Review "Characteristics of hail hazard in South Africa based on satellite detection of convection storms" by Punge et al

Reviewer 1

The authors thank the reviewer again for the comments and suggestions. All points raised by the reviewer have been taken into account in the revised version of the manuscript. Our answers below are shown in blue.

Summary

This article is the second revision of the paper with the same title. In this paper, the authors investigate the formation of hail events and hail hazard in South Africa as guided by 14 years of geostationary satellite observations storms. A multivariate stochastic model was built simulating event properties spanning 25 000 years of hail occurrence using overshooting cloud top detection to describe the spatio-temporal extent of potential hail events. Hail footprints were generated that could be used by the insurance industry for risk analyses.

The revised paper shows where revision was done and the authors took care to answer the questions from the first review. The suggestions made by the reviewer are minor and will help with the clarification of a few minor points.

Below are some questions and suggested corrections:

1. Page 1

"to estimate risk for the insurance sector (Punge et al., 2014; Radler et al.)." The date for the reference is missing. Please double-check check rest of the references for a similar problem.
An entry was missing from the bibtex file. We checked all the references but found only this incorrect entry.

2. Page 13-15

The description here is a bit fuzzy for me. I understand the differences in groupings of grid sizes for hail events and for time. However, the manner in which it is described is not clear for me.

For the number of events, the 0.3°x0.5° is combined into new grids of 3°x5° by grouping 10x10 of the smaller grid cells (I assume the 0.3°x0.5° grid cells) together. Then days are drawn from a newly formed 3°x5° grid and the surrounding 8 boxes. What are the sizes of these boxes (0.3°x0.5° or 3°x5°)?
The surrounding boxes are also the large boxes of 3°x5°. This is included in the text.

Are these 8 boxes chosen in any particular manner?
Yes, as explained below, with this method we obtain regions that "represent the scale of synoptic processes and flow patterns governing the spatial (and temporal) clustering of SCS, for example, by specific weather regimes such as Baltic blocking". To avoid confusion, we changed the word "The process" to "The spatial smoothing technique described above"

The switch between the terminology of "grid", "domain" and "boxes" becomes confusing.

Right. Domain in the revised version only refers to the entire study domain in South Africa; we changed that here and in the entire manuscript. Boxes mean grid cell; we changed accordingly.

For time, is the centre point taken in the newly formed 3°x5° grid and then taking a new grid over the centre point of the size 10°x6°
Correct; we have slightly reworded this sentence for clarity.

3. Section 3.1

Event lengths and widths are approximated with GEV over the exponential distribution due to the better fit. Its stated that the GEV fits well over the bottom tail of the distribution. But it does not fit well for low widths that are over-represented (over-estimated) and it does not fit well in the upper tails in that it gives unrealistic large values.

a. Which GEV function was used?
   Table 1 shows all three parameters of the GEV; for all three parameters, $\kappa > 0$, thus the GEV is the Frechet (Fisher-Tippett I) distribution; we included that in the text.
b. From the description, it is my understanding that the good fit of GEV is only for the bottom tail of the distribution for length of the storm. And not really anywhere else on the distributions? Should a different GEV function be applied to get a better fit? What statistical tests were performed to see the goodness-of-fit of the tested distributions?
   This cannot be said. As shown in Figure 11 (Figure 11c for event length), most parameters are reproduced very reliably by the model using the GEV across all dimensions. In almost all cases, the model results are within the error bars of the historical event set - particularly for length and width, which are the most important for damage. We added a comment on this.
c. How was the value 1.5 x largest observed value chosen as the point/place where to truncate the events? It appears this was done for the whole country – but what this checked to hold true for the whole country?
   As written in the manuscript, "length and width were truncated at 1.5 times the largest observed values at which events effectively cover the entire country (1 445 km x 677 km)". This was checked for the whole country.
d. Figures 11 a and b and g are not discussed in the text
   We included a brief discussion of all Figures from the panel, including a, b, and g.

