Response to RC1

Thank you for the helpful comments; they are much appreciated. We have prepared a new version that hopefully addresses the open issues.

General comments:

1. The distribution fit to a random process is representative only if the process is stationary and ergodic (i.e., its statistical properties can be sufficiently well determined over a given sampling period). Subsequently stationarity and ergodicity are central requirements for the validity of load extrapolations. In the present paper, these conditions are not taken into account and I am afraid that they are violated in many of the cases. As an example, for samples drawn from a Rayleigh distribution with mean μ, the standard deviation of the sample mean is defined as $\sigma_{\mu_s} = \mu\sqrt{(4-\pi)/(N\pi)}$, where N is the sample size. For μ = 8.5 m/s as in the present paper and for sample size N = 1000, the standard deviation of the sample mean wind speed will be 0.14 m/s, which means that the 95% confidence interval will be 0.55 m/s wide. In effect, samples of this size may be quite different from each other and will result in different extrapolations. This sampling uncertainty arises due to the sample size being too small to properly represent the statistical properties of the parent distribution, and not due to the quality of any subsequent distribution fit. The quality of the fit should relate instead to the realization-to-realization uncertainty for samples drawn from the same stationary process. To summarize – the uncertainty sources should be distinguished and their effect accounted for separately, or the sampling statistical uncertainty should be eliminated by drawing samples with the same wind statistics.

   The uncertainty arising from the sampling error is exactly what we wanted to highlight. Perhaps Figure 5 gave the impression that the error was computed by fitting many distributions to a single sample. We have clarified that in the caption.

   We also did not want to go into too much detail here, especially since the focus was on the consequences of the uncertainty rather than on the statistical method itself. Our goal with this paper was to come up with an extrapolation method that was easy to follow for readers struggling with statistical methods, and to challenge them with the consequences of the uncertainty through a simple example. That being said, we understand that perhaps presenting alternative approaches to ours might improve the paper. In the new version, we have gone through the following 8 workflows:

	Sampling method (distribution)	Distribution fit	Fitted only above threshold
(a)	Aggregation-before-fitting (Rayleigh)	Gumbel	No
(b)	Aggregation-before-fitting (Rayleigh)	Gumbel	Yes
(c)	Aggregation-before-fitting (Rayleigh)	GEV	No
(d)	Aggregation-before-fitting (Rayleigh)	GEV	Yes
(e)	Fitting-before-aggregation (Uniform)	Gumbel	No
(f)	Fitting-before-aggregation (Uniform)	Gumbel	Yes
(g)	Fitting-before-aggregation (Uniform)	GEV	No
(h)	Fitting-before-aggregation (Uniform)	GEV	Yes

2. The extrapolation procedure analysed by the authors is not necessarily consistent with the actual design approach used by manufacturers. According to the IEC 61400-1 standard, it is acceptable to define the long-term distribution of loads according to two approaches: 1) first carry out extrapolations for simulations binned according to wind speed, and then aggregate the long-term load distribution based on the extrapolated functions (so-called "extrapolate, then aggregate" approach).

   In the original manuscript, we decided to go with approach 2) to match the workflow with which the database was generated (Barone et al., 2012). In the new version, we have also included a fitting-before-aggregation approach to be more consistent with IEC guidelines (see point 1).

3. There is no discussion on the effect of the variation of the wind speed in the original MC sample, and in the random subsets used to test the uncertainty in extrapolation techniques. This variation means that in each sample there are a lot more simulations at wind speeds around the mean, but only few at high wind speeds – e.g., for a sample of size 1000, only 13 simulations on average will represent wind speeds above 20m/s.

   The variation in the original MC sample (from Sandia) should be clear from Figures 1 and 2. We have also added a small remark on the variation of wind speeds in any small subsample and how it compares to other sampling methods.

4. What about other load cases? Very often, if the design loads from one load case (e.g. DLC1.1) are low, another load case (e.g. emergency shutdown with gust, DLC4.2) will become design-driving. So a lower design load prediction in DLC1.1 will not necessarily lead to lower material thickness, it will simply eliminate the load case as a potential design driver.

   We did not take into account other load cases as our primary focus was on DLC 1.1. A designer still has to go through DLC 1.1 in order to discard it as a non-driving load case. Moreover, the results can also be applied to cases where loads are extrapolated from more realistic conditions with multiple variables (e.g., wind speed/wave height/atmospheric stability). In any case, the rather artificial, yet possibly design-driving DLC's in the IEC standards should not make the extrapolation problem less interesting.

5. Page 10, lines 16-20, the authors conclude: "In any case, we can conclude that the bare minimum of 300 minutes of time series, as prescribed in Appendix F of the standards (IEC, 2005), is not sufficient to produce any reasonable 50-year estimate. Based on this load set and this extrapolation produce, one should instead aim for sample sizes larger than $N = 10^5$." I do not agree with this conclusion. Due to the issues outlined in my comments above, some additional uncertainty is present in the extrapolation. If these uncertainties are eliminated, or if another extrapolation procedure (e.g. aggregation after fitting) is used, the accuracy of the extrapolation may improve and the required sample size will become smaller.

Apart from the approach used in the original manuscript, we have added some other approaches to strengthen the base for our conclusions (see point 1). We have also weakened our statement on the 300 minutes a bit. Still, we can at least stick by our conclusion that 300 minutes is a very low minimum, especially considering modern-day computing power.

Specific comments:

6. Page 3, last paragraph: "the wind speeds belonging to the 10% highest loads points towards a region well above the rated wind speed (see Figure 3)". I don't think this statement agrees with Figure 3, where the mode of the distribution $f\left(\overline{U}\middle|\hat{F} \geq 0.9\right)$ is actually at wind speeds just below rated.

   The mode of the distribution $f\left(\overline{U}\middle|\hat{F} \geq 0.9\right)$ is around 13.5 m/s, well above the rated wind speed of 11.4 m/s. In fact, if one were to take the top 1% or top 0.1% highest loads from Figure 1, they originate from increasingly higher wind speeds.

7. Page 4, line 5: Why is the GEV distribution a good candidate? In my experience, the GEV, or other 3- parameter representation as the 3-parameter Weibull, are good candidates for extrapolation in situations where only few data points are available and not all of them belong to the upper tail of the extremes distribution. However, if the amount of data and the threshold selection result in a data set which is predominantly from the upper tail which has a characteristic log-linear behaviour, I would think that a 2-parameter Gumbel distribution would provide more robust and accurate fit.

   The GEV is a near perfect match to the tail for large sample sizes. Since characteristic log-linear behavior is not always found (for example the yawing moment in Barone, 2012), we wanted to avoid the issue of having to need prior knowledge about the tail. Nevertheless, the Gumbel distribution is usually more forgiving at small sample sizes. We have included it in the new version as an alternative method (see point 1).

8. Page 5, line 5: "any extrapolated 50-year loads that are more than 50% higher than the "real" value are discarded and resampled": - This data censoring approach is quite crude. It also relies on the known true value which will not be known in practice. Given the large number of samples, it is possible to establish a confidence interval for each load level and discard the outliers, see e.g. Naess and Gaidai, Structural Safety 31 (2009), pp. 325-334.

   We have left the outliers in the new version to avoid having to know the true value.

9. Figure 9: It seems that the uncertainty for sample size N = 100 is smaller than the uncertainty for N = 1000. This is against the logic and indicates a possible sampling bias. Could the authors find an explanation for this?

