AUTHORS' RESPONSE TO REFEREE #2

Research article: "Regional tropical cyclone impact functions for globally consistent risk assessments" (Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2020-229; in review, submitted on 09 July 2020)
Authors: Samuel Eberenz, Samuel Lüthi, David N. Bresch

We thank Andrew Gettelman for his thorough review and valuable comments. The original comments from the referee are listed below directly followed by our responses in *blue and italic* and suggested changes to the manuscript in **blue and bold**.

Review of Regional tropical cyclone impact functions for globally consistent risk as- sessments by Eberenz et al.
This manuscript dives into loss functions used in a Tropical Cyclone (TC) damage model and tries to adjust damage functions by region to better match the observed record of damages. The paper is generally well written and should be published in Natural Hazards and Earth System Sciences with minor revisions.

2.1) I have some specific comments below, but I would like to see a bit more explanation for some of the figures and analysis. Especially, some of the appendix (and the two figures) could be folded into the main text.

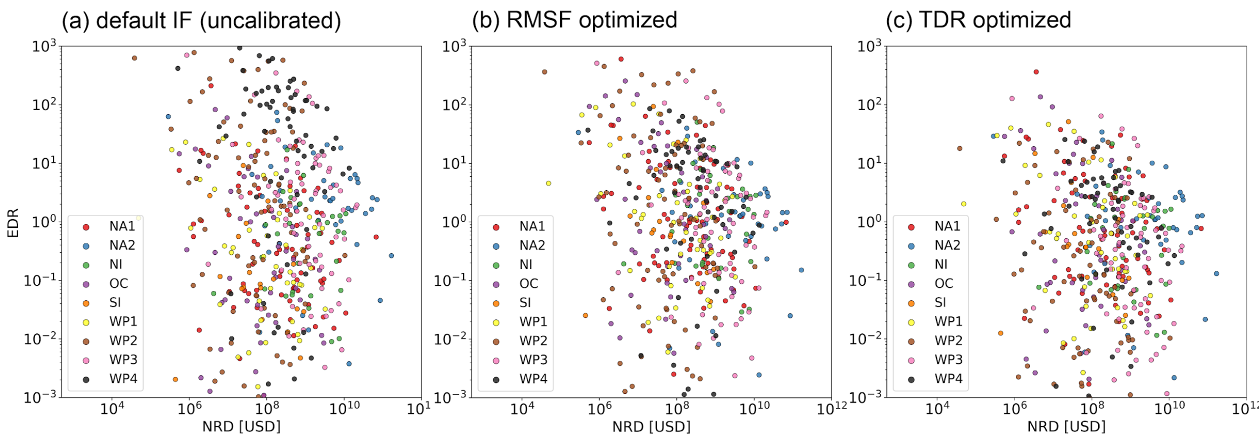
Thank you for the suggestions regarding the figures and explanations in the manuscript. They are most welcome and we will reply more in-depth and suggest changes together in our answers to the specific comments below.
Most notably, instead of folding figures A1 and A2 into the main text, we suggest a new figure (EDR boxplot per region, c.f. Comment 2.13) to be added to the results section, c.f. responses to comments 2.13, 2.17, 2.19, and 2.20.

2.2) Also, it's not clear whether the trend for simulated damages is an over or under estimation of damages before calibration and whether this is due to strong or weak storms. Maybe this is in the figures, and could be mentioned in the discussion/conclusions.

Whether damages are on average over- or underestimated is shown explicitly in the manuscript: For the average per region, Figure 5 and Table A2a provide this information as conveyed by uncalibrated total damage ratio (TDR), i.e., the grey bars in Figure 5b, next to the values after calibration. These numbers are also discussed in the text (L. 288ff, Section 3.1.1). For the single countries the spread of uncalibrated event damage ratios (EDR) information is displayed in Figure S2 in the supplementary materials. We agree that EDR per region could be shown more explicitly in the main text. The regional findings with regards to over- and underestimation before calibration form the basis for the whole calibration and are reflected both in the results and discussion section.

As for the question whether strong or weak storms are driving the average over- or underestimation (i.e. as measured by TDR), we agree to the referee that this is not yet discussed broadly in the manuscript. Our results show no significant correlation between normalized reported damage (NRD, here taken as a measure of TC severity) and EDR (measure of over-/underestimation of damage in CLIMADA). In response to this comment, we also plotted scatter plots of EDR vs NDR, finding no evidence of any significant relationship between TC severity and over- or underestimation.

We will add these scatter plots to the appendix of the manuscript as new **Figures A3***:*



"Figure A3: No significant correlation between event damage ratio (EDR) and normalized reported damage (NRD) was found. The scatter plots show the relationship for 473 TC events worldwide computed with three different sets of impact functions: (a) uncalibrated default (V_{half}=74.7 ms-1), (b) RMSF optimized, and (c) TDR optimized. The nine calibration regions are differentiated by colour."

Suggested addition to Section 3.2.2.:

*L. 298: "*The EDR values within regions show a large spread over several orders of magnitudes (Fig A1). **There is no significant correlation between EDR and NRD (Fig. A3), suggesting that the over- and underestimation of simulated event damages is not related to TC severity.** The largest spread, as expressed by the RMSF [...]*"*

To clarify that Figure 5b shows over/underestimation of average damages per region, we will explain this better in the figure caption of Figure 5:

L. 337ff (caption Figure 5):

"Figure 5: Calibration results and cost functions for nine calibration regions and all regions combined**, each shown before (grey) and after calibration (blue and red)**: (a) Vhalf: fitted impact function parameter; (b) TDR: ratio of total simulated and normalized reported damage; (c) RMSF: root-mean-squared fraction; and (d) AAD: **normalized reported (green) and simulated** annual expected damage **(AAD)**. [...]"

Furthermore, in our response to comment 2.13, we suggest to show uncalibrated EDR per region in a new figure in the beginning of the results section.

2.3) Also, the analysis focuses on tuning v-half. What would happen if you either used or added v-thresh (the minimum wind speed for damages) as an adjustment parameter? Would that help? Why or why not? Can you test it?

This comment touches upon one of the most pivotal decisions during calibration: the choice of free parameters in the impact function. Thank you for pointing out that more justification for the decision to only vary V_{half} should be provided. The very short reason is that we concluded that fitting more than one of the linear dependent parameters (c.f. L. 154f) increases the risk of overfitting. Here is the long justification.