4. Section 3.2

The authors state in this section that due to the large uncertainty the hail size estimated were not considered for the modelling approach using geostationary satellite measurements alone. And that a severity index was created as a substitute.
Yes, this was the major point of the 2nd reviewer to exclude hail size estimation from the satellite data, which we followed in previously revised manuscript (see the large deleted parts).

The section, however, does not elaborate on how the severity index was set up in terms of the range of the scale. Or some descriptives on how this scale looks like or work in terms of

the available data for South Africa. It is discussed throughout the rest of the paper but it does leave this section feeling unfinished.
The severity index is largely discussed in the Appendix. Both the reliability of the methods and the uncertainty are assessed using the Maximum Expected Size of Hail (MESH). We think this discussion is sufficient, also since the severity index is not central in the paper and the risk model.

Section 3.3 in terms discusses how the hail size can be calculated from the stochastic modelling using data from the ESWD and Severe Storms Archive. It makes the assumption that the largest hail size distributions over the continent will the same for South Africa. The authors can discuss the level of uncertainty (although not modelled explicitly) that this assumption can bring into the modelling process.
It's very difficult to estimate the resulting uncertainty in the model. Nor does a discussion of what we do not know really help the reader. However, we have included a statement on this, explaining that the resulting uncertainty can be reduced by calibrating the model using past loss events. By the way, cat models usually require calibration.

5. Section 3.5

- Page 21: "Even if there is a strong correlation between all regions, smaller regions tend to experience relatively higher variability"
  What is the definition of the regions in this context? And what is considered the larger vs the smaller regions?
  This statement refers to the regions shown in Fig. 14, i.e. KwaZulu-Natal (KZN), Highveld (HVD) and Gauteng (GAU) (added in the manuscript). We have not systematically examined event variability for smaller regions (smaller than these three); however, due to the short temporal and small spatial scales of hail events, variability is generally greater for small areas compared to larger areas (also included a brief explanation).
- Page 24: The comparison between the modelled number of hail days for Gauteng (26 days) against that of Smith et al of 69 seems like a big difference. A description follows from how the results from table 3 can change when events larger than different event sizes are viewed. But it is not related back to the 69 events of Smith et al. From Table 3 it is not clear for what events sizes (>= to what cm size) the days are valid.
  Unfortunately, Smith et al. (1998) do not provide information on the number of events within the three different categories. However, assuming an exponential hail size distribution, it can be assumed that events in the lowest class (3-10 mm, which includes sleet/graupel) dominate the statistics. Small diameter events are not included in our stochastic event set as they are not relevant to damage. We have added a few lines on this to explain the discrepancies. This insertion hopefully makes the transition to severe hail days clearer.

6. Section 4.1

- Bottom page 24 discusses applying frequency-weights for Figure 17b and c but not which frequency weights are used and where it was obtained from.

We added the frequency weights and an explanation for that: “These frequency weights are hence the inverse of the retained fraction per class, i.e. 40, 13.3, 20, 10, 1 for classes 1 to 5.”

- Figures 17, 18 – it is not clear if the number of events referred to are that of the observed geostationary data or from the modelled dataset
  All Figures show the results of the stochastically generated event set. To make this clear, we changed the heading into “Modeled event footprints”

7. Section 4.2

- Page 27: “while Fig 18d presents the same occurrence for maximum hail severity indicator greater than 2.” Should this be 2cm?
  Corrected
- “We also note that the local hail count per year is about 2 in KwaZulu-Natal maximum and 1 in the Highveld and Gauteng region so…”
  Does this sentence refer to the number of hail events per year, hail events per year over a certain hail size/severity index? This seems like a very low value per year.
  This applies to the frequency-weighed hail count per year; we included that in the text.
- Bottom of page 27: From line 465 – the event sized discussed – are these the maximum event sizes expected or the average event sizes expected per 10 year period? Where are these values compared with actually observed hail sizes as seen from newspaper/twitter reports?
  The figure shows the maximum hail size occurring once in 10 years (word maximum included in the text). We are not aware of any study or report that can be used for comparison (i.e. relating frequency to diameter). However, based on the few reports available, our estimates appear to be reliable.