   This is because how the threshold is placed at very small sample sizes. For N < 100, the data above the threshold includes more of the knee in order to have enough points for a fit. The data points

are then often lined up in a slightly downward curve, which favors a GEV with a negative ξ (thus fewer outliers). For N > 1000, the threshold has moved further up the tail, which is close to being straight and favors both positive and negative ξ.

Technical comments:

10. Page 8, lines 5-10: It does not get entirely clear how the stress criteria for accepting or rejecting a given design relate to the extrapolated loads. Please improve the explanation.

    The stress criteria was purely used as a simple example to demonstrate a decision making process. We have clarified this a little bit more in the new version.

Response to RC2

Thank you for your comments and suggestions. They have certainly helped us to prepare a new version of the manuscript.

1. What is the procedure for obtaining the 50-year load from the data? There is some uncertainty associated with this value, and the authors should attempt to estimate it (for example, from a bootstrap procedure).

   The 50-year load is obtained from a GEV fit to the tail. We have clarified that in the new version. We have also estimated the 95% confidence level around the 50-year value by bootstrapping (13.1-17.2 MN m with a median at 115.0 MN m), and have added it in the new version.

2. It is not clear how the results might differ should one choose a different distribution other than GEV for the fit. While an exhaustive study of many possible distributions is not required, some discussion and exploration of this is in order

   In the new version, we have explored some different approaches to the extrapolation problem, including a Gumbel fit. This is more robust and forgiving at small sample sizes, but requires one to assume log-linear behavior in the tail.

List of relevant changes to the manuscript

page 3

line 6-7: Added an estimate of the uncertainty around the 50-year load, given 96 years of data (RC2-1).

line 8-11: Discussed the pros and cons of a crude Monte Carlo method (RC1-3).

line 12-15: Introduced the fitting-before-aggregation method (RC1-2).

page 4

Figure 3: Added return level plot for selected bins of the fitting-before-aggregation method (RC1-2).

page 5

line 5-9: Introduced the Gumbel distribution as an alternative distribution fit (RC1-7, RC2-2).

page 7

Table 1: Added several extrapolation cases (RC1-1).

page 8

line 8-15: Briefly discussed the difficulties of extrapolating with the aggregation-before-fitting approach.

page 9

line 1-3: Briefly discussed the difficulties of extrapolating with the fitting-before-aggregation approach.

line 13-17: Selected 2 approaches for further discussion.

line 19-20: Clarified a bit that the stress criteria is a simple example of a decision making process during a design phase.

page 10

Figure 8: Added the error with respect to the “true” 50-year load for all 8 approaches.

page 11

Figure 9: Compared all approaches in terms of the RMS error at various sample sizes.

page 13

Figure 12: Repeated the results of Figure 11 to the fitting-before-aggregation approach.

page 15

Figure 16: Repeated the results of Figure 15 to the fitting-before-aggregation approach.

line 13-15: Weakened the statement of the 300-min minimum a bit (RC1-5).

page 16

line 5-10: Updated the conclusions.

The risks of extreme load extrapolation

Stefan F. van Eijk[1], René Bos[1], and Wim A. A. M. Bierbooms[1]

[1]Wind Energy Research Group, Faculty of Aerospace Engineering, Delft University of Technology, 2629 HS Delft, The Netherlands

Correspondence to: Stefan van Eijk (sfvaneijk@gmail.com)

Abstract. An important problem in wind turbine design is the prediction of the 50-year load, as set by the IEC 61400-1 Design Load Case 1.1. In most cases, designers work with limited simulation budgets and are forced into using extrapolation schemes to obtain the required return level. That this is no easy task is proven by the many studies dedicated to finding the best distribution and fitting method to capture the extreme load behavior as well as possible. However, the issue that is often overlooked is the effect that the sheer uncertainty around the 50-year load has on a design process. In this paper, we use a collection of 96 years' worth of extreme loads to perform a large number of hypothetical design problems. The results show that, even with sample sizes exceeding $N = 10^3$ ten-minute extremes, designs are often falsely rejected or falsely accepted based on an over- or underpredicted 50-year load. Therefore, designers are advised to be critical of the outcome of DLC 1.1 and should be prepared to invest in large sample sizes.

1 Introduction

Wind turbine designers are confronted with the IEC 61400-1 Design Load Case 1.1 (IEC, 2005). This evaluates the structural integrity of the major load-carrying components on the basis of a 50-year return level, plus safety factors. As prescribed by Appendix F of the standards, a minimum of 300 minutes of time series—distributed over the relevant wind speeds—will have to be evaluated and followed by an extrapolation scheme to obtain the 50-year return level. Such extrapolations produce notoriously uncertain estimates, which is why the Design Load Case (DLC) 1.1 is often avoided or at least greatly simplified in early stages of the design. However, in cases where DLC 1.1 is design-driving (e.g., for foundations and controllers), dealing with this uncertainty is unavoidable.

Many past efforts to reduce this uncertainty have focused on trying out different sampling methods (Fogle et al., 2008; Agarwal and Manuel, 2009), new modeling techniques (Moriarty et al., 2004; Bos and Veldkamp, 2016), or finding the best distribution type to match the extreme load behavior (Pandey and Sutherland, 2003; Genz et al., 2006; Freudenreich and Argyriadis, 2007; Ragan and Manuel, 2007; Natarajan and Holley, 2008; Peeringa, 2009; Lott and Cheng, 2016). Yet, because most studies deal with relatively small sample sizes (e.g., $\ll 1$ year), the actual uncertainty that surrounds the 50-year return level is often underexposed. With a 63-year data set, Barone et al. (2012b) were able to establish the 90% confidence interval around the 50-year load for sample sizes up to $N = 2,000$ ten-minute maxima. This revealed not only that the 50-year levels are clouded by high uncertainty, but also that they suffer from a considerable bias. Inevitably, this has an effect on the choices made during design.

The aim of this paper is to demonstrate this with a simple exercise, using a collection of 96 years' worth of ten-minute load maxima released by Barone et al. (2012a). The uncertainty distribution is constructed by repeatedly sampling subsets of this data set and obtaining the 50-year loads through an automated extrapolation scheme. We then simulate a problem where a hypothetical designer has to choose between two or more concepts and record how often this uncertainty leads to wrong choices. The results of this paper should help designers to estimate the required sample sizes for their problem, but also to form a critical attitude concerning the quality and reliability of extrapolated 50-year loads.

2 Methodology

Since the focus of this work is on the impact of uncertainty, rather than obtaining the highest possible quality result, the workflow is kept as simple as possible. Loads were extracted by drawing a random sample from a large set of crude Monte Carlo results and the 50-year return period is found by a graphical fit.

2.1 Loads data set

The data set that was used for this study was generated by Barone et al. (2012a). It features the onshore version of the NREL 5 MW reference wind turbine, operating for 96 years in an IEC class 1B climate (IEC, 2005).[1] Ten-minute mean wind speeds were randomly drawn from a Rayleigh distribution, bounded by the cut-in and cut-out wind speeds of 3 and 25 m/s, respectively. Turbulent wind fields were generated by TurbSim on a 20×20 grid with a width and height of 137 m and were fed to the FAST v7 aeroelastic code. Every simulation ran for eleven minutes, of which the first minute was discarded to avoid any start-up transients. More details can be found in the original paper.