The approach of this paper builds on the Master Thesis by one of the authors with the title "Applying Machine Learning Methods to the Assessment of Tropical Cyclone Impacts" (Samuel Lüthi, 2019, available at https://doi.org/10.3929/ethz-b-000398592).
In the Master Thesis, regional TC impact functions were calibrated with Bayesian optimization methods based on almost the same data set of TC tracks and EM-DAT entries per event and country. Also, the same wind field model and impact engine as implemented in CLIMADA was applied. Differences between the Lüthi (2019) and the present paper under discussion mainly lay in the definition of the regions but also further refinements and quality control of the underlying data added for the present paper: EMDAT events which produce no damage in CLIMADA had not been excluded from the analysis in Lüthi (2019). This can cause artefacts, e.g. if a Carribean TC undergoes extratropical transition and then makes "landfall" a second time as a low pressure system.

Lüthi compared regional calibration results of two multi-parameter impact functions: (i) a sigmoid function similar to the function by Emanuel (2012) with three free parameters (slope, offset, and maximum intensity) and (ii) a 12-step staircase function. Figures 3.4 and C.1 of Lüthi (2019) show the resulting impact functions and RMSF scores:

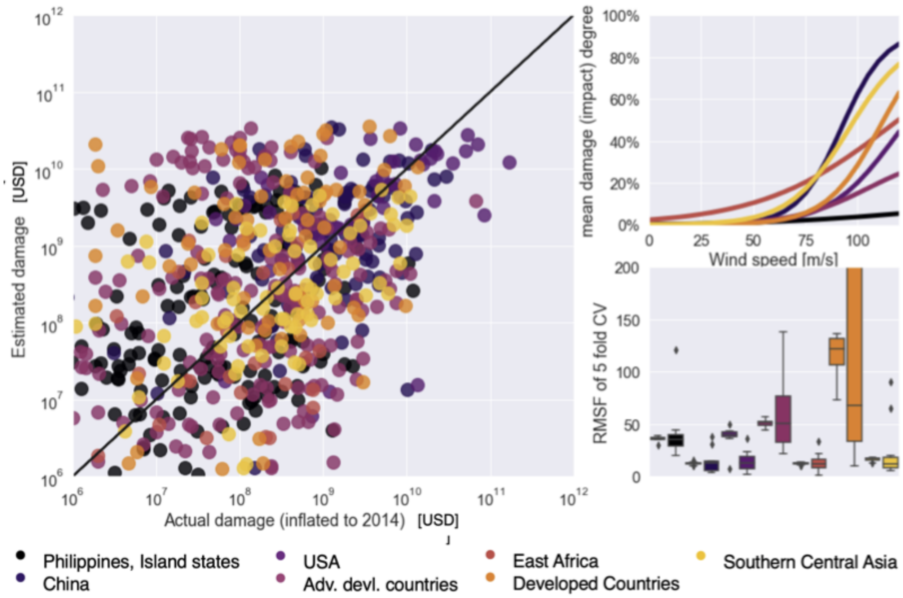


FIGURE C.1: Estimated vs. actual damages for all storms of all regions (left, log-scale) using sigmoid damage functions, including regional damage functions (top-right), and regional training (left box per region) and validation (right box per region) RMSF error (bottom-right). Damages are calculated using the regional damage functions which result from the calibration using all data per region. Training and validation RMSF are the output of a 10-fold cross validation per region. This plot corresponds to Figure 3.4.

105

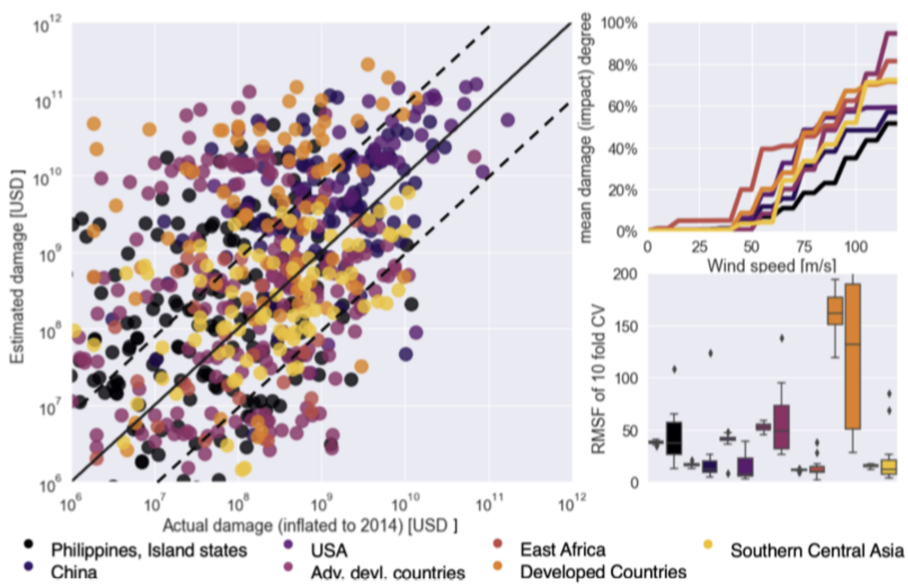


FIGURE 3.4: Estimated vs. actual damages for all storms of all regions (left, log-scale), including regional damage functions (top-right), and regional training (left box per region) and validation (right box per region) RMSF error (bottom-right). The solid line indicates correctly estimated storms, storms in between the two dotted lines are estimated in the right order of magnitude. Damages are calculated using the regional damage functions which result from the calibration using all data per region. Training and validation RMSF are the output of a 10-fold cross validation per region.

