Reply to referee 2

We thank the reviewer for reading our manuscript and giving thoughtful comments and suggestions. A detailed response to all comments is found below in blue.

The authors present a well crafted calibration exercise for storm-damage functions. They present a clear decision tree for the chosen methods and assumption. They are using a storm-based approach to statistically fit historical losses to wind speed information using different models. The basis for the calibration are high resolution insurance loss data and wind speed data covering a relatively long time period. The authors present and discuss the results in detail focusing on the high impact events and on creating a damage classifier. In the end, the authors provide a short broader discussion.

General Comments:

Comment 1

The insurance loss data and the modelled damages are obviously very skewed and mostly presented using logarithmic axis to focus on relative differenced / differences in the order of magnitude. But in the methodology, this is not incorporated as such. I would suggest the authors to change or at least expand their methodology at two points:

Thanks for the suggestions. We agree with the reviewer's point that the damages both observed and their estimates are skewed and it's not tangible to visualise it in their absolute values.

- Section 2.4.5 Ensemble mean method: As another option instead of using the arithmetic mean, I would suggest to use the mean of all the logs, as in:
  meanlog = 10 ^ (1/n * sum_i_n[log(xi)]) for a n loss estimates x.

  Since there are zero losses present in the loss estimates, an ensemble of the estimates with log transformation is not possible.

- Section 2.6 Model evaluation metric: I suggest to also calculate a metric that takes into account the very different order of magnitude. One option would be the mean absolute percentage error.

  We agree with the need of using different accuracy metrics. The mean absolute percentage error (MAPE) is indeed a dimensionless prediction accuracy metric.

However, the presence of zeros in loss values restricts us from using MAPE as the error metric. For the same reason, we chose to calculate the coefficient of variance, which is also a dimensionless accuracy metric that gives the dispersion of prediction around the mean.

Comment 2

In figure 1 the damage functions for only one municipality are shown. It is expected and written that there is a the variability of the calibrated damage functions over all municipalities, but it is not shown. It would be nice to either report the range of the calibrated parameters in a table in the supplementary material or even better reproduce a figure similar to Figure 1 showing not the points of the insurance losses but only all the calibrated damage funtions in one plot. This would provide a much better idea of the variability between the different municipalities.

We thank the reviewer for the suggestion. We have performed the proposed figure (see Fig. R1 below) and agree that it well illustrates the variability in the fits of the damage functions among the municipalities.

The fit of four damage functions analogous to Fig. 1 is shown in Fig. R1 but for all 356 municipalities. From this it is clear the fit of damage functions varies not only between models but also spatially. Figure R1a illustrates the variety of fits for the exponential damage function with very steep lines for some municipalities and much flatter lines for others. Figure R1b displays the fit for all the municipalities and highlights that the Klawa damage function doesn't increase as steeply as the other models. For the Prahl model, Fig. R1c exhibits a large variability in the fits from municipality to municipality. Figure R1e also shows that the sigmoids depicting the probability of damage occurrence have different shapes with some curves not reaching a probability of 1 within the wind speed range represented. Note that the fit can lead to negative probabilities, that we set to 0 afterwards. The shape of the magnitude term (see Eq. 8) in the Modified Prahl model can be very different among municipalities as shown in Fig. R1d with very steep or weak slopes. Therefore, we will include this figure in the supplement and some sentences on these results in the methods section.