Each output channel contains over 5 million ten-minute extremes. In this paper, we will use the tower base overturning moment, which plays a major role in the design of foundations. Figure 1 shows the entire set of loads at the respective wind speeds.

2.2 Extrapolation scheme

In many practical situations, a designer does not have the computational resources available to simulate several decades of operation. That is when the 50-year load has to be found by extrapolating.

2.2.1 Aggregation-before-fitting and fitting-before-aggregation

There are several approaches to the extrapolation problem. One method involves drawing a sample directly from the parent mean wind speed distribution. The cumulative distribution of extreme loads then follows naturally from ranking a set of N loads and assigning a plotting position:

$$\hat{F}(M_i) = \frac{i}{N+1}. \tag{1}$$

[1] The original paper specifies a class 2B site, but this has been corrected with the release of the data set (see http://energy.sandia.gov/?page_id=13173).

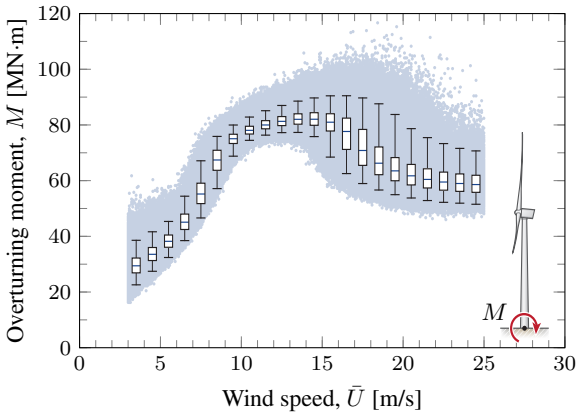

Figure 1. The data set, containing over 5 million ten-minute extreme overturning moments between the cut-in and cut-out wind speeds. The box plots indicate the scatter per 1-m/s bin, where the boxes mark the 25th and 75th percentiles, the whiskers mark the 2.5th and 97.5th percentiles, and the bar is the median.

In this case, however, the wind speeds outside of the operating range will have to be accounted for:

$$\hat{F}(M_i) = 1 - \left(1 - \frac{i}{N+1}\right) \int_{\bar{U}_{\text{cut-in}}}^{\bar{U}_{\text{cut-out}}} f(\bar{U})\,\mathrm{d}\bar{U}, \tag{2}$$

where $f(\bar{U})$ is the mean wind speed distribution. Then, plotting the entire data set yields the return level plot shown in Figure 2. Extrapolation is done by using the entire sample, called *aggregation-before-fitting*. With 96 years' worth of load data, however, the 50-year return value can be easily matched with a Generalized Extreme Value (GEV) distribution (see next subsection), which yields 115.0 MN·m. Repeating the process by randomly drawing 96-year samples from the same data set allows us to estimate the 95% confidence interval, yielding [113.1, 117.2] MN·m.

Sampling directly from the parent distribution is an example of a *crude Monte Carlo* method, which has the advantage that it gives a raw and unbiased picture of the extreme loads. However, a clear disadvantage is that the bulk of the data originates from relatively low wind speeds where the extremes loads are not expected to lie. In addition, unless a stratified sampling method is used, the wind speeds in any small subsample are not always representative of the parent distribution.

Another method, which is preferred by the IEC guidelines (IEC, 2005), requires the data to be collected in n wind speed bins of a certain width, $\Delta\bar{U}$. Data from every bin is then matched with a distribution function, after which every distribution is weighted according to

$$\hat{F}(M) = 1 - \int_{\bar{U}_{\text{cut-in}}}^{\bar{U}_{\text{cut-out}}} f(\bar{U})\,\mathrm{d}\bar{U} + \sum_{i=1}^{n} F(M|\bar{U}_i) \int_{\bar{U}_i - \frac{1}{2}\Delta\bar{U}}^{\bar{U}_i + \frac{1}{2}\Delta\bar{U}} f(\bar{U})\,\mathrm{d}\bar{U}. \tag{3}$$

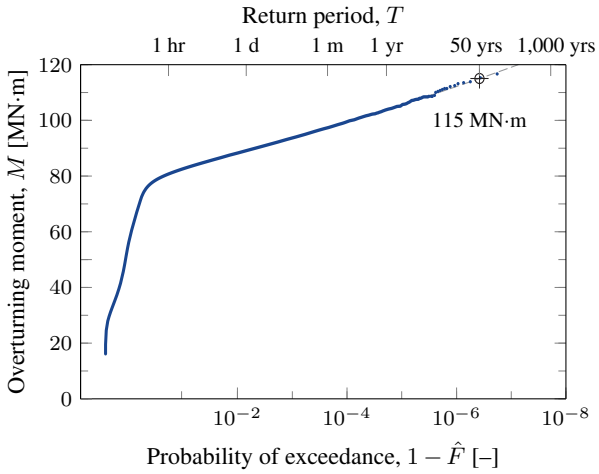

Figure 2. Return level plot of the tower base overturning moment, with the entire 96-year data set (i.e., aggregation-before-fitting). A GEV fit above the threshold given by Equation (6) yields a 50-year value of 115.0 MN·m.

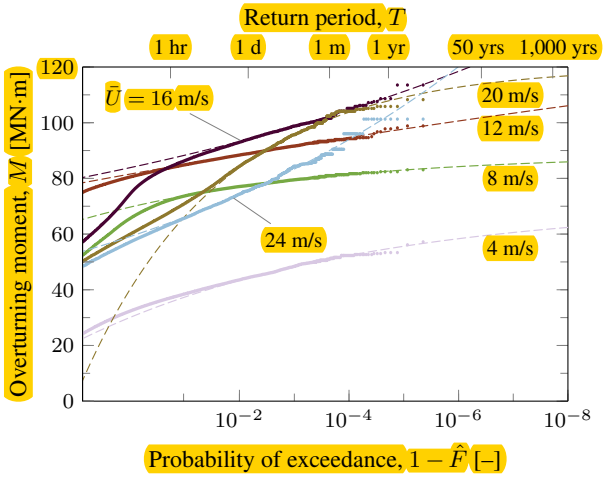

Figure 3. Return level plot of some 1-m/s wide bins with a GEV distribution fit after a threshold given by Equation (6).

This is called *fitting-before-aggregation*. It has the advantage that the data in each wind speed bin has less variation and matches closer to an underlying distribution. However, the obvious disadvantage is that a factor n fewer data points are available in each bin to fit (e.g., see Figure 3 and 4).

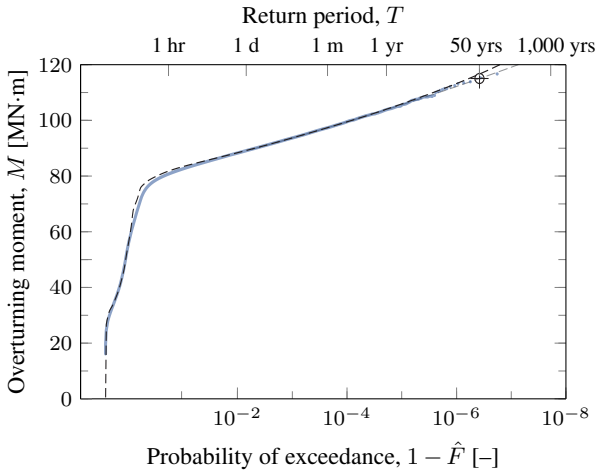

Figure 4. Return level plot of the tower base overturning moment by a weighted sum of the bins shown in Figure 3 (i.e., fitting-before-aggregation), using an equivalent 96-year sample size. The weighted distribution is given by a black dashed line, and is overlaid on Figure 2 for comparison.