Based on his results as partially shown in the two figures displayed above, Lüthi (2019) concludes: "The comparison of different calibration approaches (Section 3.3) reveals that sigmoid functions are an attractive tool. As these functions depend solely on three parameters, they can be calibrated comparatively fast and at lower computational cost. However, they can produce counter-intuitive results. As an example, the resulting damage function for East Africa (Figure C.1, appendix) shows mean damage degree values larger than zero at zero wind. On the other hand, the multi-step function shows a similar performance using twelve parameters." (p. 25) and "While the damage functions look quite different, the estimated damages and also the training and validation errors are similar. This is due to the fact that the damage functions are quite similar in the region of 30-50 m/s." (p. 41)

For the parameterization used in the present study, this implies that a shift of V_{tresh} *alone would require relatively large shifts to unrealistically low or even negative wind speeds (Indian Ocean) or very large intensities (North West Pacific region). Changing multiple parameters is problematic because the parameters* $\boldsymbol{V_{thresh}}$, $\boldsymbol{V_{half}}$, *and* ***scale*** *are not linear independent. This introduces a risk of overfitting considering the large uncertainties in the damage data underlying the calibration; Fitting a more flexible, multi-step function with larger resolution from 30-50m/s could be a valid alternative to the sigmoidal function. However, as Lüthi (2019) showed, this more computationally expensive approach does not result in any considerable improvement in skill. Therefore, we decided to keep the fitting as simple and transparent as possible as long as the other uncertainties are not reduced.*

In light of the findings by Lüthi (2019), our approach comes with the following advantages: (i) thanks to the fixed value of V_{thresh}, *it only produces physically plausible curves, (ii) allowing only 1 free parameter yields a relatively fast (computationally cheap) calibration and (iii) results are easy to interpret and compare. Comparing the RMSF values after calibration of the two studies shows that the results are generally similar; often RMSF numbers retrieved for this study are lower. In the case of China (WP3), the only region which has the same underlying countries in both studies, the calibration in this study produces lower RMSF values. It should be noted that RMSF in Lüthi (2019) are negatively affected by events which produce no damage in CLIMADA (as mentioned above, these events have been excluded in the analysis for this paper).*

In response to this comment and also comment 2.11 below, we suggest to add a brief discussions of the choice of fitting parameters with reference to Lüthi (2019) in the manuscript, that is, in Section ***2.2.3 Impact Function****, L. 165ff:*

"**In a comparison of calibration results based on a sigmoidal impact function with a more complex 12-step staircase function, Lüthi (2019) found no improvement of calibration skill with the more complex function. Therefore, we use the sigmoidal function in this study.** We define a default impact function with $V_{thresh} = 25.7\ ms^{-1}$ and $V_{half} = 74.7\ ms^{-1}$ that is used for a first, uncalibrated, simulation of global TC damages, and as a starting point for calibration. While V_{half} is fitted during the calibration process, we keep the lower threshold V_{thresh} constant throughout the study. **This is based on the finding by Lüthi (2019) that the variation of more than one of the linearly dependent parameters most likely results in an overfitting during calibration, with physically implausible values for** $\boldsymbol{V_{thresh}}$ **in some world regions.**"

2.4) In addition, US Damage is conveniently a function of wind speed for a specific reason: damage is often insured loss and that does not include flood. Can you comment on that?

Your point is well taken and we thought about this before starting our work. As for flood-related (as well as for rain and surge) damage, that's a wider field. Indeed standard US policies do not cover flood (or explicitly state an exclusion). But in recent cases (such as Katrina, 2005), the insurance commissioner forced direct insurance companies to cover damages 'as they occurred' and many (if not most) times even loss adjusters could not separate (wind/surge/rain) in hindsight, hence reported damages do often include the flooding component. But as the aim of the present effort is to provide a globally consistent and readily available impact model, calibrating to EM-DAT provides such a global yet at least regionally adjusted perspective - and it is not quite clear what EM-DAT reported damages cover in detail (some numbers look rather like total direct rather than only insured damage, some event reportings might even include total economic damage, inclusive - some - business interruption, i.e. indirect impacts). That's why we consider EM-DAT as a lower bound and rather 'best guidance' than 'ground truth' (which unfortunately cannot be established in hindsight).

2.5) Wouldn't it be wise to check the large scale data against 'small scale' engineering data based on different structure types? Or generally, why is the damage different, is it a physical reason (buildings are stronger or weaker than the US.) or a social reason: lower capital, less cost to rebuild, lower value? It would be nice to discuss this, in the conclusions if necessary.

A comparison of the large scale calibration with socio-economic indicators and engineering based impact functions would certainly be a great gain for further research and improvement of the vulnerability component of TC impact models. However, this is a rather wide and complex field: The regional impact functions are a proxy for the aggregate of multiple types of damages to all kinds of different building types. At the same time, the quality of impact data is not good enough to differentiate between these. Furthermore, engineering data is often not publicly available and if so, only available for richer countries and building codes are hard to compare across regions.

A comparison against specific bottom-up data is beyond the scope of this paper that focused on the question how much can be achieved with an event-based, top-down fitting of impact function against reported damage data.
The approach followed for GAR 2013 (Yamin et al., 2014) is definitely a good starting point to regionalize impact functions based on vulnerability indicators rather than empirically. However, our study shows the limitations of such an approach when no comparison to reported damages is done.

Following your suggestion, a welcome next step could be a study combining the empirical evidence provided by reported damage data on the one hand with building codes / socio-economic indicators on the other hand. This approach would be quite challenging, as it adds even more layers of complexity and cascading uncertainties to the calibration. Still, we agree that it could help to gain a better understanding of the drivers of the inter-regional differences in TC vulnerability. Sensitivity analysis on a more local and regional level could be a feasible starting point.

As comment 1.2 by referee #1 points into a similar direction (asking for hypotheses on the reasons for the inter-regional differences between calibrated impact functions), part of the changes to the manuscript we suggest here to take up are duplicates of our suggestions in reaction to comment 1.2. Please refer to AR1 for the full answer to 1.2.

Proposed changes in the manuscript:

In Section 1 (Introduction):
"While the attribution of vulnerability to regional drivers is outside the scope of this study, the results can serve as a starting point for further research dissecting the socio-economic and physical drivers and factors determining vulnerability to TC impacts locally and across the globe."

In Section 6 (Conclusion and Outlook):
"**The substantial over-estimation of TC damages in the North West Pacific with the default impact function opens the question for the drivers of the apparently lower vulnerability in this region. Considering the inability of the model setup to directly represent the impacts from TC surge and pluvial flooding, one would rather expect aggregated calibrated impact functions to be steeper than the default wind impact function. Therefore, we suggest investigating interregional differences in possible other drivers, including protection and construction quality and standards – but also damage reporting practices. A study combining the empirical evidence provided by reported damage data on the one hand with socio-economic indicators on the other hand would be desirable but rather challenging, as this would add even more layers of complexity and cascading uncertainties to the calibration, especially on a global level."**

2.6) Finally, would this be applicable to other models beyond CLIMADA? Why or why not?

