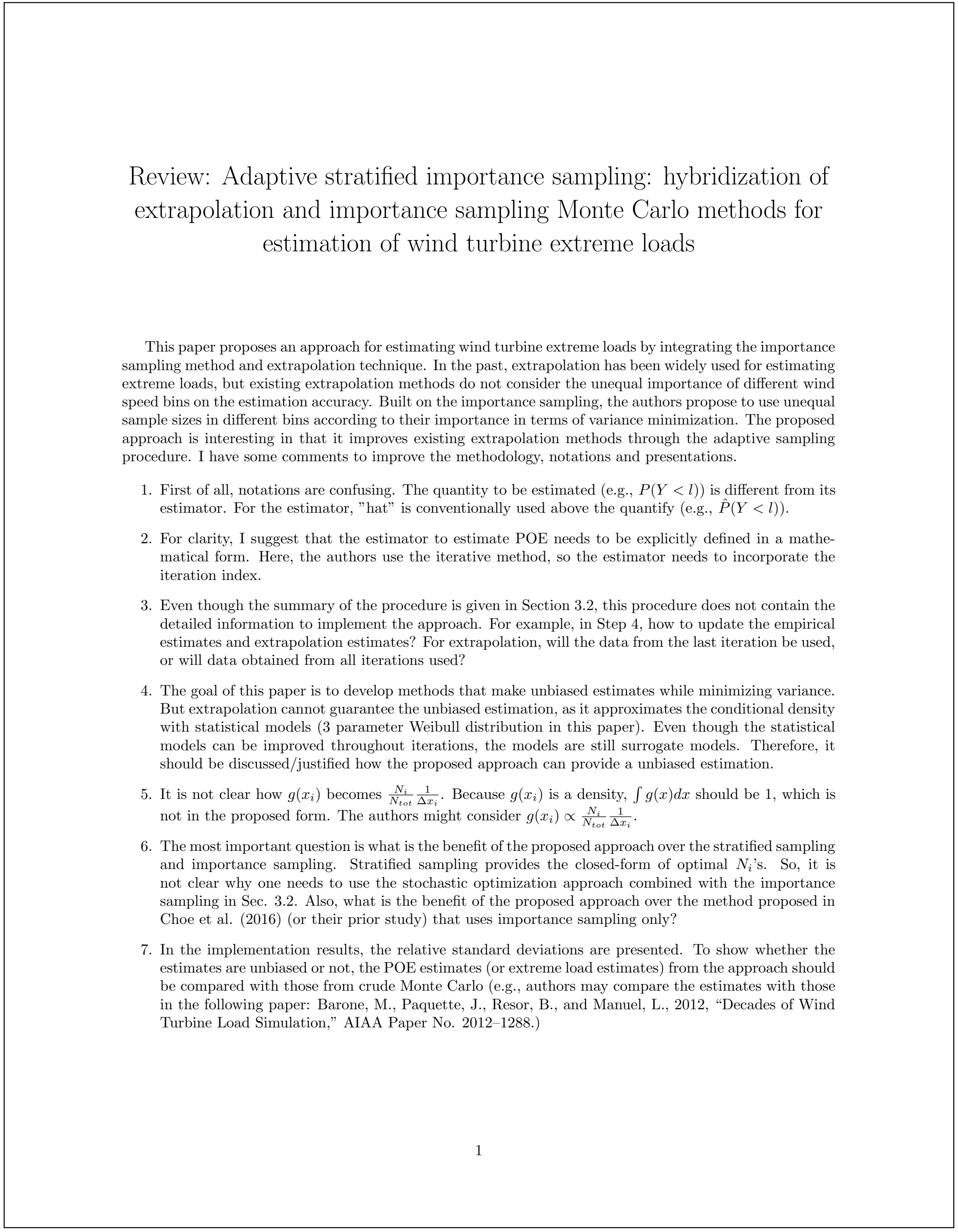
Dear Reviewer (“referee 2”),

Thank you very much for reviewing our paper, “Adaptive stratified importance sampling: hybridization of extrapolation and importance sampling Monte Carlo methods for estimation of wind turbine extreme loads”. We include your comments for reference, followed by point-by-point replies.



