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 baseline 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 refernce? 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 sparcely 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 2^{nd} and 3^{rd} 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 $\boldsymbol{\beta}$, 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	