2.2.2 Choice of distribution function

The tail behavior is matched with a distribution function, for example by least-squares fitting. A good candidate for this is the generalized extreme value distribution (GEV):

$$G(M;\mu,\sigma,\xi) = \exp\left[-\left(1+\xi\frac{M-\mu}{\sigma}\right)^{-1/\xi}\right], \tag{4}$$

where μ is the location parameter, σ the scale parameter, and ξ the shape parameter. A possible alternative is to fix $\xi = 0$, which produces a two-parameter Gumbel distribution:

$$G(M;\mu,\sigma) = \exp\left[-\exp\left(-\frac{M-\mu}{\sigma}\right)\right]. \tag{5}$$

A Gumbel distribution appears as a perfectly straight line on *Gumbel paper* (i.e., on a double-logarithmic scale as in Figure 2) and is often a good first guess of the tail behavior.

2.2.3 Identifying the distribution tail

The tail of the distribution shows a characteristic bend, or "knee", that hints that more than one process is at work. Indeed, tracing back the wind speeds belonging to the 10% highest loads points towards a region well above the rated wind speed (see Figure 5). It turns out that this is due to a particular controller response to negative gust amplitudes (e.g., Bos et al., 2015; Bos and Veldkamp, 2016), which also explains the shape of the scatter plot in Figure 1.

To estimate the uncertainty that comes from repeatedly extrapolating different sets of loads, this process has to be automated. However, the difficult part is then to decide where the tail exactly starts under varying sample size. A simple solution that seems

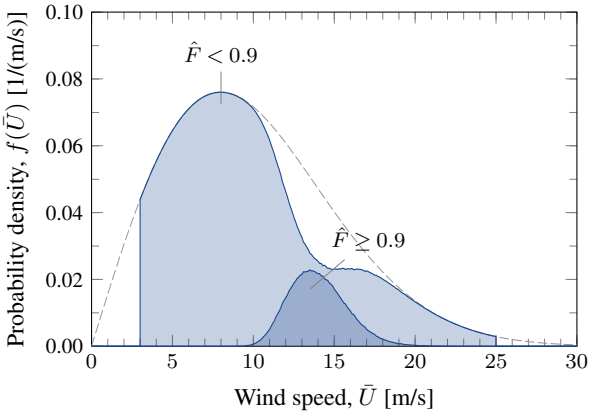

Figure 5. Histogram of the sampled wind speeds, where the dashed line marks the Rayleigh mean wind speed distribution belonging to an IEC class 1B climate. The light and dark filled areas correspond to the lowest 90% and highest 10% loads, respectively.

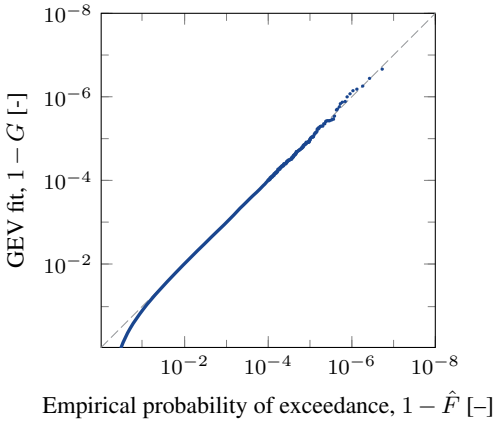

Figure 6. Q-Q plot, showing how well the empirical tail behavior matches with an generalized extreme value distribution ($\mu = 78.2$ MN·m, $\sigma = 2.05$ MN·m, $\xi = 0.026$).

to work in most cases is to assume that the tail covers the second half of the distribution when drawn on Gumbel paper; i.e., above a threshold

$$-\ln\left[-\ln\left[\hat{F}\right]\right] > -\tfrac{1}{2}\ln\left[-\ln\left[\hat{F}(M_1)\right]\right] - \tfrac{1}{2}\ln\left[-\ln\left[\hat{F}(M_N)\right]\right]. \tag{6}$$

For the full data set, this means that a distribution is fitted to the upper 0.07% of the data. For a GEV fit, this results in the Q-Q plot shown in Figure 6.

Table 1. Extrapolation cases.

	Sampling method (*parent distribution*)	Distribution fit	Fitted only above threshold
(a)	Aggregation-before-fitting (*Rayleigh*)	Gumbel	No
(b)	Aggregation-before-fitting (*Rayleigh*)	Gumbel	Yes
(c)	Aggregation-before-fitting (*Rayleigh*)	GEV	No
(d)	Aggregation-before-fitting (*Rayleigh*)	GEV	Yes
(e)	Fitting-before-aggregation (*Uniform*)	Gumbel	No
(f)	Fitting-before-aggregation (*Uniform*)	Gumbel	Yes
(g)	Fitting-before-aggregation (*Uniform*)	GEV	No
(h)	Fitting-before-aggregation (*Uniform*)	GEV	Yes

2.3 Workflow

The different approaches discussed in the previous subsection are used to set up eight extrapolation cases (see Table 1). The sampling method describes whether the data is drawn directly from its parent Rayleigh distribution and fitted to the entire set (*aggregation-before-fitting*), or if it is constructed from 1-m/s wide bins with an equal number of data points per bin (*fitting-before-aggregation*). Extrapolation is done by matching the data points with either a *Gumbel* or *GEV* distribution by least-squares fit. The final column states whether the entire sample is used (*No*) or only the points that lie above the threshold given by Equation (6) (*Yes*).

These cases are automated for $k = 10^3$ sets of loads, which yields a collection of 50-year return levels, distributed according to $f(\hat{M}_{50\text{ yrs}})$:

$$
\begin{array}{c}
\\
\leftarrow set \\
\\
\\
\\
\\
\\
\\
\\
\end{array}
\begin{array}{c}
\overset{rank \rightarrow}{} \\
\\
\\
\\
\\
\\
\\
\\
\\
\end{array}
\begin{array}{ccccccccc}
M_{1,1} & \leq & M_{2,1} & \leq & \dots & \leq & M_{N,1} & \rightarrow & \hat{M}_{50\text{ yrs},1} \\
M_{1,2} & \leq & M_{2,2} & \leq & \dots & \leq & M_{N,2} & \rightarrow & \hat{M}_{50\text{ yrs},2} \\
\vdots & & \vdots & & & & & & \vdots \\
M_{1,j} & \leq & M_{2,j} & \leq & \dots & \leq & M_{N,j} & \rightarrow & \hat{M}_{50\text{ yrs},j} \\
\vdots & & \vdots & & & & & & \vdots \\
M_{1,k} & \leq & M_{2,k} & \leq & \dots & \leq & M_{N,k} & \rightarrow & \hat{M}_{50\text{ yrs},k} \\
 & & & & & & & & \downarrow \\
 & & & & & & & & f(\hat{M}_{50\text{ yrs}})
\end{array}
$$

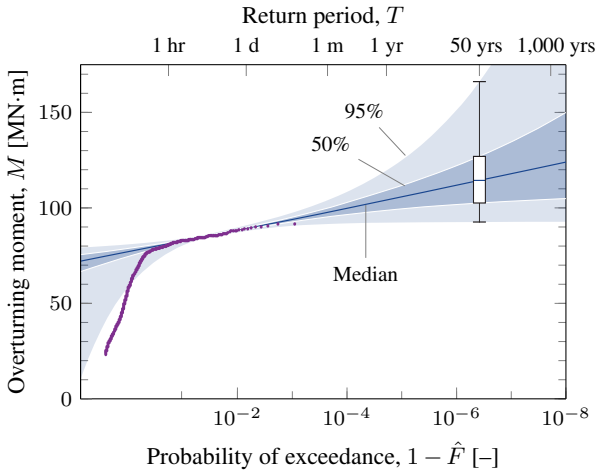

Figure 7. Return level plot of the tower base overturning moment, showing all the $k = 10^3$ GEV fits with a sample size of $N = 10^3$. The dot markers belong to one of the 10^3 samples.