Generally yes, the impact functions are not specific to the CLIMADA modeling framework. Also, we would expect the relative inter-regional differences to be robust with other TC impact modeling set-ups, i.e. that the North West Pacific shows the lowest vulnerability as expressed by the flat impact function. The precise shape and scaling of the calibrated impact functions are, however, to a certain degree specific to the decisions and components of the modeling setup, including most prominently:

1. *The representation of hazard by wind alone: an explicit representation of surge and rain would require different impact functions, also for wind.*
2. *The choice of free parameters in the impact function, as already discussed in response to your comment 2.3.*
3. *The value of total asset values (TAV, c.f. Table A3): impact functions would scale differently with a different assumed total inventory value of exposed assets.*
4. *Spatial resolution: The impact functions are calibrated for 10 km resolution. Parameters could change, if hazard and exposure is represented on a higher or lower resolution.*

The first point is now already reflected upon in the manuscript, most prominently in the discussion: "The regionalized impact functions presented here were calibrated for wind-based damage modelling on a spatially aggregated level. Model setups with an explicit representation of related sub-perils like storm surge or torrential rain require different (i.e. flatter) impact functions for the wind-induced share of TC damage, as well as additional impact functions for all sub-perils. Likewise, impact models with an explicit representation of building types and agricultural assets require a more differentiated set of impact functions." (L. 562ff)

However, we agree that points 2, 3 and 4 could be stated more explicitly, to support the use of the study results outside CLIMADA. Therefore, we will add the following limitations to the discussion:

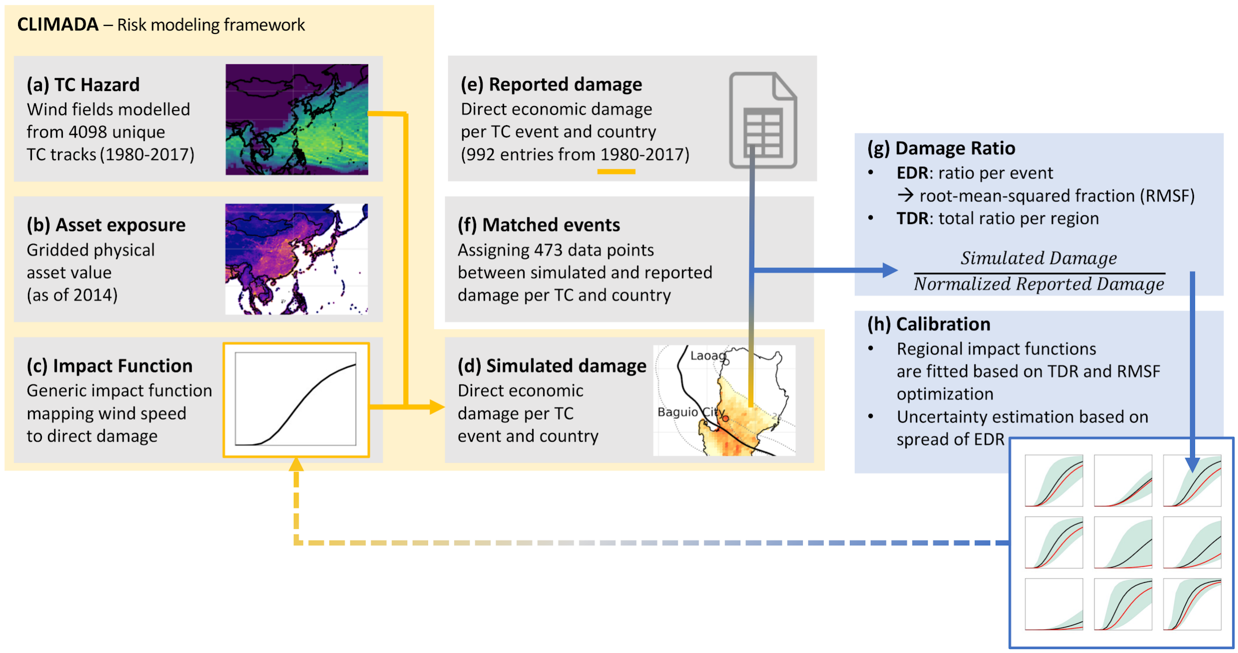
L. 562ff (Section 5.2 Uncertainties and limitations):
"**While the results of this study are not specific to the CLIMADA modeling framework, the precise shape and scaling of the calibrated impact functions are, however, to a certain degree specific to the choices and input data of the modeling setup: (1) The choice of free parameters in the impact function (c.f. Section 2.2.3 and Lüthi, 2019); (2) The TAVs (c.f. Table A3): impact functions would scale differently with a different assumed inventory of exposed assets; (3) spatial resolution; and (4) the representation of hazard intensity:** The regionalized impact functions presented here were calibrated for wind-based damage modelling on a spatially aggregated level. [...]"

Specific Comments:
2.7a) Page 3, L81: Figure 1 is hard to follow. I suggest that perhaps each panel can be labeled a,b,c, etc, and then referred to in the text, rather than focusing on the section numbers. I had to read this 3 times to follow it, and only some of the panels are discussed.
2.7b) Page 3, L90: figure 1: careful with the arrows. For example the arrows in the second row probably point the wrong way. You want 2.2.4 and 'simulated damage' to point to 2.3.2. Not 2.3.1 pointing to them. This highlights the nomenclature problem with the figure.easier to label a,b,c, etc.

Thank you for the well thought out suggestions to make Figure 1 easier to follow. We are taking up both suggestions and propose the following adjustments to Figure 1 and its caption and reference in the manuscript to improve readability:

L. 86:

Figure 1: (1) replace section numbers in panels by labels (a) to (h); (2) remove arrow heads pointing towards panels (e) and (d).