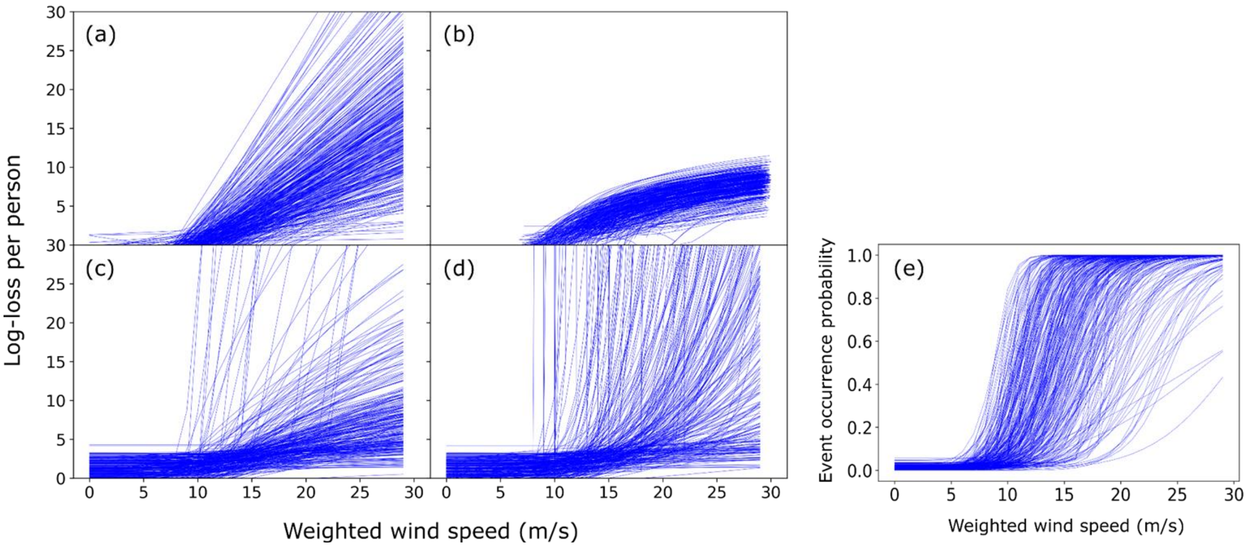

Figure R1: Shapes of the damage functions for all municipalities for (a) the exponential damage function, (b) the cubic excess over threshold damage function, (c) the magnitude term in the probabilistic damage function by Prahl, and (d) the magnitude term in the modified Prahl probabilistic damage function, (e) sigmoid function that estimates the probability of an event occurrence. Note that the y-axis for (a)-(d) represents the log-loss per person with units of log NOK.

Comment 3

The data is split into testing and training set, which is a very important practice. I suggest to use a cross-validation approach, especially as the results in Table 1 are reported only on the unseen testing data, and municipalities have to be excluded from the evaluation due to lack of data.

We agree with the reviewer on the importance of cross-validation. Therefore, we have decided to perform a seven-fold cross-validation. The 36-year loss data is split into 7 groups in chronological order with each group containing five years of loss data (1985-1989, 1990-1994, 1995-1999, 2000-2004, 2005-2009, 2010-2014, 2015-2019) and the loss data of year 2020 is not included in any of the groups. Now taking each group as testing data, the damage functions are trained on the rest of the data. The predictive skills of the damage functions are evaluated on the testing data. The large spread in the model skill metrics (i.e., MAE and CV) indicates that the performance of damage functions is highly dependent on the choice of the training data (Fig. R2). For each model, the spread in the number of municipalities showing the smallest MAE (such as done in Table 1) remains relatively low across all loss classes, as defined in section 3.1 of the manuscript (Fig. R2 c, f, i). The black dots in Fig. R2 show the results present in the manuscript (Table 1) obtained with another set of training and testing data. We notice that they often lie outside the interquartile range, especially when considering all loss days and the extreme loss days (top and bottom rows in Fig. R2), and are sometimes even outside the range of the seven-fold cross-validation analysis, emphasising again the strong dependence of the results to the chosen training and testing periods. In light of this new analysis, we will include these results in the manuscript and figure R2 in the supplement.

We will add some sentences in the manuscript on this topic, such as:

In section 2.6 (model evaluation section):
The damage functions are sensitive to extreme loss observations and the presence of few extreme events can heavily alter the damage functions' shape. Therefore, different training data sets may result in differing damage function fits. Cross-validation is an effective method to estimate the uncertainties involved in the choice of the testing and training data. We perform a seven-fold cross-validation by splitting data into seven with each set of testing data having five consecutive years of data. So, in the first fold the testing period is 1985-1989 and the training period is 1990-2020, in the second fold the testing period is 1990-1994 and the remaining years are in the training dataset, and so on.

In section 3.2:
The seven-fold cross-validation reveals that the parameters in the storm-damage functions obtained during the fitting step depend on the choice of the training dataset. Moreover, the loss estimates are also highly dependent on the choice of training dataset (see the range in the model evaluation metrics shown in Fig. R2 a,b,d,e,g,h). However, whatever the training dataset, the Klawa and exponential models still have the best skill in most of the municipalities (as also shown in Table 1) across the different loss classes (see Fig. R2 c,f,i).