Publish with minor revision that the editor can check.

Reviewer 2

The authors thank the reviewer again for his/her extensive work and the additional comments and suggestions. All points raised by the reviewer have been taken into account in the revised version of the manuscript. Our answers below are shown in blue.

The authors clearly expended a lot of effort addressing many of my and the other reviewers' concerns. I appreciate their attention to detail. The new phrasing makes it clear an explicit OT temperatures – hail size relationship is not being established herein, which was my main concern in my previous review. Added text, figures, and appendices help clarify the process and provides verification context. I still have some concerns remaining, mainly requests for additional clarification and requests for reader cautioning, but they are not insurmountable to address.

Note: all line numbers herein refer to the author tracked changes document.

Major comments:

My major comments fall into two categories: requests for additional clarification, and points to make to caution the future reader.

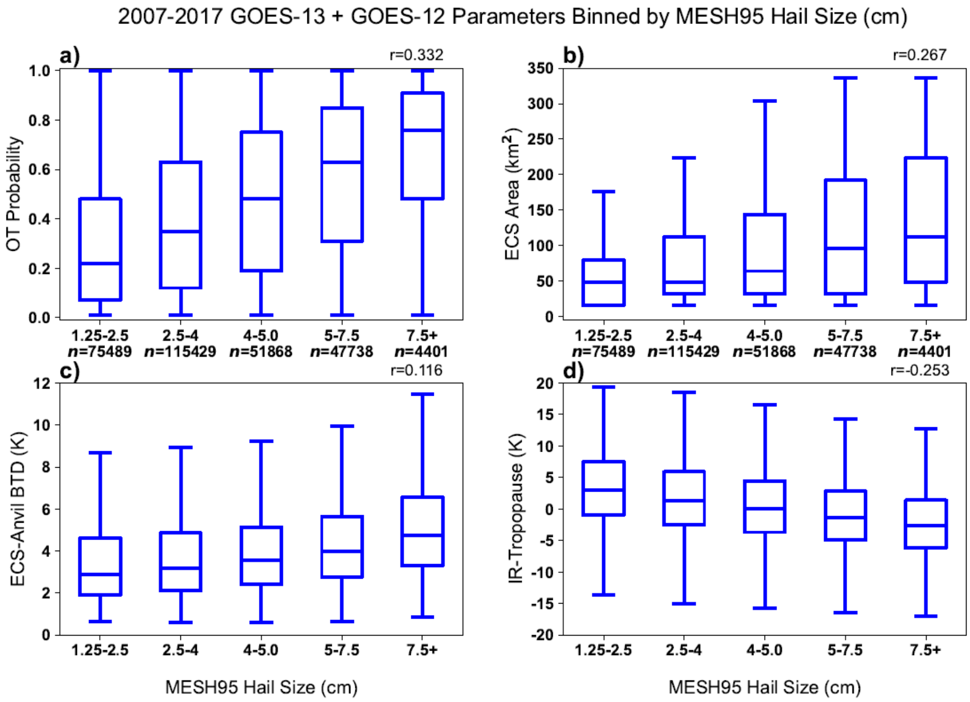
Additional clarification:

- Lines 187- 202: What time ERA5 file is used for the insurance claims, considering the insurance data doesn't have a time of occurrence? Give the odd distribution of parameters for the claims data in Figs. 5a, 5b, it would seem a possibly unrepresentative time was chosen.
  ERA5 were chosen at 12 UTC. While CAPE is dependent on the time of day, sensitivity tests have shown that shear and melt level are not very sensitive. The somewhat odd distributions are due to heavily population biased sampling locations. We have included an explanation in the text.
- Lines 225-228: This additional explanation helps, thank you. That being said, Fig. 7 is still not very clear. Why not separate it into ~3 subfigures over different time intervals, so the three events can be separately shown?
  As the insurance claims have no time, they cannot be separated and aligned to the different OTs/events. In the event definition and also in the stochastic event set, we do not model single streaks, but several streaks on a given day. Separating Fig. 7 into the three different events would imply that each streak is stochastically modelled and not the number of streaks per day.
- Lines 241 - 250: I appreciate the added text, but a bit more clarification is still required. Line 243 introduces the phrase "potential hail events", an excellent addition, but further sentences don't use the phase. "Historic events" here could refer to historic hail detections from TRMM/GPM, S. African claims data, or the Australian/European reports.
  We went through the whole manuscript and changed OT events to "potential hail events" when referring to the historical event set. Btw, the word "OT detection" is now used only when referring to single OTs without event clustering or stochastic modeling.
- I understand that "potential historic hail events" is a mouthful to repeat multiple times; possibly "historic OT events" could be used instead. (And is used, in subsequent sections.)
  In that formulation, "historic" is not required (see explanation above).

- Lines 261-264: A Gaussian distribution is fit to each 3° by 5° box, correct? I'm having trouble following what the phrases in parentheses ("mean number of events " and "summer peak) mean.
  We reformulated the sentence "Depending on the mean number of OT events at a certain location…" and deleted the summer peak
- Lines 266-270: While I appreciate the extra text, I'm still having trouble following this explanation. I understand a random drawing of the day of year occurrence of N events in a grid box, and from each of the surrounding boxes. But are not all of these events retained? What are "blocks of N ⅓"? Most importantly, what N is chosen (and why?)
  N is the grid cell's number of events in the 250 year batch derived from the modelled event frequency. Indeed, only 1 out of 9 events is retained. The exponent (N^(1/3) is tuned to fit observations. 1 would have all events on one day, 0 would retain the original date for each event. We have included an explanation in the text.
- Line 276: No blocks of N⅓ here?
  No, because events don't tend to cluster on hours as much as on days, and sample sizes per day would be quite small per day for reasonable tuning.
- Lines 316- 317: While I appreciate the addition, the application of the nomenclature to this specific case is still a bit murky. What does the objective function predict? The sample data probabilities of.... a specific length and width?
  The objective function is -ln of the product of the probabilities of each sample length given the assumed (GEV) distribution of lengths. Then the distribution parameters were varied to find the maximum of the objective function. We replaced this extended explanation with the former one.
- Lines 349- 375: While this section is clearer, some additional clarification can be provided. State up front at line 361 what statistical correlation relationships are preserved both in the historical dataset and the model. From Punge et al (2014, P14 hereafter) it seems there is a first step that moves from correlations between length-width-hail size to length-width-OT temp difference before the historical dataset can be constructed, is that correct? This explanation will also keep the reader from jumping to conclusions that some sort of OT temp difference and hail size scaling equation exists.
  The correlations refer to all four parameters length, width, area, and severity; correlatios are calculated for each combination (i.e., 16 pairs). We reformulated the sentence accordingly.
- A very quick recap (or reference to specific section of P14) for how track area is determined would also be helpful.
  We added a recap that this refers to the ellipse area determined by length and width and included a reference to Sect. 3.1, where also the formula is given.
- I appreciate the authors uploading P14 to Researchgate so I (and future readers) can review it for answers to these questions.
  You're welcome
- Finally, why are some of the correlations so different between the historical and modeled length/width and event to storm area ratio? Are these values still within the realm of reasonability?
  Larger differences are more or less limited for larger event lengths. However, the number of those events decreases almost exponentially. Thus, small differences have a larger impact on the results. For event lengths below about 200 km, the agreement between historical and modelled events is good, and this is the case for the majority of events (note the logarithmic scale of the color bar).