The medians and other quantiles are then estimated by sorting. For example in the case #4 (the crude Monte Carlo method with a GEV fitted above a threshold), the end result is the situation depicted in Figure 7, which can be repeated for different sample sizes.

3 Results

Based on the load set and the extrapolation scheme, we can estimate how far a single 50-year load prediction would be off from the true value.

3.1 Uncertainty surrounding the 50-year overturning moment

Figure 8 shows how the median and confidence intervals around the 50-year level vary as a function of sample size, N. Evidently, the larger the sample size, the smaller the error. In addition, there are some interesting differences between the 8 approaches. Comparing the left column (a, c, e, g) to the right column (b, d, f, h), it seems that a threshold is indeed needed to establish a reliable fit of the distibution tail. Even when the data point are sampled from 1-m/s wide bins, the distributions are often bent and hardly match with any single distribution. The extreme case is when the full empirical load distribution of the aggregation-before-fitting approach (e.g., see Figure 2) is fitted to a Gumbel (a), producing errors well in excess of +100%. In the case of the GEV (c), the fit often takes on strongly negative values of ξ. This leads to a reversed Weibull distribution with an upper bound, which produces a negative bias.

The fitting-before-aggregation approach (e–h) tends to suffer from a positive bias. Likely, this is because most of the partial distributions have a slight downwardly curved tail. Such a shape requires a large enough sample size to fully establish. Small sample sizes, on the other hand, tend to result in a fit with a too large slope that overpredicts the 50-year load.

The tail of the full data set has a slight upwardly curved tail that matches best to a GEV distribution with a small positive ξ (see Figures 2 and 6). However, the Gumbel distribution is clearly more forgiving at small sample sizes. A fixed $\xi = 0$ has the advantage that the fit always stays close to the ideal value, which is especially helpful if the tail of the empirical distribution only contains a few data points.

In addition, the root-mean-squared (RMS) error provides a single measure for the quality of the result:

$$\varepsilon_{\text{RMS}} = \sqrt{\frac{1}{k}\sum_{j=1}^{k}\left(\hat{M}_{50\text{ yrs},j} - M_{50\text{ yrs}}\right)^2}, \tag{7}$$

where $M_{50\text{ yrs}}$ is the "true" 50-year level. As shown in Figure 9, the aggregation-before-fitting approach with a Gumbel fitted above a threshold (b) produces the lowest RMS error. Most of the other approaches show a clear improvement with increasing sample size. Ultimately, the RMS error of (d) falls into the classic $1/\sqrt{N}$ rule that is often found with Monte Carlo methods.

Of course, there are many other approaches to the extrapolation problem that lead to different quality results. In this paper, however, the focus is more on demonstrating how these errors affect a design process. Out of the 8 approaches presented here, 2 are selected. The first is the aggregation-before-fitting approach with a GEV fit above a threshold (d), which has a relatively small bias but a large spread. The second is the fitting-before-aggregation approach with a Gumbel fitted above a threshold (f), which has a large bias but a smaller spread.

3.2 Effect on decision making

How this uncertainty affects the decision making process is demonstrated here with a very simple example, where the choice between 2 designs is based on a stress level. Say that a new concept is proposed that is an exact copy of the original NREL 5 MW machine, but with a different wall thickness at the base of the tower. The second moment of area then changes according to

$$I_{yy} = \frac{\pi}{4}\left[r^4 - (r-t)^4\right], \tag{8}$$

where $r = 3$ m is the base radius and $t = 35$ mm is the original wall thickness (Jonkman et al., 2009). An extreme overturning moment would cause a compressive stress of

$$\sigma_z = \frac{Mr}{I_{yy}} + \frac{mg}{A}, \tag{9}$$

where $mg = 6.82$ MN is the total weight of the wind turbine and

$$A = \pi\left[r^2 - (r-t)^2\right], \tag{10}$$

is the cross-sectional area of the tower base section.

