The authors thank the two reviewers for their helpful and detailed comments. We have considered all the points they listed. We went through the entire paper again (several times), trying to improve clarity and readability. We have deleted two of the original figures that we felt were not really necessary, but have added additional figures in a new appendix.
The added text in the revised manuscript with track highlighting is marked in red.

Author comments to reviewer comments RC1:

Review for NHESS-2021-342

Characteristics of hail hazard in South Africa based on satellite detection of convective storms
Heinz Jurgen Punge1, Kristopher M. Bedka2, Michael Kunz1, Sarah D. Bang3, and Kyle F. Itterly4

Summary.
This paper reviews and analyses the hail climatology for South Africa and regions within South Africa using satellite detection of convective storms. The paper investigates 14 years of geostationary satellite observations of convective storms and generated a spatiotemporal multivariate stochastic model representing 25000 years. The historic, stochastic and observed insurance exposure and vulnerability data are analysed to identify the expected hail damage for return periods of 200 years.
General Comments
Scientifically the authors have done an excellent job considering the literature review, data collection and modelling that have gone into the research described in this article. One slight drawback is that in the paper itself, the authors attempt to address all the whole multiple complex modelling processes in a relatively short and succinct manner. At times, unfortunately, the description of the process followed does not do the modelling process justice and made it difficult to follow what was done. Some examples will be given below.
The format of the paper also made reading the paper very difficult for me. At times, the figures and tables mentioned in the text were not next to or near the text referencing them. This caused a lot of scrolling up and down in pdf (and eventually I just printed out the text in frustration). As times, figures were even placed to appear to be part of a previous section e.g. Fig. 14 seems to form part of the end of Section 3.3 but is part of Section 3.4 that starts underneath Fig. 14. The format used by the authors may be due to format instructions fromthe journal. If not, please reconsider the placements of figures and tables to be as close as possible to the relevant text to improve the reading flow of the paper.

The authors thank the reviewer for the helpful and extensive comments. We considered all points listed below in the revised version of the manuscript (replies in blue).

As suggested, we improved the description of the modelling process. We are sorry for the wrong placement of the figures and Tables. We rearranged their appearance in order to improve readability of the paper.

> *Specific comments*
> *25. The reference to Grieser and Hill (2019). Did Grieser and Hill focus on hailpad derived metrics for South Africa, another country or just in general?*

The study of Grieser and Hilll uses data from CoCoRaHS, a volunteer-based network of weather observations in the United States (Doesken and Reges 2011; Reges et al. 2016). Data of this kind has been collected and published only for very few locations, and no recent study was found for South Africa.We have specified the data origin in the revised version of the manuscript.

> *35. "hailstorm formation is often related to local and meso-scale processes related to, for example" Perhaps do not use the word related twice in one sentence.*

Sentence rephrased ("...hailstorm often form by…").

> *Figure 1 "a title: SRTM" acronym is not defined. Please check whole text for acronyms.*

We have defined SRTM, but deleted the acronym as it is only used here. In addition, we double checked the manuscript to make sure all acronyms are defined in the revised version.

> *95. "Based on past experience, only OTs detected with a probability >50% and with a surrounding anvil cloud (green and yellow colors in Fig. 1a; the IR anvil detection index, a rating based on an anvil detection model accounting for viewing situations, greater than 10; see also Scarino et al., 2020) are used in this work." Make the sentence in brackets a sentence on its own or add to figure description.*

As suggested, we moved the sentence in the brackets to the figure caption.

> *110 - 115. "A uniformly distributed random number between -0.5 and 0.5 was added to each reported hail diameter to compensate" I assume the -0.5 and 0.5 is also in mm?*

Sorry that we did not included the unit; the random value accounting for the coarse classification of hail sizes in the database is actually between -0.5 and +0.5 cm. We have corrected this.

155 "As 9.5% of the OTs occur at a melting level of less than 2 400 m, but ony 3.5% of the microwave hail detections and and 2.5% of the claims, a lower threshold of 2 400 m was introduced for this parameter." . Spelling. Sentence seems incomplete.

We corrected both the spelling and grammar.

> *160. "The latter feature is due to the minimum freezing level condition and remains to be confirmed by independent observation." Independent observation from whom?*

Very few information on hail occurrence was found for that region; the most suitable method of ground verification would be a network of hail pads or sensors covering multiple regions of South Africa. This was added to the sentence.

> *190: "complemented with hail size information from reporting." Reporting? You mean the insurance reporting?*

In this case hail reports such as those registered in the hail databases are meant. This is now clarified in the text.