L. 89 (caption Figure 1):

"Figure 1: Schematic overview of the data and methods applied to calibrate regional TC impact functions in a globally consistent manner. From left to right: TC event damages are first simulated within the CLIMADA framework based on TC ~~tracks~~ **hazard (a)**, asset exposure **(b)**, and a default impact function **(c), c.f. Sect. 2.1 to 2.2.3** ~~(a, Sect. 2.1 to 2.2.3)~~. Resulting simulated damages **(d)** are ~~matched and~~ compared to reported damage data from EM-DAT **(e) for 473 matched TC events (f) by means of the damage ratio (g), c.f. Sect. 2.2.4 to 2.3.2** ~~(b, 2.2.4 to 2.3.2)~~. During calibration **(h), steps (c) to (g) are repeated several times with varied impact functions for each region, optimizing the cost functions TDR and RMSF (c.f. Sect. 2.3.3)** ~~impact modelling and damage comparison are repeated several times for regional impact functions with varied slope (2.3.3)~~. The result is a set of best fitting impact functions for nine world regions **(Sect. 3.2)**~~(c)~~. Finally, the calibrated impact functions are plugged into CLIMADA **once more** (d**ashed arrow**) to compute annual average damage per region **(Sect. 3.3)**."

L. 80ff:

"To regionally calibrate TC impact functions, simulated damages are compared to reported damages, as illustrated in Figure 1: In a first step, direct economic damage caused by TCs are simulated in the impact modelling framework CLIMADA (**Fig. 1a-d,** Sect. 2.1 to 2.2.2) with one single default impact function applied globally to start from (Sect. 2.2.3). Then, damage data points per country and storm are assigned to entries of reported damage (**Fig. 1e-f,** Sect. 2.3.1). For the matched events, the ratio between simulated and reported damage is calculated (**Fig. 1g**, Sect. 2.3.2). For calibration, countries are clustered into regions and two complementary cost functions are optimized based on the damage ratios, by regionally fitting the slope of the impact function (**Fig. 1h**, Sect. 2.3.3)."

2.8) Page 3, L103: chosen resolution.

Suggested change in L.103:
“The setup ~~does~~ work**s** equally well at higher **chosen** resolution~~s~~ [...]”

2.9) Page 3, L106: this description is a little confusing, and I think it is because you need to be clear about terminology. What is a hazard? What is exposure data? Maybe start from the concept of damage = exposed assets x damage ratio, and damage ratio is an impact function x hazard intensity. I think those are the correct terms.

Suggested change in L.103ff:
“**In the CLIMADA framework, damage is defined as the product of exposed assets and a damage ratio. In our case,** ~~Simulated~~ damage per TC event and country is ~~computed~~ **simulated** as ~~following~~ **follows: For each grid cell and event, damage is calculated as the product of total exposed asset values and the mean damage ratio. The mean damage ratio (0 to 100%) results from plugging the hazard intensity (maximum sustained wind speed) into the impact function. Finally, damage per event is aggregated over all grid cells within the country.** ~~(1) For each grid cell and event, the mean damage ratio (0 to 100%) is determined by plugging the maximum sustained wind speed (hazard intensity) into an impact function. (2) Absolute damage per grid cell is computed by multiplying the mean damage ratio with the value of exposed assets at the grid cell. (3) The total damage per country and event is computed as the sum over all grid cells within the country.~~”

2.10) Page 4, L145 : can be constrained

*We will correct the typo in L. 145 (“*constraint*” to “***constrained***”).*

2.11) Page 5, L175: is v-thresh fixed? Seems like you just vary v-half, but I can see how v-thresh depends on building type. I.e wood v. Stone.

As already mentioned in response to comment 2.5, the regional impact functions are a proxy for the aggregate of multiple types of damages to all kinds of different building types. Due to the global and aggregated scope of this study, our approach did not start from differences of v_thresh with regards to building types but rather explored how far we get with a top-down approach that is “blind” to bottom-up specifications.
Still, we agree that a variation of v_thres could be worthwhile. We have explained our reasons to vary v_half only in the answer to comment 2.3 above. Please refer there for the detailed response and also the additional clarifications with reference to Lüthi (2019) we suggest to add the manuscript before publication.

2.12) Page 7, L226: is the 20% difference significant? Or is the goal here to make sure the 58% number from climada matches the 76% value from observations?

In Section 2.3.1, we state percentages to examine and illustrate what share of simulated and reported TC damages is represented by the matched TC events, that is, considered in the analysis. The fact that *both normalized reported damages and simulated damages represent more than half of the total damage inventory of the two data sets gives us confidence that a representative calibration can be based on the matched events. If the shares were much lower, we would be much less confident that the calibration results are representative.*

In light of the reason to calculate these shares, the difference of 20% has not been of major interest or concern to us: it simply reflects that the two datasets are both not necessarily complete inventories of damaging TCs, and also that the reported damage data comes with substantial uncertainties as discussed in Section 5.2. The mismatch between the event inventories of the two data sets was one of the reasons to limit calibration on those events with a validated match between the TC track from IBTrACs and the data point from EM-DAT.

2.13) Page 7, L234: what does the distribution of EDR look like un optimized? Can you plot it?

The distribution of un-optimized EDR can indeed be shown in a more concise fashion. The distribution of uncalibrated EDR per calibration region is shown below as boxplots, highlighting the differences between the regions. Furthermore, we plotted histograms for global EDR in response to the referee's request: The histograms of the global distribution of uncalibrated EDR show a large spread, as already discussed in the manuscript (L. 298ff), with both over- and underestimation of simulated damages occurring. Calibration reduces the spread (Fig. S5b-c) to a certain degree, placing more than half of events in the EDR range from 0.1 to 10, that is, simulated event damage is of the same order of magnitude as normalized reported damage.