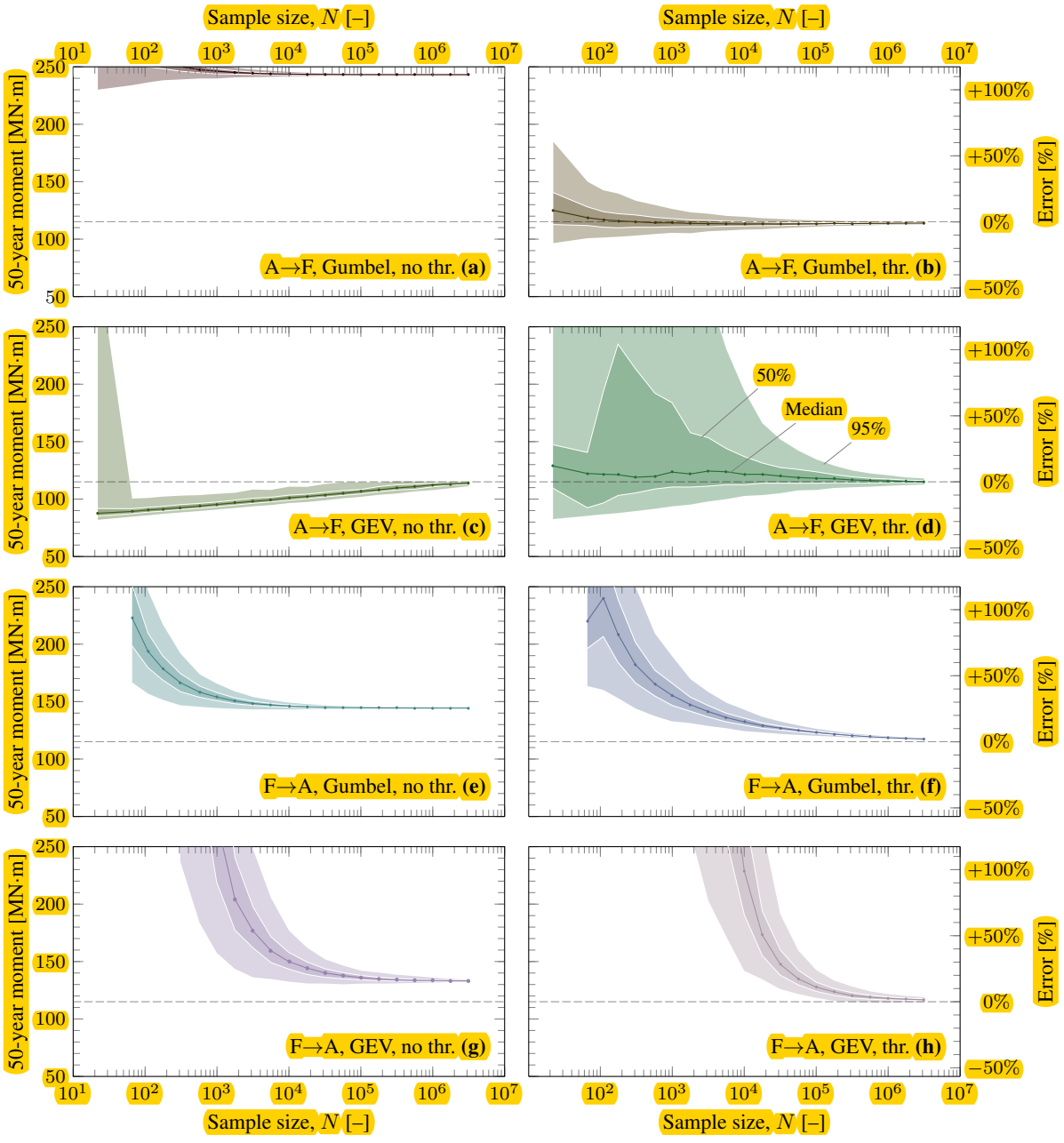

Figure 8. Error in the extrapolated 50-year overturning moment, as a function of sample size.

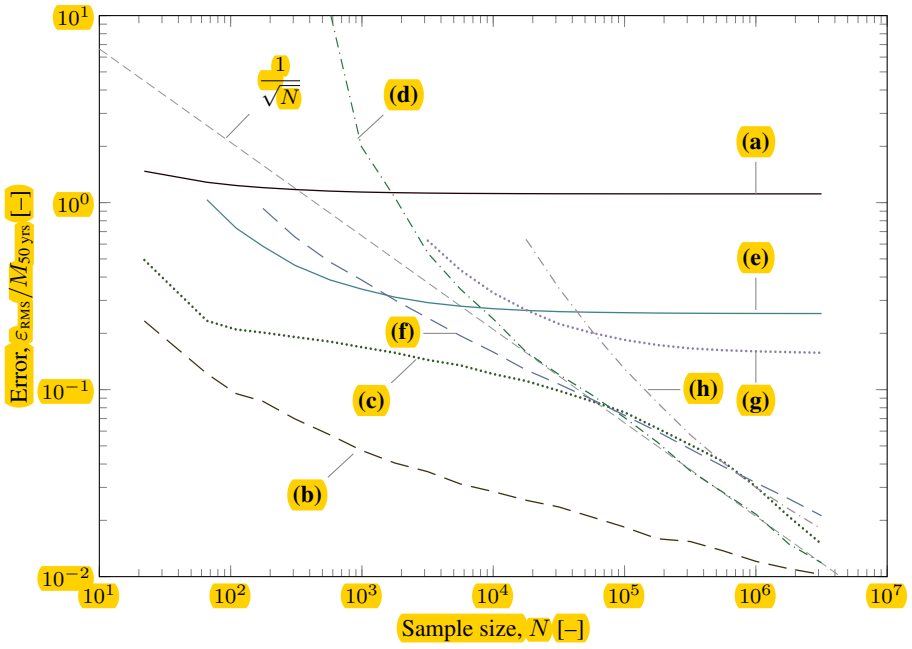

Figure 9. Root-mean-squared error in the extrapolated 50-year overturning moment, as a function of sample size.

The objective of the exercise is to find a new wall thickness to reduce the 50-year stress levels; i.e.,

$$\sigma_{z,\text{new}} < \sigma_{z,\text{old}}. \tag{11}$$

This might seem trivial at this point—any thicker wall is guaranteed to reduce the stresses—but the actual difficulty is to determine the 50-year moment. Whereas the original design has already gone through an extensive load analysis from which the 50-year load level is known, any new concept has to go through this process again.[2]

Due to the uncertainty that surrounds this 50-year level, the new design can be falsely rejected or falsely accepted. Figure 10, for example, shows how often this happens when the load analysis is carried out with the aggregation-before-fitting approach and a sample size of $N = 10^3$. When the wall thickness is reduced by 10% to 31.5 mm, the new design will appear to have lower stresses in 18% of the cases (i.e., the *false positives*). On the other hand, even when the wall thickness is increased by 10% to 38.5 mm, the new design has a 47% chance to still be rejected (i.e., the *false negatives*).

The closer a new design is to the original, the larger the required sample size (see Figure 11). In the case of a large positive bias (see Figure 12), a new concept is nearly always rejected, even if it has exactly the same thickness as the original.

Another case is a comparison between several concepts, where the 50-year stress levels contain the same degree of uncertainty. Five concepts, from 25- to 45-mm wall thickness, are ranked among each other, such that

$$\sigma_{z,1} \leq \sigma_{z,2} \leq \sigma_{z,3} \leq \sigma_{z,4} \leq \sigma_{z,5}. \tag{12}$$

[2]Maybe not for something like a new wall thickness, but more so for different control schemes or for rigorous changes to the blade design.

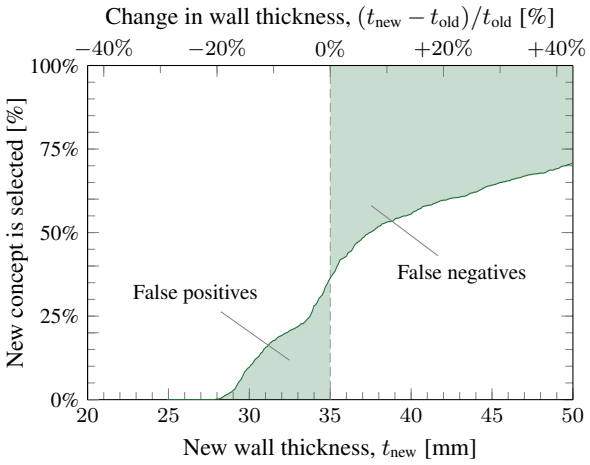

Figure 10. Outcome of a concept selection, where a new wall thickness if either accepted or rejected on the basis of a 50-year stress level using an aggregation-before-fitting approach with a GEV fitted above a threshold with a sample size of $N = 10^3$.

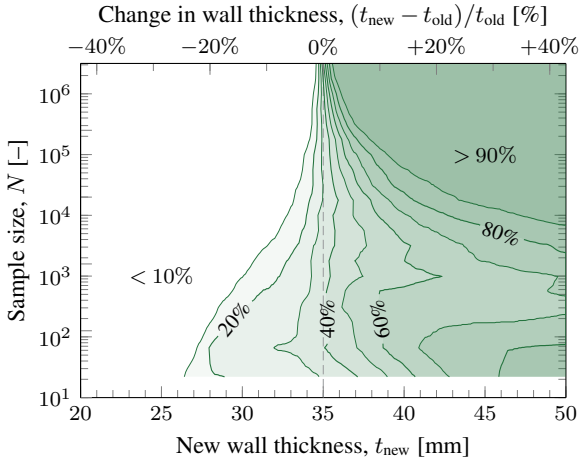

Figure 11. Relative number of times a new wall thickness is accepted on the basis of a 50-year stress level, predicted with an aggregation-before-fitting approach with a GEV fitted above a threshold.