Review: Adaptive stratified importance sampling: hybridization of extrapolation and importance sampling Monte Carlo methods for estimation of wind turbine extreme loads

This paper proposes an approach for estimating wind turbine extreme loads by integrating the importance sampling method and extrapolation technique. In the past, extrapolation has been widely used for estimating extreme loads, but existing extrapolation methods do not consider the unequal importance of different wind speed bins on the estimation accuracy. Built on the importance sampling, the authors propose to use unequal sample sizes in different bins according to their importance in terms of variance minimization. The proposed approach is interesting in that it improves existing extrapolation methods through the adaptive sampling procedure. I have some comments to improve the methodology, notations and presentations.

1. First of all, notations are confusing. The quantity to be estimated (e.g., $P(Y < l)$) is different from its estimator. For the estimator, "hat" is conventionally used above the quantify (e.g., $\hat{P}(Y < l)$).
2. For clarity, I suggest that the estimator to estimate POE needs to be explicitly defined in a mathematical form. Here, the authors use the iterative method, so the estimator needs to incorporate the iteration index.
3. Even though the summary of the procedure is given in Section 3.2, this procedure does not contain the detailed information to implement the approach. For example, in Step 4, how to update the empirical estimates and extrapolation estimates? For extrapolation, will the data from the last iteration be used, or will data obtained from all iterations used?
4. The goal of this paper is to develop methods that make unbiased estimates while minimizing variance. But extrapolation cannot guarantee the unbiased estimation, as it approximates the conditional density with statistical models (3 parameter Weibull distribution in this paper). Even though the statistical models can be improved throughout iterations, the models are still surrogate models. Therefore, it should be discussed/justified how the proposed approach can provide a unbiased estimation.
5. It is not clear how $g(x_i)$ becomes $\frac{N_i}{N_{tot}}\frac{1}{\Delta x_i}$. Because $g(x_i)$ is a density, $\int g(x)dx$ should be 1, which is not in the proposed form. The authors might consider $g(x_i) \propto \frac{N_i}{N_{tot}}\frac{1}{\Delta x_i}$.
6. The most important question is what is the benefit of the proposed approach over the stratified sampling and importance sampling. Stratified sampling provides the closed-form of optimal N_i's. So, it is not clear why one needs to use the stochastic optimization approach combined with the importance sampling in Sec. 3.2. Also, what is the benefit of the proposed approach over the method proposed in Choe et al. (2016) (or their prior study) that uses importance sampling only?
7. In the implementation results, the relative standard deviations are presented. To show whether the estimates are unbiased or not, the POE estimates (or extreme load estimates) from the approach should be compared with those from crude Monte Carlo (e.g., authors may compare the estimates with those in the following paper: Barone, M., Paquette, J., Resor, B., and Manuel, L., 2012, “Decades of Wind Turbine Load Simulation,” AIAA Paper No. 2012–1288.)

Replies:

1. Thank you for pointing out the convention, and I am sorry if the current notation is not clear. I am aware of the convention to use “hat”. However, I do notice, even in well-regarded statistical texts (e.g. *Monte Carlo Statistical Methods*, Robert and Casella), a certain lack of consistency, that seems to derive largely from the difference between statements such as 1) “The estimate of interest is $X = E_f[Y] = \int Y(x)f(x)dx \sim \sum_i Y(x_i)f(x_i)$”, where we have switched from exact to estimate in the middle of the series of steps, where it would be wrong to start with “$\hat{X} = ...$”, and 2) “The estimate of $X = E_f[Y] = \int Y(x)f(x)dx$ is $\hat{X} = \sum_i Y(x_i)f(x_i)$”. In an effort to be consistent, and a personal preference for less rather than more symbols, and for using statements such as 1) above to keep the equations “flowing”, I have simply omitted “hats”. I have 1) gone through the text and tried to be sure we are clearly indicating when something is an estimate versus an exact quantity and 2) rephrased the initial description of “exceedance probability” to include the “hat”.

2. We have added language referring to the explicit equations one would use to form the empirical estimate at each iterations. I agree that we could also label the estimates at each iteration with an iteration index, e.g., something like $P^k(Y < l)$, where k is the ASIS iteration. I think this, like the hats, represents the introduction of more symbols than are necessary for the level of detail we are currently targeting. We plan to write a shorter “recipe-like” companion paper in which the algorithm is very clearly stated in terms that the reader could readily implement.