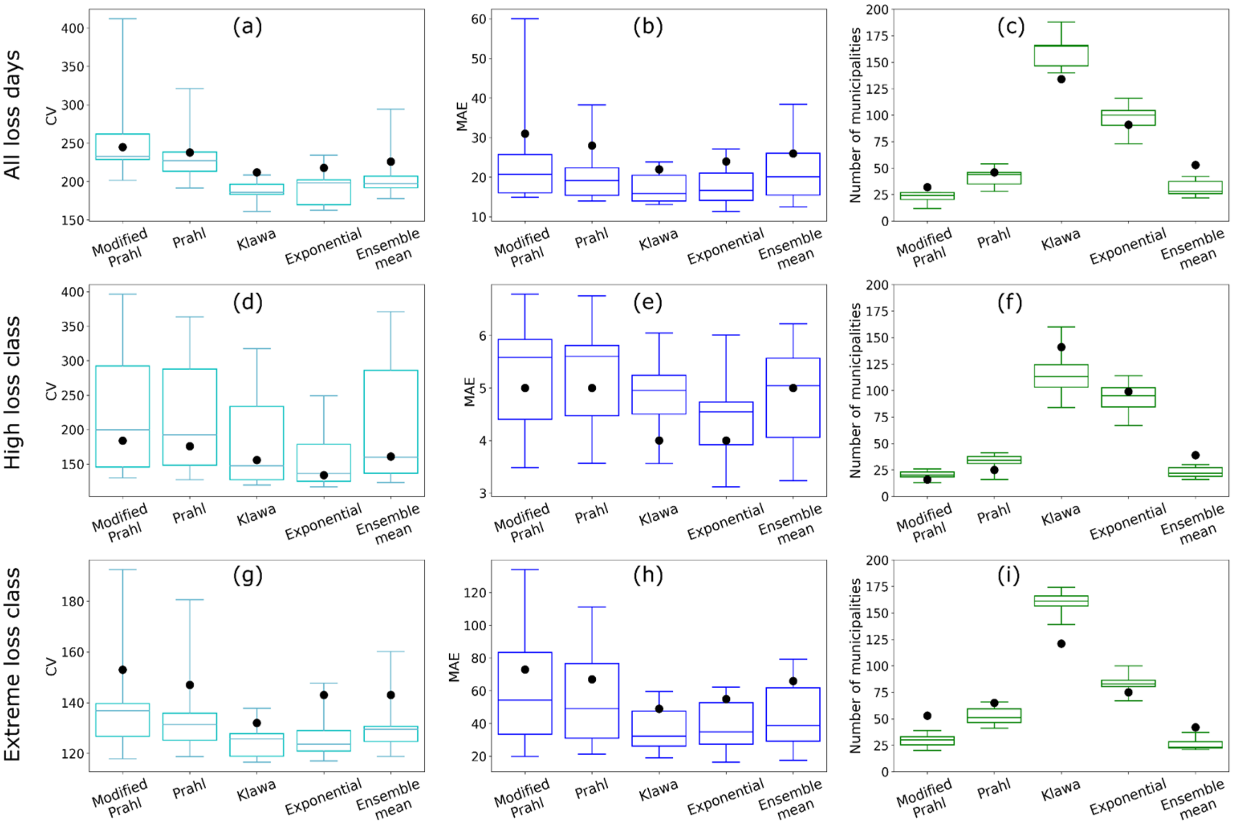

Figure R2: Distribution of model performance metrics from cross-validation (a) coefficient of variance (CV), (b) mean absolute error (MAE), (c) number of municipalities with smallest MAE for four damage functions and their ensemble mean for all loss days. (d), (e) and (f) same as (a), (b) and (c) but for the high loss class. (g), (h) and (i) same as (a), (b) and (c) but for the extreme loss class. The boxes represent the interquartile range, the horizontal line represents the median, the whiskers represent the minimum and maximum and the black dots represent the results from Table 1 in the manuscript.

Comment 4

A short broader discussion is done in the section 4. “Conclusion”. I would suggest to also discuss the following aspects:

- Discuss windspeed as explanatory value for damage model, is it able to represent the randomness of gust occurance? This is relevant for “low intensity” events, where damages are caused by infrequent and hard to predict stronger gusts. This is relevant for both damage classifier as well as estimating low intensity impacts)
- The population distribution is not changing in the chosen model setup only the total number of people. Could it be that “where people live” did not only change in scale (represented in the model) but also in location (not represented). If yes, how would this influence the model performance?
- What is the purpose of this calibration exercise, is there a foreseen application? If yes, it would be nice for the reader to know, especially what function would be the chosen

for this specific application? Even better would be a general discussion of the metrics: Which function would be the best for which type of application?

- Other readers might be needing to do a similar calibration exercise, but might lack the sample size used in this study or might face other shortcomings. It would be very interesting to discuss some learnings that could be generalized for other calibration exercises.

Thank you for the thoughtful suggestions. Here below, we discuss the four points raised by the reviewer. We will extend our discussion in the revised manuscript.