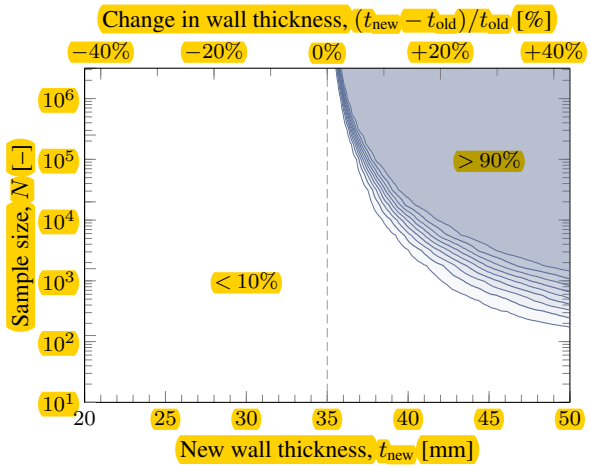

Figure 12. Relative number of times a new wall thickness is accepted on the basis of a 50-year stress level, predicted with a fitting-before-aggregation approach with a Gumbel fitted above a threshold.

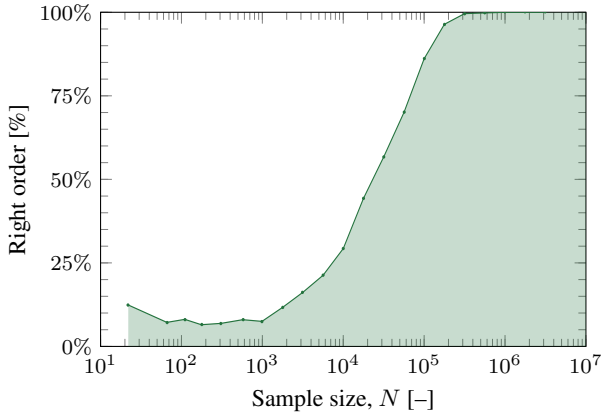

Figure 13. How often five concepts, ranging from 25- to 45-mm wall thickness, are ranked in the right order from lowest stress to highest using an aggregation-before-fitting approach with a GEV fitted above a threshold.

In the ideal case, the 45-mm wall thickness should end up at rank 1, the 40-mm one at rank 2, etc. However, how often this ideal ranking happens in practice is shown in Figures 13 and 14. This is where a small spread is preferred over a small bias. As long as the concepts are close to the same mean value, they can still be effectively compared. After $N = 10^4$, the uncertainty is small enough for the order to be right roughly 100% of the time in the case of fitting-before-aggregation (Figure 14). For aggregation-before-fitting (Figure 13), this is not until $N = 3 \cdot 10^5$.

How often each rank is assigned to each concept is shown in Figures 15 and 16. Clearly, the 45-mm wall thickness does not always appear the best and the 25-mm wall thickness does not always appear the worst. In fact, there are cases where the 45-mm wall thickness ends up being the worst of the 5 concepts (see Figure 15 for $N = 10^2$).

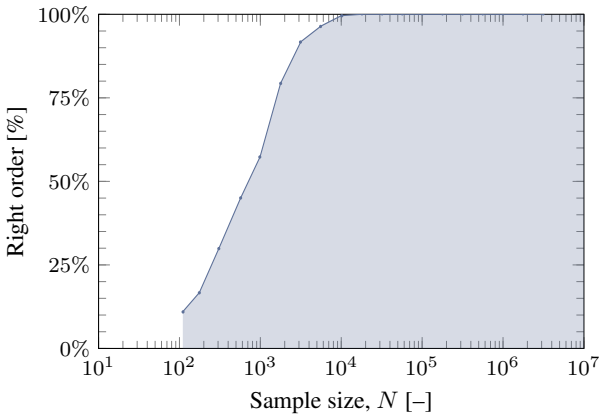

Figure 14. How often five concepts, ranging from 25- to 45-mm wall thickness, are ranked in the right order from lowest stress to highest using a fitting-before-aggregation approach with a Gumbel fitted above a threshold.

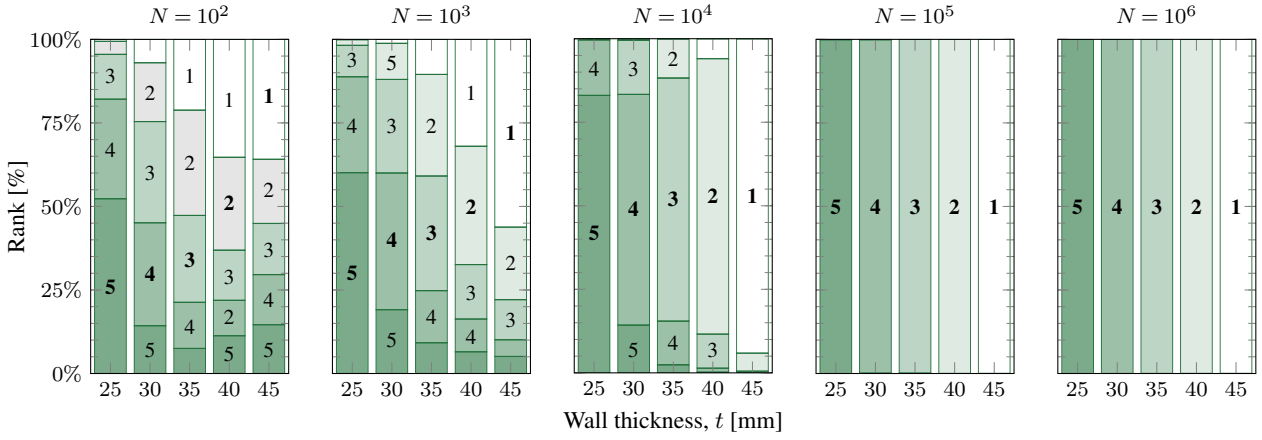

Figure 15. Ranking of five wall thicknesses on the basis of a 50-year stress level using an aggregation-before-fitting approach with a GEV fitted above a threshold with different sample sizes.

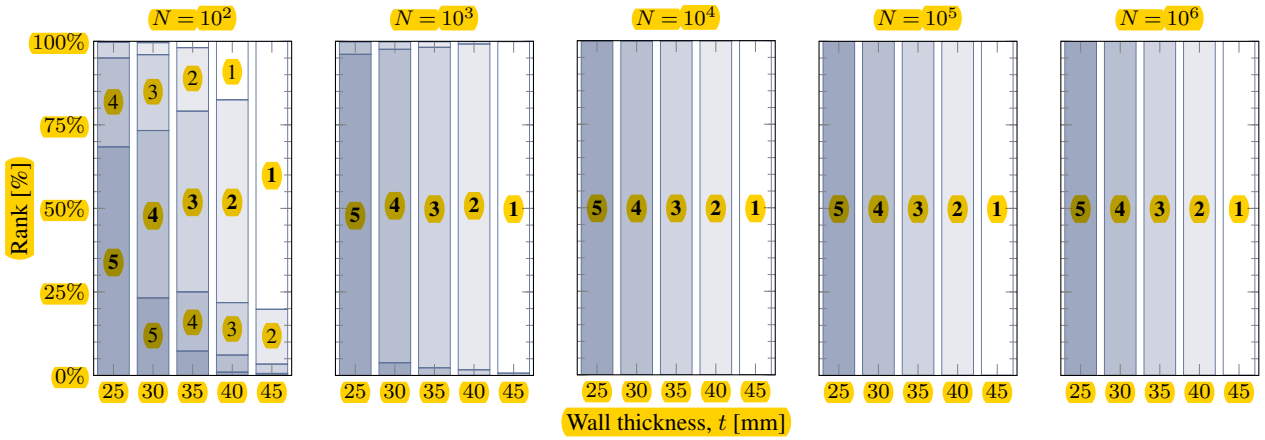

Figure 16. Ranking of five wall thicknesses on the basis of a 50-year stress level using a fitting-before-aggregation approach with a Gumbel fitted above a threshold with different sample sizes.

