, S_{Np}]$ are unknown coefficients which need to be determined, and $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_M]$ are functions of $\mathbf{X}$ as defined in eq. (10). The most straightforward way of determining $\mathbf{S}$ is minimizing the variance of the residual ε using a

least-squares regression approach:

$$\mathbf{S} = \min \left\{ \frac{1}{N_e} \left| \sum_{i=1}^{N_e} \left[g(\boldsymbol{\xi}^{(i)}) - \sum_{j=0}^{N_p-1} S_j \Psi_{\boldsymbol{\alpha},j}(\boldsymbol{\xi}^{(i)}) \right]^2 \right| \right\} \tag{A6}$$

where N_p is the number of polynomial coefficients in the PCE and N_e is the number of sampling points in the experimental design. For this purpose, a design experiment has to be set up and the so-called design matrix $\boldsymbol{\Psi}$ needs to be constructed:

$$\boldsymbol{\Psi}_{ij} = \Psi_{\boldsymbol{\alpha},j}(\xi^{(i)}); \qquad i = 1,\ldots,N_e, \qquad j = 1,\ldots,N_p. \tag{A7}$$

Plugging the definition of $\boldsymbol{\Psi}$ in eq. (A5), the PCE can be expressed as $\mathbf{y} = \boldsymbol{\Psi}\mathbf{S}$. Under the condition that the residuals ϵ are (approximately) normally-distributed, the solution for $\mathbf{S}$ which minimizes the sum of residuals is given by

$$\mathbf{S} = (\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi})^{-1} \cdot \boldsymbol{\Psi}^{\mathrm{T}} \cdot \mathbf{y}, \tag{A8}$$

with $\mathbf{y} = g(\boldsymbol{\xi}^{(i)})$ being a vector with the outcomes of the functional realizations obtained from the design experiment, where $i = 1 \ldots N_e$.

The solution of eq. (A8) requires that the so-called information matrix $(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi})$ is well-conditioned, which normally requires that the number of collocation points N_e is significantly larger than the number of expansion coefficients N_p. Subsequently, the problem grows steeply in size when M and p increase. In such situations, it may be desirable to limit the number of active coefficients by carrying out a Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1996), which regularizes the model by penalizing the sum of the absolute value of model coefficients:

$$\mathbf{S} = \min \left\{ \left| \frac{1}{2N_e} \sum_{i=1}^{N_e} \left[g(\boldsymbol{\xi}^{(i)}) - \sum_{j=0}^{N_p-1} S_j \Psi_{\boldsymbol{\alpha},j}(\boldsymbol{\xi}^{(i)}) \right]^2 + \lambda \sum_{j=0}^{N_p-1} |S_j| \right| \right\} \tag{A9}$$

where λ is a positive regularization parameter; larger values of λ increase the penalty and reduce the absolute sum of the model coefficients, while $\lambda = 0$ is equivalent to ordinary least-squares regression.

A2 Kriging

Kriging (Sacks et al., 1989; Santher et al., 2003) is a stochastic interpolation technique which assumes the interpolated variable follows a Gaussian process. A Kriging metamodel is described (Sacks et al., 1989) by

$$Y(\mathbf{X}) = \mathbf{f}(\mathbf{X})^T \boldsymbol{\beta} + Z(\mathbf{X}), \tag{A10}$$

where $\mathbf{X}$ represents the input variables, and $Y(\mathbf{X})$ is the output. The term $\mathbf{f}(\mathbf{X})^T \boldsymbol{\beta}$ is the mean value of the Gaussian process (a.k.a. the "trend") represented as a set of basis functions $\mathbf{f}(\mathbf{X}) = [f_1(\mathbf{X}),\ldots,f_P(\mathbf{X})]$ and regression coefficients

$\boldsymbol{\beta} = [\beta_1, \ldots, \beta_P]$; $Z(\mathbf{X})$ is a stationary, zero-mean Gaussian process. The probability distribution of the Gaussian process is characterized by its covariance, which for two distinct points in the domain, $\mathbf{x}$ and $\mathbf{w}$ is

$$V(\mathbf{w},\mathbf{x}) = \sigma^2 R(\mathbf{w},\mathbf{x},\boldsymbol{\theta}) \tag{A11}$$

where σ^2 is the overall process variance which is assumed to be constant, and $R(\mathbf{w},\mathbf{x},\boldsymbol{\theta})$ is the correlation between $Z(\mathbf{x})$ and $Z(\mathbf{w})$. The hyperparameters $\boldsymbol{\theta}$ define the correlation behavior, in terms of e.g. a correlation length. Given a set of points $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N]$ where the true function values $\mathbf{y} = Y(\mathbf{X})$ are known, the aim is to obtain a model prediction at a new point, $\mathbf{x}'$. Based on Gaussian theory, the N known values $Y(\mathbf{X})$ and the Kriging predictor $\hat{Y}(\mathbf{x}')$ will be jointly Gaussian distributed:

$$\left\{ \begin{array}{c} Y(\mathbf{x}') \\ Y(\mathbf{X}) \end{array} \right\} \sim \mathcal{N}_{N+1}\left(\left[\begin{array}{c} \mathbf{f}(\mathbf{x}')^T \boldsymbol{\beta} \\ \boldsymbol{\Psi}\boldsymbol{\beta} \end{array} \right], \sigma^2 \left[\begin{array}{cc} 1 & \mathbf{r}^T(\mathbf{x}') \\ \mathbf{r}(\mathbf{x}') & \mathbf{R} \end{array} \right] \right) \tag{A12}$$