3. This is admittedly just a sketch of the algorithm. To answer your specific question, the data used at each iteration is cumulative. We have added language to that effect to the outline. This paper is about opening new ground (specifically, seeing how an initial bin-based conception can be transformed into stochastic search for the optimal importance density), and the details of the tuned version of this algorithm are still being worked out. As stated above, at this point we would intend to write another paper in which is much closer to a simple recipe for how to implement the recommended algorithm.

4. Perhaps we have not been clear. The ASIS method is a rigorously justified unbiased importance sampling estimate. As you point out, the extrapolative estimates are emphatically not unbiased estimates. We will make sure this is clearly stated in the text.

5. We claim $g(x)$ is in fact normalized:

$$\begin{aligned}
\int g(x)dx &= \sum_{i=1}^{N_{bins}} \int_{x_i}^{x_{i+1}} g(x_i)dx \\
&= \sum_{i=1}^{N_{bins}} \int_{x_i}^{x_{i+1}} \frac{N_i}{N}\frac{1}{\Delta x}dx \\
&= \sum_{i=1}^{N_{bins}} \frac{N_i}{N}\frac{1}{\Delta x}\Delta x \\
&= \sum_{i=1}^{N_{bins}} \frac{N_i}{N} \\
&= 1.
\end{aligned}$$

We now explain this in the text.

6. In Choe et al (2016 and prior work), one will notice that the closed form optimal N_i contains a term, $s(x) = P(Y > l|X = x)$, that is not known in practice, and in fact is the quantity that they are trying to estimate. To overcome this difficulty, Choe et al run a certain (in there case I believe 250) number of simulation, from which they build a surrogate model of this function $s(x)$, only after which they can use their optimal N_i values. Also, the optimal N_i depends on what load value l they are interested in. Importance sampling is a way of estimating with respect to one distribution by sampling from another. But adaptive importance sampling is made difficult by the need to keep track of the exact importance distribution even while it is changing. The use of a bin-based ("stratified") formulation allows for precisely defining the importance distribution at each iteration. This amounts to a form of "parameterization" of the importance distribution. Our main contribution is not as much in the optimization of the bin distribution but more in deriving a formulation that explicit converts any bin distribution into an unbiased IS estimate.

7. We have done this in our earlier paper: Graf, P., Damiani, R., Dykes, K., and Jonkman, J.: Advances in the Assessment of Wind Turbine Operating Extreme Loads via More Efficient Calculation Approaches, in: AIAA SciTech 2017.

Again, thank you very much for the careful and thoughtful review of our paper. I hope I have addressed your concerns. Your comments helped improve the paper considerably.

Sincerely,

Peter Graf, on behalf of the authors.

Dear Dr. Dimitrov ("referee 3"),

Thank you very much for the thoughtful and thorough review of our paper, "Adaptive stratified importance sampling: hybridization of extrapolation and importance sampling Monte Carlo methods for estimation of wind turbine extreme loads". I know it takes a lot of time to provide such feedback. I have really learned from your comments. In particular, connecting work on wind turbine extreme loads to the larger body of literature on structural reliability is important, and I would aspire to do a better job of that going forward. What follows are your specific comments, followed by my replies.

1) As the authors say on Page 2, line 9: "Monte Carlo methods: exceedance probabilities are written as expectations of indicator functions". This is in fact a passing definition for structural reliability problems, regardless of the method used for the integration (the computation of the expected value). We can therefore make a parallel between the statistical extrapolation and the structural reliability analysis problems. This means we can take inspiration from literature where e.g. adaptive importance sampling is used as a tool for reliability analysis. Some examples include [1], [2] where search-based (adaptive) importance sampling is formulated and also [3] where it is applied to a dynamical system.

Certainly, our work is in the category of structural reliability, so these references are relevant. I want to be clear that the ASIS method is NOT extrapolation. The common ground is the bin-based formulation that allows for samples that would "traditionally" be used for extrapolation to also be used for importance sampling, which are non-extrapolative unbiased estimates. Thank you for providing these excellent background references on adaptive importance sampling. I am including these in the background material.

2) The current results section mainly shows the behaviour of the ASIS algorithm, but it is impossible to judge how it will compare to standard extrapolation procedures using the same number of FAST simulations. The algorithm suggested by the authors is essentially an approach for "guided" sampling from the joint distribution of environmental conditions. It would be really interesting to see a comparison with e.g. direct sampling from the long-term distribution using a pseudo Monte-Carlo simulation with low-discrepancy series, followed by extrapolation directly from the empirical CDF of the long-term load distribution.