Wind speed is the most common variable used to estimate storm damages. A drawback of this approach is that the same wind speed at the municipality level resolution may cause small damages in some cases or no damages in most cases. Such inconsistencies occur mainly due to extremely local high wind gusts and incorrect reporting of damages. As a consequence, the lower end of the wind speed damage relations becomes noisy, thus making it very difficult to model. To check how the wind gust from NORA3 compares to the wind speed, we performed the same population weighting exercise with wind gusts and found a high correlation in the 98th percentiles calculated from wind speeds and wind gusts (Fig. S4b). With insurance data being at a coarser resolution than the wind gusts, which are very local (a few hundred metres), hard to predict and very transient (less than a minute), it is difficult to use wind gusts to fit damage functions.

Due to the unavailability of the gridded population data for the earlier part (1985-1999) of our study period, we had to use constant population to weight the wind speed at every grid point. Therefore, we cannot take into account the spatial change in population density, such as the spatial expansion of cities with time. This is a source of uncertainty in our storm-damage fits.

Foreseen applications can be the damage assessment right after an event, assessment of future losses in the context of climate change, and for impact-based forecasting of damages. For the future assessment, the Klawa damage function could be an obvious choice because it has an adaptation component by updating the 98th percentile with the future wind speed and it is the most common model used in the literature, which makes it easier for comparison. For the damage assessment shortly after an event and for the future assessment of losses, a set of different damage functions can be used in order to get a measure of the uncertainty in the monetary loss amount. For forecasting purposes, an ideal starting point would be to apply a damage classifier to distinguish between damaging and non-damaging winds, as part of an early warning system, followed by a prediction of losses using a variety of damage functions.

High quality data on loss and wind speed is necessary for the calibration of damage functions. A long time series of loss data is desired to reduce uncertainties and increase accuracy of model fitting and predictions. However, loss information as used in this study is rarely available, unfortunately. In such cases, a general approach is to approximate the losses using population of the respective regions and then quantify the impact of windstorms (Donat et al., 2011). In addition, there are open source climate risk assessment models such as CLIMADA (Aznar-Siguan and Bresch 2011), which can be coupled with the loss data in hand for damage estimation.

Aznar-Siguan, G. and Bresch, D. N. (2011) CLIMADA v1: a global weather and climate risk assessment platform, Geoscientific Model Development, 12, 3085-3097, https://doi.org/10.5194/gmd-12-3085-2019

Donat, M. G., Leckebusch, G. C., Wild, S., and Ulbrich, U.: Future changes in European winter storm losses and extreme wind speeds inferred from GCM and RCM multi-model simulations, Nat. Hazards Earth Syst. Sci., 11, 1351–1370, doi:10.5194/nhess-11-1351-2011, 2011

Specific comments:

Comment 1

Title: is the phrase "in the complex terrain" justified? The only terrain specific methodology used in this paper is the usage of a population density that follows topographic features. Would "taking into account heterogenic population density" be more describing?

We agree with the reviewer that our methodology does not directly involve the Norwegian topography, which is only implicitly taken into account in the population data, with generally more people living along the sea and in the valleys than over the mountains. Therefore, following the reviewer's suggestion, we will change the manuscript title in order to better reflect our methodology, to *Assessment of wind-damage relations for Norway using 36 years of daily insurance data.*

Comment 2

L29: quick impact estimation for response planning directly after the event is also an important application (see Welker et al. 2021)

Thanks for mentioning this relevant point and bringing this paper to our attention. We will change the sentence to something like:

Establishing robust windstorm-damage relations may help predict storm damage risk in the weather forecasting context (Merz et al., 2020), roughly estimate the storm impact directly after it occurred in order to better plan the emergency response (Welker et al. 2021), and evaluate the change in risk on the longer term in conjunction with climate change.

Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D. N., Domeisen, D. I., Feser, F., Koszalka, I., Kreibich, H., et al.: Impact forecasting to support emergency management of natural hazards, Reviews of Geophysics, 58, e2020RG000 704, 475 https://doi.org/10.1029/2020RG000704 2020.

Welker, C., Röösli, T., and Bresch, D. N.: Comparing an insurer's perspective on building damages with modelled damages from pan-European winter windstorm event sets: a case study from Zurich, Switzerland, Nat. Hazards Earth Syst. Sci., 21, 279–299, https://doi.org/10.5194/nhess-21-279-2021, 2021.