Here

$\boldsymbol{\Psi}$ is the design matrix collecting the terms constituting the basis functions $\mathbf{f}(\mathbf{X})$,

$\Psi_{ij} = f_j(\mathbf{x}_i)$ **for** $i = 1\ldots N$ **and** $j = 1\ldots P$

where N is the number of samples and P is the total number of terms output from the basis functions - which may be different than the number of dimensions M as a basis function (e.g. a higher-order polynomial) can return more than one term per variable;

$\mathbf{r}(\mathbf{x}')$ is the vector of cross-correlations between the prediction point $\mathbf{x}'$ and the known points $\mathbf{X}$; and

$\mathbf{R}$ is the correlation matrix of the known points,

$R_{ij} = R(\mathbf{x}_i, \mathbf{x}_j, \theta)$ for $i, j = 1, \ldots, N$.

It follows that the model prediction $\hat{Y}(\mathbf{x}')$ will have the following mean and variance (Santher et al., 2003):

$$\mu_{\hat{Y}}(\mathbf{x}') = \mathbf{f}(\mathbf{x}')^T \boldsymbol{\beta} + \mathbf{r}(\mathbf{x}')^T \mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\beta})$$
$$\sigma_{\hat{Y}}^2(\mathbf{x}') = \sigma^2(1 - \mathbf{r}(\mathbf{x}')^T \mathbf{R}^{-1}\mathbf{r}(\mathbf{x}') + u(\mathbf{x}')^T [\boldsymbol{\Psi}^T \mathbf{R}^{-1}\boldsymbol{\Psi})^{-1} u(\mathbf{x}')]. \tag{A13}$$

where $u(\mathbf{x}') = \boldsymbol{\Psi}^T \mathbf{R}^{-1}\mathbf{r}(\mathbf{x}') - \mathbf{f}(\mathbf{x}')$. Using the predictor functions above requires determining the regression coefficients ($\boldsymbol{\beta}$), the field variance (σ^2), and the correlation hyperparameters ($\boldsymbol{\theta}$). A suitable approach is to find the values of $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$ which maximize the likelihood of $\mathbf{y}$, (Lataniotis et al., 2015):

$$\mathcal{L}(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = \frac{\det(\mathbf{R})^{-1/2}}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\beta})^T \mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\beta})\right]. \tag{A14}$$

Here the hyperparameters, $\boldsymbol{\theta}$, appear within the correlation matrix $\mathbf{R}$. Having set up the design matrix $\boldsymbol{\Psi}$, the expansion coefficients $\boldsymbol{\beta}$ can be determined with the least-squares approach, by solving the equation $\mathrm{d}(-\log\mathcal{L})/\mathrm{d}\boldsymbol{\beta} = 0$ for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = \beta(\boldsymbol{\theta}) = (\boldsymbol{\Psi}^T \mathbf{R}^{-1} \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^T \mathbf{R}^{-1} \mathbf{y}. \tag{A15}$$

Similarly, by solving $\mathrm{d}(-\log\mathcal{L})/\mathrm{d}\sigma^2 = 0$ for σ^2, the field variance is obtained as

$$\sigma^2 = \sigma^2(\boldsymbol{\theta}) = \frac{1}{N}(\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\beta})^T \mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\Psi}\boldsymbol{\beta}). \tag{A16}$$

From eq. (A15) and (A16) it follows that $\boldsymbol{\beta}$ and σ^2 can be expressed as functions of $\boldsymbol{\theta}$. Therefore, calibrating the Kriging model amounts to finding the values of $\boldsymbol{\theta}$ which maximize the likelihood. By combining eqns. A14–A16 this leads to the optimization problem

$$\boldsymbol{\theta} = \arg\min_{\theta} \left(\frac{1}{2} \log\left(\det(\mathbf{R})\right) + \frac{N}{2}\log\left(2\pi\sigma^2\right) + \frac{N}{2} \right). \tag{A17}$$

For a problem with M dimensions, we assume that the correlation between sample points can be modelled using an anisotropic separable correlation function ((Sacks et al., 1989; Lataniotis et al., 2015), which assumes a different correlation parameter for each dimension. The total correlation is expressed as the product of the individual one-dimensional correlation functions,

$$R(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}) = \prod_{i=1}^{M} R(x_i, x_i', \theta_i). \tag{A18}$$