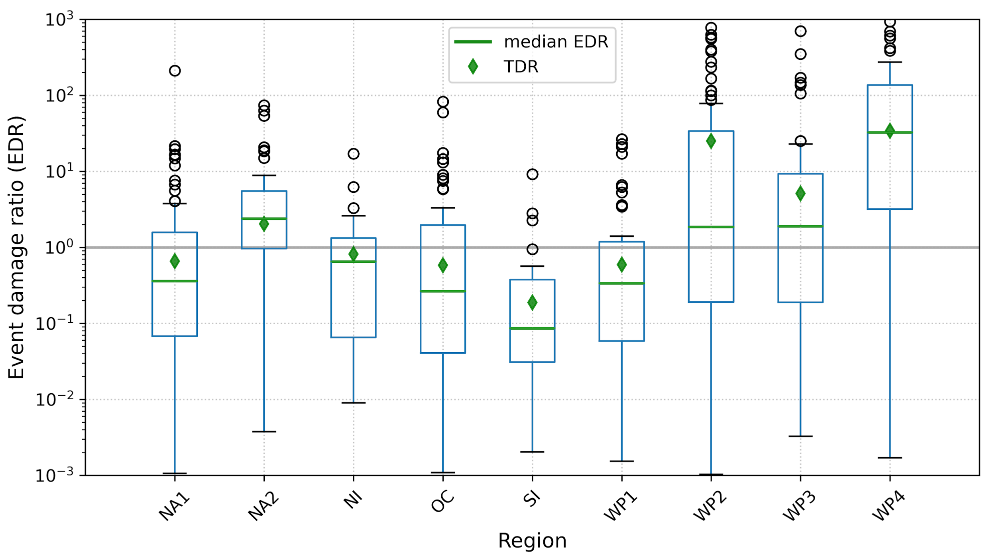


Figure: Spread of event damage ratio (EDR, boxplot) and total damage ratio (TDR) per region before calibration (V_{half}=74.7 ms^{-1}) per region. The plots are based on data from 473 TC events affecting 53 countries. The EDR boxplots show the median (green line), the first and third quartiles (IQR, blue box), data points outside the IQR but not more than 1.5·IQR distance from either the first or the third quartile (black whiskers), and outliers (black circles). The additional markers show TDR before calibration (green diamond).

For the distribution of EDR per country please refer to the boxplots in the Supplement (URL, updated figures also attached to the end of this document).
Change to the manuscript: We suggest to add the regional boxplots of EDR to the results Section 3.1 (L. 283, “Damage ratio with default impact function”*) the histogram to the supplement of the manuscript.*

2.14) Page 7, L238: a plot of the TDR by country would be useful too.

We have already provided plots of the distribution of uncalibrated EDR per country in Figure S2 in the supplement. To satisfy your request for TDR per country, we added TDR (which can be read as a weighted average of EDR) to these plots, both before and after calibration.

The updated Figures S2a-f are attached below, at the end of this document, for your consideration.

Change in Section 2.3.2 of the manuscript:
L. 238f:
“The distribution of EDR and TDR before calibration as well as TDR after calibration is shown per country in Figures S2a-f in the Supplement.”

2.15) Page 7, L245: is a data point a matched storm event? I.e 43 of the 376 events have damage in the USA and Canada?

Yes, the number of matched events per region is listed in Table A2.
To clarify L.245, we suggest the following change in the manuscript:

L.245:
“a minimum desired number of 30 data points **(matched TC events)** per region”

2.16) Page 8, L281: what if you fit v-thresh instead or in addition? Might this help? Why or why not?

To avoid redundancy between the responses, please refer to the replies to comments 2.3 and 2.11 for a discussion of the choice of free parameters in the impact function.

2.17) Page 9, L286:figure a1 and a2 should be part of the main text. Comment further please on the uncertainties. Is damage higher or lower? Where? What are the general issues?

The referee has a valid point in suggesting to add a figure showing the spread of EDR to the main text. In our opinion, it is not necessarily required to add Figures A1 and A2 to the main text to answer the questions asked with this comment, not to overload the manuscript with figures. The question whether and where damage is higher or lower before and after calibration is now already answered on an aggregated and more digestible level in Figure 5 within the main manuscript: Figure 5b compares TDR with and without calibration for each region and globally, Figure 5d shows normalized reported damage and total simulated damage per region, again with and without calibration. As for the spread of EDR (damage ratio of single events), we agree that this should be shown in the main text. We suggest *showing this in the form of boxplots per region, as they also show the spread of EDR to get a better feeling for the uncertainties. The suggested Figure is shown in response to comment 2.13.* *Ad uncertainties: The uncertainties are commented on quite extensively in Section 5.6 ("Uncertainties and limitations") already. The spread of uncalibrated EDR within each region shown in Figure A1 now already illustrates the quantitative extent of the uncertainties, this will be further improved with the suggested figure. Furthermore, the discrepancy between TDR and RMSF calibration gives a hint on how robust the calibration is for each region (Figures 5b and 5c). This is already discussed in the manuscript already, i.e. in line 319: "The comparison of complementary calibration approaches gives an indication of the robustness of the calibration per region." and line 520f: "The deviation between the results of the two calibration approaches indicates how robust the calibration is with regards to the model's ability to represent the correct order of magnitude of single event damage. Whereas the model setup returns reasonable risk estimates and consistent calibration results for Central and North America, we found an extensive spread in EDR and calibration results for other regions, especially in East Asia."*

Suggested changes to the manuscript: c.f. Comment 2.13.

2.18) Page 9, L291: higher or lower?

There are both cases, regions with higher and with lower simulated damages. As described in lines 291ff, uncalibrated TDR is below 1 in some regions and above 1 in others, c.f. answer to comment 2.2 above.

To clarify, we suggest the following reformulation in Section 3.2.1:
L. 291:
"For most regions, ~~TDR is less than one order of magnitude different from one~~ **total simulated and normalized reported damage deviates less than one order of magnitude.**"

2.19) Page 9, L300: maybe it's just showing and figA1 and A2, but I in you should show one more step before figure 5. It's too hard to interpret what the optimization is doing and whether simulated damage is generally too high or too low.

This is a valid point: insights into the optimization and its consequences for damage ratios is a crucial aspect of this publication. As already mentioned in our responses to comments 2.2 and 2.19, Figure 5b does indeed show where simulated damages are generally too high or too low as compared to reported damages, both before and after calibration.
As for your request to get more insight in what the optimization is doing on a more detailed level, we hope that our responses and propositions in response to comments 2.2, 2.13, and 2.14 allow for a more thorough interpretation of what the implications of the calibration on a more detailed level, e.g. with regards to TC severity and per country. In our opinion, the additional figure suggested in response to comment 2.13 serves to show the additional step before Figure 5 as requested by the referee.

A further, more complex insight into the optimization on the basis of fitted V_{half} is provided in Figure S3 in the supplement, providing insight in the robustness of TDR and RMSF to changes in the impact function. We agree that this could be mentioned more explicitly in the text, though, as proposed below.

Proposed changes in the manuscript:

L. 316f: "~~Plots~~ **The robustness** of TDR and RMSF per region ~~as functions of~~ **to changes in** V_{half} ~~are provided~~ **is visualized** in the Supplement**: Regions with a large uncertainty, i.e. a large spread of EDR, generally show a relatively low robustness of the cost functions** (Fig. S3)."