Cautioning the reader:

- Lines 177-180: This phrasing makes it sound like the extensive calibration of the OT detection algorithm has been for improving its severe weather detection capabilities, but in my opinion has actually been for improving OT detection compared to human ID. The studies cited here found OTs to correlated with severe weather but were not explicitly looking at hail. Punge et al. (2014, 2017) and Jurkovic et al (2015) would be better hail-OT connection citations. It should also be acknowledged here that these sources found OTs near only about 50% of hail events, and Bedka et al (2018) noted that large percentages of OTs do not produce hail. I understand use of OTs is the best option the authors have, but all appropriate caveats need to be acknowledged up front.
  We agree that this issue has to be discussed in more detail. However, as the reviewer also mentioned, it need to be acknowledged up front. Therefore, we extended the discussion of the caveats of OT detections in the introduction and deleted the few sentences in Sect. 2.5. "Still, in some cases, OT features may have been misdetected or may not have produced hail on the ground, for example, due to melting of hailstones during fall through a deep column of warm air. This is acknowledged in the studies of Punge et al. (2014, 2017) and by Bedka et al. (2018) who noted the large percentages of OTs without hail on the ground. However, in addition to the hazard modeling purpose, the focus of our study is on the identification of larger spatial SCS clusters with an increased potential of hail production during the lifetime of the event, rather than detecting each individual storm with enhanced hail potential. These large-scale hail-producing outbreaks can cause by far the largest part of the damage registered by insurers, and can induce solvency issues when the risk was not properly estimated."
- Appendix A: Fantastic addition to the article. It addresses many of the concerns I noted earlier about the relationship between MESH and OT probability. I would ask the authors to provide a few cautioning statements for the reader. Fig. A1, while convincingly establishing a link between increased MESH-estimated hail size and increased OT probability, does not establish a relationship with observed hail size at the ground. MESH is not observing hail fall but is instead a proxy for hailfall based on a storm's ability to loft condensate - essentially, updraft strength, much like OT temperature difference. The relationship between a storm's updraft strength and hail size produced at the surface is not linear, and at larger updraft speeds may in fact be inversely proportional (see Fig. 6 of Lin and Kumjian 2022). Readers should be cautioned against assuming Fig. A1 implies a similar connection with observed surface hail size.
  An upcoming provisionally-accepted paper (Minor revisions, Scarino et al. Artificial Intelligence for the Earth Systems, 2023) expands upon the analyses presented in this section, and shows high uncertainty associated with hail size reports on the ground over the U.S. This is a known issue in the hail research world, was it a shilling, or nickel, "hen egg", "teacup", "golfball", "cricket ball", "softball", etc.. size, and what do these objects correlate to in true physical dimension? The Scarino et al study shows is almost no correlation with any NWP or satellite parameter with observed hail size, whereas there is notable correlation of many of these parameters with MESH. The attached graphic shows this general lack of correlation between these parameters and reported hail size (SPC Hail Size in graphic). The lack of correlation of satellite parameters with observed size is also reflected in the results of Murillo and Homeyer (JAMC, 2019).



We appreciate your concern about a reader inferring that our analysis might suggest that there is a linear relationship between satellite-inferred intensity and hail size.
To address this, we have included the following text in Appendix A: "Therefore, IR-anvil BT difference is a suitable parameter, independent of any reliance on a numerical model, for purposes of modeling the expected hail severity at the ground. Though these results suggest a quasi-linear relationship between MESH and satellite-derived updraft intensity proxies, the true relationship between such proxies and hail size encountered on the ground is unknown, primarily due to known uncertainties with hail size reporting."
We also updated Figure A1 with the one shown above.

- Lines 329- 340: While I appreciate the change in some of the phraseology, the text here still is connecting increased updraft speed with the ability to produce larger hail. While this could be true for smaller hail and/or weaker updrafts, this relationship doesn't hold for stronger updrafts, as Lin and Kumjian (2022) makes clear. (Marion et al. was about tornadoes so is not relevant here.)
  Please caution the reader that updraft strength has been shown to not be directly related with increases in hail size, and for stronger updrafts in particular the relationship potentially reverses. However, given the lack of other available data sauces, OT temperature difference here will be used as an estimate of storm severity, and will be connected to hail size via the reports databases., etc. etc. (I would avoid the term "updraft intensity", as it isn't clear if it means strength, area, or both.)
  As suggested, we modified the statement of the Marion et al. (2019) paper and omitted the term updraft by relying on temperature difference solely. However, the reviewer has to be aware that the intensity measure does not enter our stochastic risk model.