Comment 3

Figure 1 a) mark the threshold in the plot (compare with L170:"…,we chose the 95th percentile of the wind speed […] as the threshold." If a modelled damage of zero would be assumed for events with a wind speed below the threshold, it would be nice to show a doted line at zero similar to Figure 1 b)

We will modify the figure as suggested by the reviewer. Although we only use the wind speed bins above the 95th percentile of the wind speed to calculate the fit, the obtained exponential model can also be applied to the wind speeds below the 95th percentile and we can get loss estimates for those wind speeds as well, as shown in Fig. 1a (red curve to the left of the 95th percentile line). However, if the estimated loss is negative, we set it to 0.

Comment 4

L172: "…are split into ten equally spaced bins…" I can only see 6 bins in Figure 1 a). I assume it is because 4 bins did not include the minimum of at least 5 loss days. If that is the case, I would be nice to state that for the reader. If this is not the case it would be even more important to state this.

The reviewer is right, it is because 4 bins did not include the minimum of at least 5 loss days that Fig. 1a only displays 6 bins out of the 10. We will add a sentence in line 173 to make this aspect clearer, such as: *Note that Fig. 1a only displays 6 bins because the 4 other bins do not include the minimum of 5 loss days required in each bin.*

Comment 5

L307ff: “The extreme loss class represents 31 and 9 ….”. The structure of this sentence makes it hard to understand. Please consider a revision.

We agree and we will rephrase this sentence and better connect it with the previous sentences.

Comment 6

Figure 3 b) and c). Please provide a reasoning for the skewed percentiles of the non-linear class boundaries. Why are there no 5th and 10th percentile?

Figure 3a shows that most of the municipalities have a MAE between 0 and 100, whereas only a few have higher MAE. Therefore, the distribution of the MAE is highly skewed towards low MAEs and the lowest percentiles will be very similar. It is already visible in Fig. 3b, for example, with the difference between the 40th and 20th percentiles being of only 13 (28-15) in contrast to the difference between the 95th and 90th percentiles being 139 (406-267) so more than 10 times larger. Therefore, we chose to highlight the larger percentiles rather than the lower percentiles. However, we have chosen to follow the reviewer’s suggestion and the updated figure is shown below:

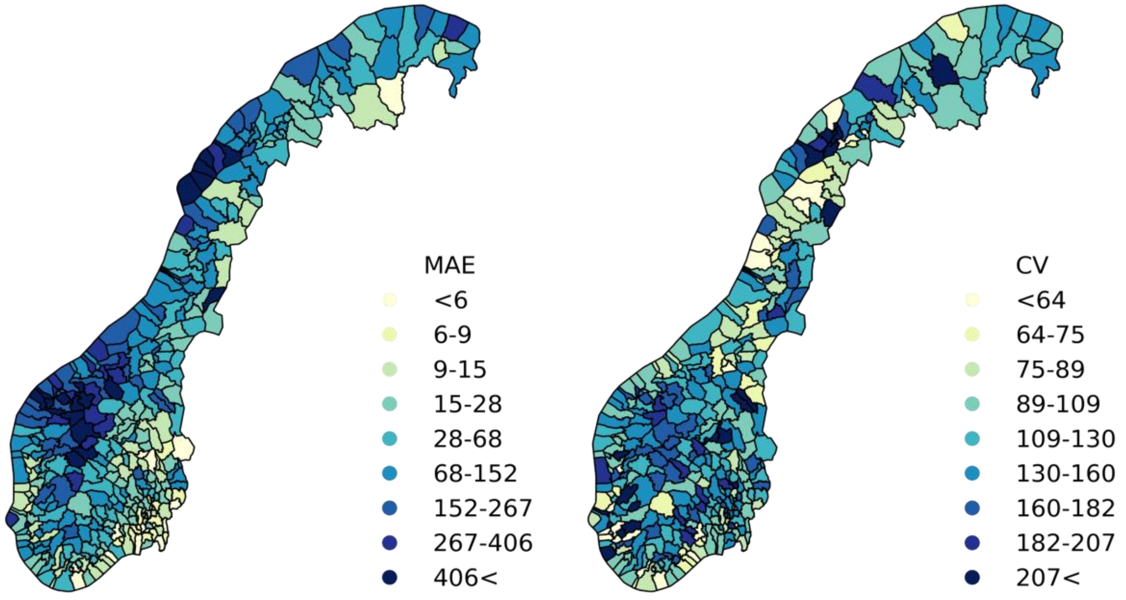

Figure R3: (Left) Map of the smallest MAE among the five models in the extreme loss class fitted on the test data and (right) the corresponding coefficient of variation of the root mean square error. The legends have non-linear class boundaries at the 5th, 10th, 20th, 40th, 60th, 80th, 90th and 95th percentiles. Note that the results are based on the performances on the unseen testing data.