2.20) Page 10, L324:is there a clearer way (eg fig A1) to show h overestimation?

The over- and underestimation of TC damages as expressed by TDR (aggregated level) and EDR (single events) is shown in Figure 5 and Figures A1 and A2. For the aggregated level, TDR before and after calibration shows the average under- and overestimation aggregated per region in Figure (5b), c.f. Also our response to comment 2.2.
For the event level, please refer to the EDR plots per country as shown in Figure S2 in the supplement. Beyond this, we hope that our explanations and additional figures proposed in response to your comments 2.2 and 2.13 help to clarify the communication of over- and underestimation of TC damages.

2.21) Page 19, L591: can you speculate a little more on what next steps might be? Modify or add a flood risk component?

We would strongly suggest to work towards future TC risk assessments based on modelled TC events and an explicit representation of surge and rain. Adding a storm surge component, requires high resolution to resolve topography. Representation of torrential rain, requires to consider transition speed of the TC, as in the case of Hurricane Harvey that stayed stationary over Houston for a long time, dumping more rain than expected in the same area. Also, interaction with other weather phenomena like monsoon need to be considered, as in the case of the Philippines.

To further improve the impact functions, it would be worthwhile to combine damage data-based calibration with socio-economic and engineering type data, as discussed in response to comment 2.5. Especially if impact functions for different sub-perils (wind, surge, rain) need to be combined, more information and knowledge than the one provided by reported damage data is required to constrain calibration. This could also involve expert judgement and engineering-based impact functions.

The last point can be stressed more in the manuscript, we therefore suggest to add the following sentences in the outlook (Section 6):

L. 589ff:
"When modeling multiple TC sub-perils, aggregated reported damage data are not sufficient to constrain impact function calibration. This might be resolved by consulting socio-economic and engineering type data and knowledge."

Appendix to AC2: Reworked Figures from the Supplement (Figures S2, c.f. comment 2.14 above):

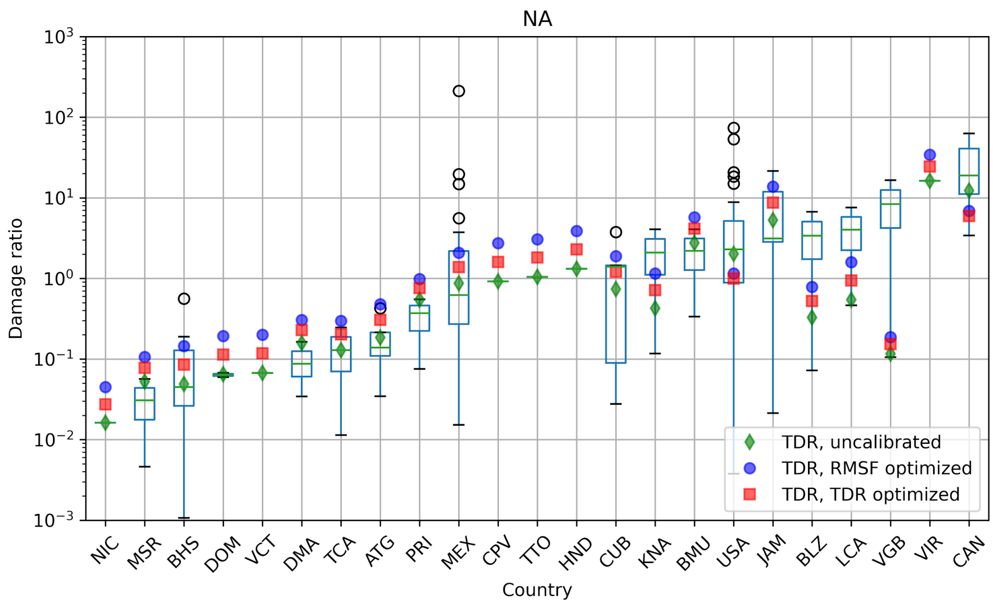


Figure S2a: Spread of event damage ratio (EDR, uncalibrated) and total damage ratio (TDR) per country in the North Atlantic and North East Pacific basin (NA). The plots are based on data from 23 countries. The EDR boxplots show the median (green line), the first and third quartiles (IQR, blue box), data points outside the IQR but not more than 1.5·IQR distance from either the first or the third quartile (black whiskers), and outliers (black circles). The additional markers show TDR before calibrated (green diamond) and after calibration (blue circle: RMSF optimized and red squares: TDR optimized).

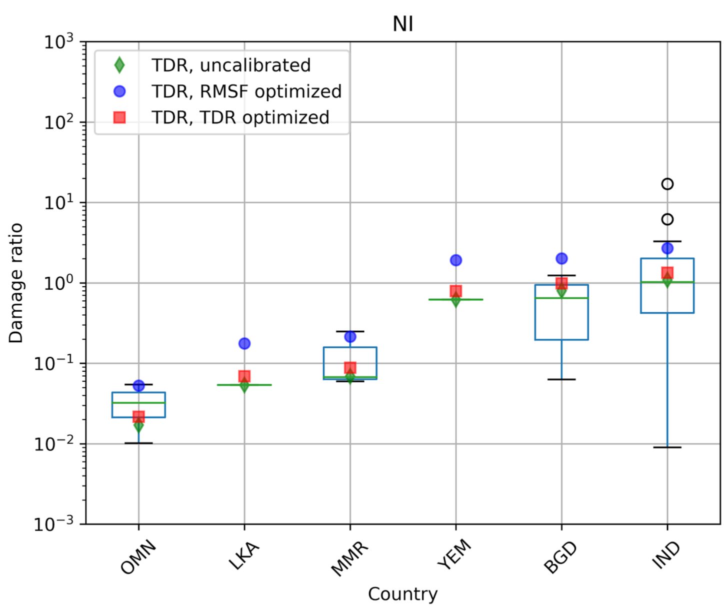


Figure S2b: Spread of event damage ratio (EDR, uncalibrated) and total damage ratio (TDR) per country in the North Indian Ocean basin (NI). The plots are based on data from six countries. The EDR boxplots show the median (green line), the first and third quartiles (IQR, blue box), data points outside the IQR but not more than 1.5·IQR distance from either the first or the third quartile (black whiskers), and outliers (black circles). The additional markers show TDR before calibrated (green diamond) and after calibration (blue circle: RMSF optimized and red squares: TDR optimized).