Minor comments:

- Line 10: Damage is not limited to large hail. Large quantities of small hail can be equally problematic, as can almost any size of windblown hail.
  That's right. But we refer in this sentence not to damage at all, but to the significant contribution to natural hazards. In the very detailed insurance loss data we got for a specific

region we see that small hail only cause light damage (< 1% of the insurance losses). Therefore, it makes sense to refer here to large hail.

- Line 25: Cf is used for comparison, but only one figure is listed - perhaps e.g. was meant instead? Also note the reference is to their figure.
  We changed into "see"
- Line 25-26: What methods did Smith et al. use to derive their frequency estimate?
  Same methods and data as Amirat et al. (1985). We included that here, but also give some more details about the reports in the previous sentence.
- Lines 28- 32: Oddly phrased. What problems did Grieser and Hill (2019) face that leads the authors to conclude that hail pad and hail report data aren't sufficient? I'm assuming the difference is quantity of data in South Africa vs. the U.S., but phrasing could be improved.
  Hailpad data and hail reports are always insufficient to quantify hail risk for insurance purposes, because that requires to estimate the damage of a 1 in 200 year event (therefore requires stochastic modeling as later explained). We completely reformulated the two sentence and moved it to the previous paragraph because it does not refer to South Africa. *"Grieser and Hill (2019) used volunteer-collected hail observations in the United States to model the rate of hailstones hitting the ground per unit area, time, and hailstone size bin during the passage of a hailstorm. Based on that data, they set up a model to calculate the vulnerability of subjects at risk as a function of the diameter of the largest hailstone, which can be transferred to other regions."*
- Lines 45- 46: Again, an odd transition. Based on just these sentences, a radar data climatology In S. Africa seems possible. Perhaps adding "but is not available over other large portions of the country" at the end of these sentences.
  To improve the transition between the two phrases, we included further – and no unimportant – information: "*However, the South African radar network does not cover the entire country.*"
- Line 47: "for hail" → "for hail detection"
  We changed into *"proxy for hailstorm detection"*
- Line 57-58: would rephrase to"... an appropriate proxy to assess individual severe convective storms (SCSs) and large-scale outbreaks for the potential of hail production. Large-scale hail-producing outbreaks can cause by far..."
  Changed as suggested
- Line 99, Fig. 2b: Determining where the green colors start in Fig. 2b is difficult. Adding a black outline showing anvil detection would be helpful. Also, adding a sentence pointing out the Great Escarpment and the Drakensberg in the topography map would be useful for later references in the text.
  Very good point. We even thought that it might be helpful for the reader to add a short subsection (now new subsection 2.1) where both the climate and the specific topographic situation of South Africa is described. Here we also explain the Great Escarpment and the Drakensberg.
  Fig. 2b is just a snapshot as an example of an SCS and the spatial extent; we think that details are unimportant for the reader.
- Lines 121- 122: I'd keep the mention of the Sandmæl algorithm but note that South Africa is not continuously covered by visible imagery.
  The point here is that the new Khlopenkov et al. (2021) algorithm goes beyond the Sandmæl algorithm. For this reason and not because of the problem with visible imagery we deleted this reference here.
- Line 124: " ... scanned for hailstorms..." → " .... scanned for OTs..."
  changed