4 Discussion

The uncertainty around the 50-year level clearly has a very large impact on decision making. In this paper, we have focused on wall thickness in order to produce results that are counter-intuitive. This is to demonstrate that extrapolated 50-year values can be misleading and can easily trick the designer into making bad choices. In this case, the bad choices are obvious. However, they can be very difficult to spot in many other cases. For example, when choosing between different foundation types, tuning the gain settings for a controller, or even deciding whether DLC 1.1 is driving over other IEC load cases. The positive bias that is often present in extreme load extrapolations (e.g., Barone et al., 2012b; Van Eijk, 2016) makes it particularly difficult to prove that new designs are capable to reduce 50-year levels. However, it will take an immense computational effort to completely remove the uncertainty from the design process. It is therefore very important that the designer is skeptical enough of their own results.

Situations where good designs are wrongly discarded or where bad designs are wrongly accepted have a high chance of occurring when the sample sizes are small, especially during the initial design phases. In any case, we can conclude that the bare minimum of 300 minutes of time series, as prescribed in Appendix F of the standards (IEC, 2005), is not sufficient to produce any reasonable 50-year estimate (at least not when using one of the 8 approaches in this paper). One should instead aim for sample sizes larger than $N = 10^4$, and preferably larger than $N = 10^5$. The effects of changes in wall thickness that in the order of more than 10% are then easily recognizable.

The most obvious solution to reduce the uncertainty is to use high-performance computers in order to run extensive simulation campaigns (e.g., Barone et al., 2012a, b). An alternative remedy is to rely on importance sampling, which is a well-known variance reduction method that allows the user to allocate the computational resources for the most severe conditions (e.g., Bos et al., 2015; Bos and Veldkamp, 2016).

5 Conclusions

The goal of this paper was to demonstrate the effects of the uncertainty around extrapolated 50-year loads. It showed that, unless very large sample sizes are used, DLC 1.1 is a very unreliable measure for the performance of a design. This uncertainty has a pronounced effect on early phases of the design, when computational resources are often scarce.

One should always take into account that it is very time-consuming to prove that concepts are able to reduce the 50-year load, unless the design changes are very radical. In one example using an aggregation-before-fitting approach, where the bottom tower wall thickness of the NREL 5 MW reference turbine was varied, a 10% increase in wall thickness was identified as a way to reduce the stress in only 53% of the cases with a sample size of $N = 10^3$. In fact, more than 10^5 simulations were required to decrease the probability of a false rejection to 10%. Using a fitting-before-aggregation approach instead led to a strong positive bias and a rejection of the new concept in most cases.

These results show that a critical attitude is required when judging extrapolated extreme loads. When DLC 1.1 is not the design-driver, it might be best to avoid it altogether in early phases of the design. Otherwise, using high-performance computing or importance sampling methods will be the best approach.

References

Agarwal, P., Manuel, L.: Simulation of offshore wind turbine response for long-term extreme load prediction, Eng. Struct., 31(10), 2236–2246, 2009, doi:10.1016/j.engstruct.2009.04.002.

Barone, M., Paquette, J., Resor, B., and Manuel, L.: Decades of wind turbine load simulation, 50th AIAA Aerosp. Sci. Meet. Incl. New Horizons Forum Aerosp. Expo., Nashville, TN, United States, 9–12 January, 2012, doi:10.2514/6.2012-1288.

Barone, M., Paquette, J., Resor, B., Manuel, L., and Nguyen, H.: Simulating the entire life of an offshore wind turbine, EWEA Annual Event, Copenhagen, Denmark, 16–19 April, 2012.

Bos, R., Bierbooms, W. A. A. M. , van Bussel, G. J. W.: Importance sampling of severe wind gusts, 11th PhD Semin. Wind Energy Eur., Stuttgart, Germany, 23–25 September, 2015.

Bos, R., Veldkamp, H. F.: A method to find the 50-year extreme load during production, J. Phys. Conf. Ser., 753, 42021, 2016, doi:10.1088/1742-6596/753/4/042021.

van Eijk, S. F.: Wind turbine load extrapolation, Master's thesis, Delft University of Technology, 2016.

Fogle, J., Agarwal, P., Manuel, L.: Towards an improved understanding of statistical extrapolation for wind turbine extreme loads, Wind Energy, 11, 613–635, 2008, doi:10.1002/we.303.

Freudenreich, K., Argyriadis, K.: The load level of modern wind turbines according to IEC 61400-1, J. Phys. Conf. Ser., 75, 12075, 2007, doi:10.1088/1742-6596/75/1/012075.

Genz, R., Nielsen, K. B., Madsen, P. H.: An investigation of load extrapolation according to IEC 61400-1 Ed. 3, Eur. Wind Energy Conf., Athens, Greece, 27 February–2 March, 2006.

IEC: 61400-1 Wind turbines – Part 1: Design requirements, 3rd edn., International Electrotechnical Commission, Geneva, Switzerland, 2005.

Jonkman, J. M., Butterfield, S., Musial, W., and Scott, G.: Definition of a 5-MW reference wind turbine for offshore system development, Technical Report NREL/TP-500-38060, National Renewable Energy Laboratory, Golden, CO, United States.

Lott, S., Cheng, P.-W.: Load extrapolations based on measurements from an offshore wind turbine at alpha ventus, J. Phys. Conf. Ser., 753, 072004, 2016, doi:10.1088/1742-6596/753/7/072004.

Moriarty, P. J., Holley, W. E., Butterfield, S. P.: Extrapolation of extreme and fatigue loads using probabilistic methods, Tech. rep. NREL/TP-500-34421, National Renewable Energy Laboratory, Golden, CO, United States, 2004.

Natarajan, A., Holley, W. E.: Statistical extreme load extrapolation with quadratic distortions for wind turbines, J. Sol. Energy Eng., 130(3), 031017, 2008, doi:10.1115/1.2931513.

Pandey, M. D., Sutherland, H.J.: Probabilistic analysis of LIST data for the estimation of extreme design loads for wind turbine components, J. Sol. Energy. Eng., 125(4) 2003;125:531. doi:10.1115/1.1626128.

Peeringa, J.: Comparison of extreme load extrapolations using measured and calculated loads of a MW wind turbine, Eur. Wind Energy Conf., Marseille, France, 16–19 March, 2009.

Ragan, P., Manuel, L.: Statistical extrapolation methods for estimating wind turbine extreme loads, 45th Aerosp. Sci. Meet. Exhib., Reno, NV, United States, 8–11 January, 2007.