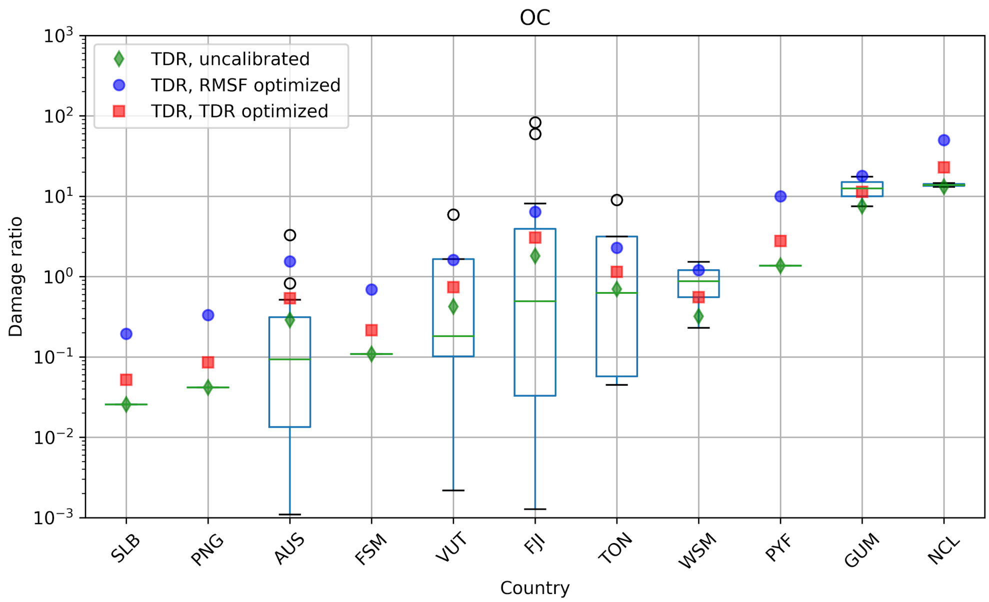


Figure S2c: Spread of event damage ratio (EDR, uncalibrated) and total damage ratio (TDR) per country in Oceania with Australia (OC). The plots are based on data from 11 countries. The EDR boxplots show the median (green line), the first and third quartiles (IQR, blue box), data points outside the IQR but not more than 1.5·IQR distance from either the first or the third quartile (black whiskers), and outliers (black circles). The additional markers show TDR before calibrated (green diamond) and after calibration (blue circle: RMSF optimized and red squares: TDR optimized).

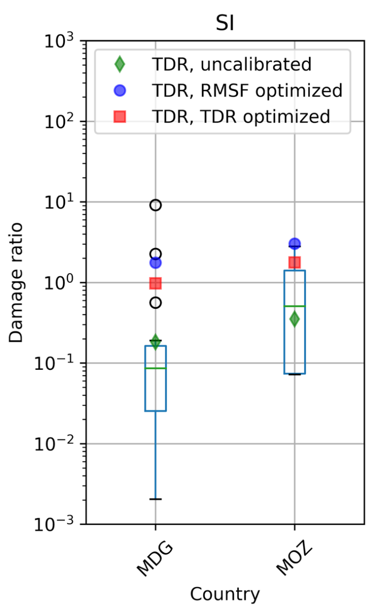


Figure S2c: Spread of event damage ratio (EDR, uncalibrated) and total damage ratio (TDR) per country in the South Indian Ocean basin (SI). The plots are based on data from two countries. The EDR boxplots show the median (green line), the first and third quartiles (IQR, blue box), data points outside the IQR but not more than 1.5·IQR distance from either the first or the third quartile (black whiskers), and outliers (black circles). The additional markers show TDR before calibrated (green diamond) and after calibration (blue circle: RMSF optimized and red squares: TDR optimized).

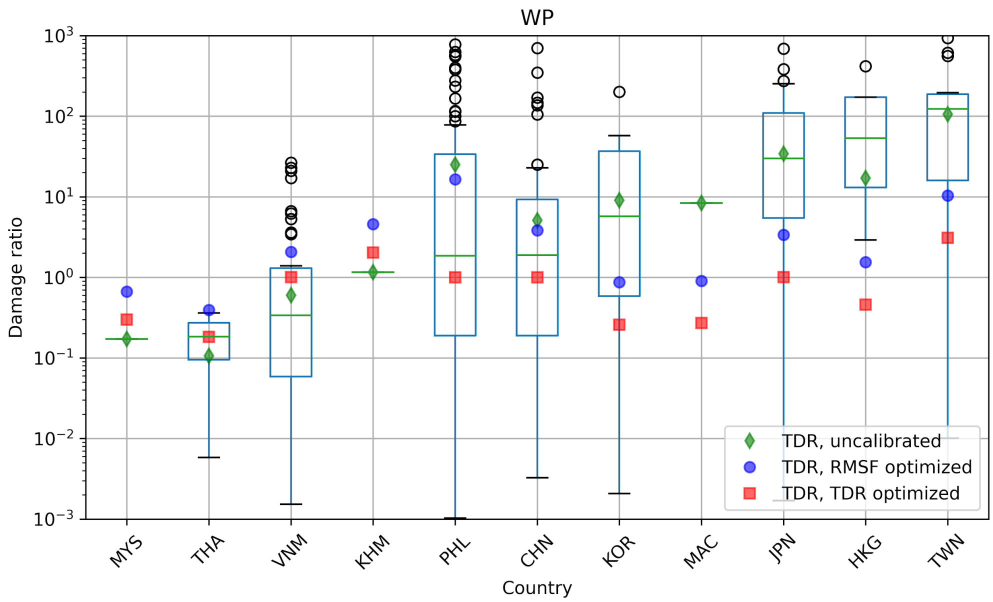


Figure S2d: Spread of event damage ratio (EDR, uncalibrated) and total damage ratio (TDR) per country in the North West Pacific basin (WP). The plots are based on data from 11 countries. The EDR boxplots show the median (green line), the first and third quartiles (IQR, blue box), data points outside the IQR but not more than 1.5·IQR distance from either the first or the third quartile (black whiskers), and outliers (black circles). The additional markers show TDR before calibrated (green diamond) and after calibration (blue circle: RMSF optimized and red squares: TDR optimized).