- Line 72: Typically, 20% is used as a threshold probability in hail detection (e.g., Bang and Cecil 2019, 2021). Why the change here?
  We assume this comment refers to line 172 and not to 72 (Introduction). The cited studies were designed to represent significant hail (and threshold tuned to match radar MESH), but our OT methodology also covers smaller hail diameters. We included a statement on that around line 172.
- Lines 145- 146: Prein and Holland (2018) focused on comparison of the distribution of hail environments across the globe and hail detection, not hail size spectra (and they weren't particularly successful in global application.) What publications have focused specifically on observed hail size distributions across the globe?
  Even if the answer is none, I think arguing from scarcity is reasonable enough, it just should be presented with the necessary caveats.
  We added some references that estimated hail size spectra in different regions / continents and deleted the Prein and Holland (2018) reference.
- Lines 147- 149: Any reason not to include reports from the US? With inclusion of MPing and COCORAHS sources, hail smaller than 2 cm could be included in the spectra calculation.
  There is no particular reason why we did not include the data from the USA. However, since our sample already includes about 30,000 reports, we believe that including them further does not really change the results - especially since we only used the data for stochastic hail modeling.
- Lines 182-184: Is the filter based on environments associated with the insurance claims? These sentences and lines 79-80 make it seem like that is the case, but such connections aren't described in this section.
  Yes, we used insurance claims but also microwave detections. We added that to the sentence.
  Because Punge et al. (2017) is behind a paywall (and Bedka et al (2018) mainly just cites Punge et al.) please provide a brief recap here.
  As suggested, we included a brief recap of the principals of the filter algorithm.
- Lines 203-205, Figs. 5c-e: Adding thicker black lines where the filter threshold were chosen would be helpful.
  Added
- Figs. 3,4, 6: I'd prefer grouping these figures all in one figure, to allow for easier, direct comparison.
  We have combined Figs. 3 and 6 into one panel, but not Fig. 4, so as not to give the impression that microwave detection could be used to derive hail frequency.
- Lines 538-539: Which criteria are used in the method described herein, ECS or OT? If OT, why, given that it seems like it misses a lot of hail - producing storms? (Perhaps because of false alarms, which could be indicated in another table column in Table A1.)
  The purpose of this analysis is to demonstrate how detectable radar-observed hail cores can be with satellite cloud top signals. Cooney et al. (JGR, 2021) shows that there can be many ECS' in infrared (IR) imagery around a true updraft core. This is routinely evident in every severe thunderstorm, where an updraft core ejects cold outflow that could appear like a weak OT feature in the absence of other spatial context. The OT detection algorithm has been validated with human OT identifications and OT detections using precip radar echo top by Cooney et al and Khlopenkov et al (JGR, 2021). While some weak OT features can be missed, or a weak cold spot can be mis-interpreted to be a true OT updraft core, we feel that OT detections using the OT probability >= 0.5 is our most reliable way for depicting cell tracks that are necessary for the South Africa hail modeling described in this paper. You are correct in noting that we more frequently have ECS detections near MESH events than OT

detections. What this says is that some typically weaker hailstorms can have very little temperature perturbations in their tops. Using ECS's as a proxy for hailstorms though would result in many anvil regions being identified that are not truly updrafts which would be adverse to our model development. It is unclear to us how a "false alarm" would be characterized in Table A1; would any OT that doesn't have a severe MESH value be considered a false alarm? That would not be very informative because it is well known that many thunderstorms throughout the world routinely produce OTs but not hail. This is why we begin with what we feel are reliable OT updraft core detections, then filter detections to eliminate updrafts in environments not at all supportive of hail, then spatially cluster updrafts to form storm tracks, and then use the satellite-derived storm intensity as a proxy for maximum hail size for our modeling purposes.