The one-dimensional correlation functions are assumed to follow an exponential relation to the distance $h = (x_i - x_i')$ between points,

$$R(h, \theta) = \exp\left(-\frac{|h|}{\theta}\right). \tag{A19}$$

One of the possibilities for tuning the performance of a Kriging model is selecting different trend functions. If the trend function is replaced by a constant (i.e. the mean of the field) the resulting model is referred to as simple Kriging. A linear trend is denoted as ordinary Kriging, while with any other more advanced function the model is called universal Kriging. In universal Kriging, the functional form of the mean field $\mathbf{f}(\mathbf{x})^T\boldsymbol{\beta}$ is identical to the generalized PCE defined in eq. (A8), meaning that the PCE is a possible candidate model for the mean in a Kriging interpolation. We adopt this approach and define the Kriging mean as a PCE with properties as described in section 4.3.

The main practical difference between regression- or expansion-type models such as regular PCE and the Kriging approach is in the way the training sample is used in the model: in the pure regression-based approaches the training sample is used to only calibrate the regression coefficients, while in Kriging as in other interpolation techniques the training sample is retained and used in every new model evaluation. As a result the Kriging model may have an advantage in accuracy since the model

error tends to zero in the vicinity of the training points; however this comes at the expense of an increase in the computational demands for new model evaluations. The extra computational burden is mainly the time necessary to assemble $\mathbf{r}(\mathbf{x}')$, the matrix of cross-correlations between the prediction points and the training sample, and the time to multiply $\mathbf{r}(\mathbf{x}')$ with other equation terms. Thus, while for a PCE the model evaluation time $t(N)$ for a sample of size N would follow $t(N) = \mathcal{O}(N)$, for

5 a Kriging model $t(N) = \mathcal{O}(N^2)$. For a Kriging model, a gain in accuracy over the model represented by the trend function will only materialize in problems where there is a noticeable correlation between the residual values at different training points. In a situation where the model error is independent from point to point (as e.g. in the case when the error is only due to seed-to-seed variations in turbulence) the inferred correlation length will tend to zero and the Kriging estimator will be represented by the trend function alone.

10 **A3 Sobol indices from the PCE**

One useful corollary of the orthogonality in the PCE polynomial basis is that the total variance of the expansion can be expressed as the sum of the contributions from individual terms (Sudret, 2008),

$$\mathrm{Var}\left[\tilde{S}(\boldsymbol{\xi})\right] = \mathrm{Var}\left[\sum_{j=0}^{N_p-1} S_j \Psi_{\gamma,j}(\boldsymbol{\xi})\right] = \sum_{j=1}^{N_p-1} S_j^2 \mathrm{E}\left[\Psi_{\gamma,j}^2(\boldsymbol{\xi})\right]. \tag{A20}$$

Each of the terms in the sum in eq. (A20) represents the contribution of the variables contained in the respective multivariate

15 polynomials $\Psi_{\gamma,j}$ where $j = 0 \ldots N_p - 1$. This property can be used for eliminating polynomials which do not contribute significantly to the variance of the output, thus achieving a sparse, more computationally efficient reduced model. By combining the variance truncation and the LASSO regression technique in eq. (17), a model reduction of an order of magnitude or more can be achieved. For example, for a 9-dimensional PCE of order 6, using LASSO regularization parameter $\lambda = 1$ and retaining the polynomials which have a total variance contribution of 99.5%, resulted in a reduction of the number of polynomials from

20 5005 to about 200.

Denoting by $\mathscr{F}_{i_1,\ldots,i_s}$ the set of all polynomials dependent upon a specific combination of input variables $\{i_1,\ldots,i_s\}$ (and only on them), the sum of variance contributions over $\mathscr{F}_{i_1,\ldots,i_s}$ normalized by the total variance represents the PCE-based Sobol index with respect to variable set $\mathscr{F}_{i_1,\ldots,i_s}$ (Sudret, 2008):

$$SU_{i_1,\ldots,i_s} = \left(\sum_{j \in \mathscr{F}_{i_1,\ldots,i_s}} S_j^2 \mathrm{E}\left[\Psi_j^2(\boldsymbol{\xi})\right]\right) \cdot \left(\mathrm{Var}\left[\tilde{S}(\boldsymbol{\xi})\right]\right)^{-1}. \tag{A21}$$

25 Based on eq. (A21) it is also straightforward to obtain the total Sobol indices for a given variable γ by summing all $SU_{i_1,\ldots,i_s}$ where $\gamma \in (i_1,\ldots,i_s)$. Note that since each variable appears in multiple cross-terms in the expansion, the contributions of some polynomial coefficients are included multiple times in the total Sobol indices, meaning that the sum of the total indices will typically exceed 1.

The Sobol indices estimated using the above procedure represent the relative contribution to the model variance from

30 variables following the joint input distribution used to calibrate the PCE. In the present case, this distribution would span the

51

uniform variable space of the high-fidelity database defined in Section 2, and the indices will correspond to the load variation within the entire variable ranges as defined in Table 1.