Comparing ASIS to "standard extrapolation" is a bit of an apples-to-oranges comparison. Extrapolation relies on assuming and fitting an extreme value distribution, after which loads corresponding to arbitrarily small exceedance probabilities can be estimated. Whereas, ASIS is specifically formulated to provide unbiased (Monte Carlo) estimates of the exceedance probabilities, where the price we pay for the lack of assumption of any particular extreme value distribution is that we have to "earn" the low exceedance probabilities by perhaps more samples. I think your suggestion ("*pseudo Monte-Carlo simulation with low-discrepancy series, followed by extrapolation directly from the empirical CDF"*) represents the best "competition" to ASIS (or, in general, any form of adaptive sampling) from a non-adaptive approach, so indeed it would be interesting to try. However, this experiment is beyond the scope of the *present* paper, where the intent was largely to show how the bin-based approaches that are typically associated with extrapolation could be adapted for use in importance sampling, rather than to claim we have found the strictly best approach.

3) *Statistical extrapolation is a very academic exercise. In the last years the wind energy scientific society has published multiple papers which present attempts at improving the efficiency and reducing the uncertainty in load extrapolation. But the majority of these solutions represent highly complex algorithms with many parameters to tune and with tricky implementation which can easily be done in a wrong way. My general concern with statistical extrapolation is that this complexity hinders the adoption of new research in industry as it requires a significant level of very specific expertise in order to achieve a correct and robust implementation. As a confirmation to this observation, the coming IEC 61400-1 ed.4 standard actually suggests methods for doing extrapolation with even lower complexity (and most probably higher uncertainty) than what was suggested in ed.3 of the same standard. This should not be viewed as a criticism to the present paper or its authors – but in my opinion the focus of future research in statistical extrapolation should not only be towards improving accuracy or efficiency, but also towards achieving robust and easy to implement solutions. This is something the authors may want to give a thought to and eventually discuss in their paper.*

This is a great point. However, 1) the paper *is* an academic exercise, 2) the codes that simulate the dynamics of the turbine response are also very complex. Perhaps what is needed is not a simplified procedure but a set of open source community codes that perform the statistical analysis. That said, although this paper is reaching to develop a new method, resources permitting, we intend to write a companion paper that distills ASIS to a simple, easily implementable, “recipe”.

4) Page 5, line 24: The authors use 10 peaks per 10-minute simulation. However, they do not seem to take any measures for ensuring statistical independence between the peaks (or there is no description of any measures they have taken). Thus, there is a very high chance that some of the events they have identified as peaks are correlated (e.g. they have small time separation). The result is most likely a (slight) over-prediction of the exceedance probabilities. A simple and straightforward measure for reducing dependency could be enforcing some minimum time separation between successive 1-minute peaks. Typically time separation equivalent to the time for 1-2 rotor revolutions should be sufficient to eliminate dependence almost entirely.

In an effort to prevent the analysis from being more complicated than it already is, we have made the assumption that peaks from each 1 minute segment are independent. We would argue that the error in the resulting estimates resulting from peaks that are less than 1-2 rotations apart but happen to fall on either side of a 1 minute division is not significant compared to the variation due to 1) stochastic turbine response 2) lack of full convergence of the Monte Carlo estimates.

5) Figure 1: The amount of variability in the peaks is also a consequence of the authors’ choice to use one-minute peaks. Drawing the same plot for the 120 ten-minute peaks per bin would show lower variation.

This is a fair point, given that 9 out of our 10 one-minute peaks will be smaller than the ten-minute peak. But we argue (anecdotally) that the reduction in variability using only ten-minute peaks is much smaller than the variability itself. There is a limit, we admit; for example, using the peaks every five-seconds would be misleading (there would be lots of small peaks), but one-minute peaks are defensible, and helps us develop method faster because we get 10 times as much data to analyze.

6) Section 2.2: Some good additional references to consider in this section are [4] which presents an excellent method for accounting for independence between peaks, [5] which is an extensive parametric study on extrapolation considering, among other things, the effect of number of peaks on extrapolation accuracy, and [6] where the seed-to-seed variability is addressed explicitly, though with focus on fatigue loads.