- Lines 219-220: What's the temporal resolution of the MSG data used for OT detection? That fact should probably be included in section 2.1.
  Temporal resolution is 15 min. This was already mentioned in Sect. 2.1 (former line 124-125).
- Lines 238-239: Probably should note a filter for these erroneous groupings is being developed, in future work. ☺
  I refrain from mentioning such a filter as this is not really a major issue. However, it's simple to exclude such event types. We have included a short sentence on that.
- Line 241: What model?
  Hail hazard model included
- Line 251: frequency of filtered OTs, correct?
  right, included
- Line 278: "observed and modeled.... OT events"
  No, here we considered events, i.e., grouped OTs. To avoid any confusion, we made this clear at the beginning of Sect. 4 by including: *"This section refers to events of grouped OTs according to the event definition in Sect. 2.4"*
- Fig 10: These distributions are for the entire domain shown in Fig. 9, correct? Please note in caption.
  Correct and included
- Lines 292-293: But both the "historic" and "modeled" events set here are of (filtered) OT detections, correct? So the chance of missing hail reports on the ground is, in this specific context, irrelevant.
  Yes, all analyses and modeling tasks are based on the filtered OTs. To avoid any conclusions, we added in Sect. 2.3 the statement that *"All subsequent analyses presented in the next sections are based on the filtered OT dataset."*
  And yes, we did not further considered missing hail reports. This effect cannot be quantified because of the large number of unreported hail events (we checked that in previous studies for other countries where we got geo-referenced hail damage data).
- Line 305: "covered by OTs/hail streaks" → covered by individual OTs"
  This refers not to individual, but to grouped OTs (i.e., events). We changed that to "OT events". Further, we hope this is now clear with the explanation added above.
- Line 306: While I understand your meaning here, since this ratio has an inherent upper bound of 1 the phrasing could be better.
  Yes, this sentence is a bit strange. We split that into two sentences.
- Line 307: Isn't this product f^2?
  No, f refers to the area and not length and width separately.
  We cannot follow why this should be proportional to the square of f?
- Line 308: " > 105 km², not shown"
  It's the logarithm in the former Fig. 11e, thus 10^5 is correct.

- Line 328: "Storm's severity" → "tornadic intensity". "Storm severity" is too nebulous a term, since it could be interpreted as meaning "updraft strength", which is not necessarily correlated with severe impacts on the ground.
  Changed accordingly
- Lines 381- 384: Why was this figure removed? A comparison of the frequency of occurrence of hail events (not including size) across the country in the historical vs. stochastic datasets seems of prime importance. If the figure is not retained, then discussion about it should be eliminated.
  This Figure was removed to save space, but if that is not a prime concern - happy to include it again based on the reviewer feedback.
- Section 3.6: Much improved, and an interesting result when compared to the Smith et al. study. If I am reading correctly, the stochastic event set underestimates the occurrence of hail days in the region, but potentially overestimates severe hail. It's possible the break down in the updraft strength - large hail relationship is causing these large biased "severe" hail numbers (also possible large hail is underreported, as you note.) Any idea what could be causing the overestimation of hail in general?
- Indeed, while we didn't mention it explicitly, in the model every other hail event has 3 or more cm, and that would mean more than 7 such severe hail days per year for the Gauteng region, much higher than Smith's result. Besides underreporting of large hail in the Smith study, we also need to consider underreporting of small hail in the model's severity data, which could explain an overall overestimation of large hail.
- Lines 505- 507: While I agree with these statements, I would shift them earlier in the conclusions as they are awkwardly placed here.
  We have moved the two sentences to the beginning of the Conclusion section.
- Fig. B1: Great addition. Can these plots be normalized by total number of detections and plotted in the same plot for easier comparison?
  We exchanged Fig. B1 with a normalized graph.
- Lines 570- 580: Excellent addition. I'd rearrange the text (or figure) so the subfigures are referenced in order.
  We rearranged the text as suggested, but also added a short statement for Fig. B1a and e.

Grammatical:

All suggestions / corrections were considered

- Line 47: "this" → "these"
- Line 131:" ... east of..." → " .. in east..."
- Line 206: Add "Fig. 6" after Fig. 3, so the comparison has an object.
- Line 510: "95." → "95th"
- line 221: "of the event, or grouped, OTs"
- Line 267: "of" → "from"
- Fig 11: years → years'
- Line 417: "hail hazard" → "the hail hazard"