These are excellent and appreciated suggestions. The literature on peak independence, numbers of peaks, thresholds for peak-over-threshold, appropriateness of difference statistical distributions, etc., is indeed extensive. Thank you for the suggestions, we have added the references suggested.

7) Section 2.4: I think this section can be removed or shortened significantly without reducing the quality of the paper. To me, what the authors try to define in the section is that the extrapolation per wind speed represents a set of conditional approximations which can best be described by a Rosenblatt transformation [7]. The IFORM method is related to the Rosenblatt transformation as the Rosenblatt transformation is a convenient way to draw the IFORM contours. Further, the IFORM method is meant to be applied to static quantities as e.g. 10-minute statistics of environmental conditions. The inventor of the method (S. Winterstein) has also worked on re-formulating the IFORM for dynamic problems as the one considered by the authors of the present paper [8]. Another relevant method is the tail-equivalent linearization method (TELM) [9], which employs FORM for approximation of the extreme response of dynamical systems. In general my feeling is that discussing FORM or IFORM in the current version of this paper is a distraction from the main scope – but if it is necessary to retain it, the TELM method could be a possible bridge.

We have explicitly called out this section as “optional”, as you are not the only reviewer to think it does not fit in that well. The point was about response variability: if the same environmental conditions can result in very different response (due to “random seeds” used, e.g., to generate turbulent inflow), then there is very little that can be done to reduce variance of estimates other than run large numbers of simulations. In our discussion, this is related to IFORM through the “environmental contour” method. Thank you for the suggestion regarding TELM, that is an interesting approach. Undoubtedly many interesting connections could be made between the question of “number of peaks to use” which is a proxy for “threshold” in peak detection, and the TELM method, because, like TELM, we are using only the “tail” data to fit our extrapolation distribution. This field, we feel, has the potential for much of this sort of unification of methods previously assumed to be disparate (this is exactly what we have done in in building a “bridge” from bin-based extrapolation to importance sampling), but it is beyond the scope of the present paper.

8) Page 11, lines 15-25: As the authors state, this algorithm will converge to a local optimum, and measures should be taken to ensure that the global optimum is found. We could again make the parallel to structural reliability, where multiple failure modes lead to a system reliability problem with multiple design points ("local minima") and the failure domain may be non-convex. When employing (adaptive) importance sampling, the system reliability problem is normally approached by using multiple "seeds" of importance sampling densities which are initially placed in the different parts of the variable domain. Melchers with multiple design points ("local minima") and the failure domain may be non-convex. When employing (adaptive) importance sampling, the system reliability problem is normally approached by using multiple "seeds" of importance sampling densities which are initially placed in the different parts of the variable domain. The authors may want to consider such an approach for their problem – it can potentially also help with the issue of the largest variance being at different wind speeds for different load channels.

The non-convexity of the search for the optimal bin-distribution is different from (but probably complicated by) the interest in making estimates for different load channels at the same time. Our interest in this paper has been more toward formulating the extreme loads estimation problem as stochastic/global search, not necessarily in providing the optimal solution algorithm. Therefore, we have simply adopted the common strategy for global search (e.g., simulated annealing) of allocating a certain percentage of the search steps to "exploration" rather than "exploitation" (gradient descent). It is certainly a good and intuitive idea to build the overall IS distribution from individual distributions targeted toward different load channels.

9) Figure 2: A good reference for the effect of number of peaks (time series) on the extrapolation is [5], where this is investigated for several types of extrapolation methods, distribution fits, and with up to 1000 minutes of time series per extrapolation. Another relevant study is the one by Zwick & Muskulus [6] where they thoroughly investigate the seed-to-seed variability problem, though without considering extrapolation.

Thanks for these suggestions. We refer to these now at appropriate points in the text.

Technical comments:
*10) Page 7, line 28: Y*f] is probably a typo. Did the authors mean [Y*f]?*

Indeed, this is a typo, thank you.

Thank you very much for taking the time to prepare such a thoughtful review of our paper, in particular for providing such a wealth of references for us to learn from. They have certainly allowed us to improve the level of discussion in the paper. I apologize for not being able to pursue all your suggestions thoroughly in the present paper, but they will be very helpful going forward.

Sincerely,

Peter Graf, on behalf of the authors.