> *200. "Following Punge et al. (2014), both annual and daily cycles are modeled with Gaussian distributions. For the day of year, domains of 3° × 5° are considered, and depending on the..." Why Gaussian distribution? What is the statistical justification for it?*

The occurrence of hail is linked to conditions on a set of variables that need to be fulfilled and can therefore be described as a convolution of the distributions of these variables. As such distributions, e.g. of insolation, are themselves usually continuous and often normal, their convolution can be expected to be normal as well.
According to literature, both the diurnal and annual cycles can be well approximated by a normal distribution. We added an explanation with references to the manuscript.

> *Not sure how these grid definitions relates to the previously defined rectangular grids of 0.3 x 0.5 mentioned on page 10.*

The grids of 3 x 5 group 10 x 10 of the smaller grids, with the intention of increasing the number of observations in each cell, large enough to derive characteristics of the distribution. This is now explained in the manuscript.

> *205: "Days are drawn from the boxes distribution for the..." It is not clear to what the boxes distributions refer to.*

It refers to the 3° x 5° boxes; this is specified in the revised version.

> *205: "...Finally, the day is retained only for N/9 events at random. This procedure has been found empirically to approximate the observed space-time distribution of days in a satisfactory manner. " Is this procedure self-developed or taken from somewhere? Why n^(1/3) and N/9 - those specific values? What is the proof of empirically proof behind it?*

The method is self-developed. We realise the description was somewhat imprecise. A division by nine is required as we draw nine times the required number of events: for the box concerned and the eight surrounding ones (queen criterion). SCS/hail events preferably cluster on this scale (15°x 9°) due to synoptic processes. The empirical proof is that the distributions in Fig. 17 represent the observed distributions quite well with a single tuning parameter. We improved the description to make this more clear and added two references on the relation between synoptic processes and SCS clusters.

> *210: "from a region of 10° × 6° around.." Why the double grid size? Is this to also represent the 8 neighbouring grid cells? Paragraphs 205 and 210 can be extended to make the spatial construction more clear. In the current format is it difficult to follow, and relate back to standard spatial weight matrixes using the queen criterion.*

The 10°x 6° region turned out to be a good choice for this parameter. The time of day is correlated on a smaller scale spatially as in a series of events, the later ones are shifted spatially with respect to the earlier ones. We have added this explanation.

> *215: "Also note the secondary maximum in fall (around days 100–150, i.e. April and May) during nighttime, represented in the model. " Does this represent a local maximum? How do you see from the graphs it is in the night?*

This maximum is a local maximum in the historic events and occurs at around day of year 140 and 6 UTC and can be discerned as an area of orange shades in that region of the plot. We have added this explanation to the text.

> *215: " It is shifted towards fall over the Southern Ocean." ??? Are you modelling that far away from the shores of the country as well? And will it have any landfall impact?*

Off-shore events need to be represented as they can extend to the onshore coastal region. An impact of a far off-shore event is quite unlikely and will be marginal at this distance, but has been included for completeness. We have added an explanation.

> *220: "Time, slightly earlier than Smith et al. (1998, 5–6 pm) but consistent with Olivier (1990) (Fig. 11b). The daily cycle is most pronounced..." What is the possibility of there being a shift in these times from the 1990's to now? In that case, would the results be comparable?*

A shift in the diurnal distribution of severe convective storms cannot be excluded or proven with the data at hand and has sometimes been discussed in the context of climate change. A more likely explanation would be a different sensitivity to hail size in the two methodologies. Larger hail has a tendency to peak later in the day than small hail and Olivier (1990) may miss some of the smaller hail events. As this explanation is speculative, we haven't included it in the manuscript.

Figure 10: I assume the day of the year for 1 to 365 represents 1 Jan to 31 Dec. Perhaps add that to the title to indirectly show the difference in expected hail occurrences for northern and southern hemisphere?

The assumption is correct and was added to the figure caption.

> *Figure 11: I'm struggling with what number of days each bar represents. It seems the number 50 falls on the 4th bar?? This will only work if each bar represents 12.5 days?*

The days of year have been grouped in classes of 14 days in this figure. "50" hence roughly corresponds to weeks 7 and 8. We added an explanation to the caption.

> *225: "The distributions are well approximated by the GEVs". GEV is an extreme distribution that requires a "limit" (e.g. peaks-over-threshold or block-maxima) in the data over which you are modelling events? What was that limit and how was it obtained?*

There is a possible misunderstanding here in that we are not applying extreme value theory here. Instead, we use the GEV to approximate the distribution of *all* events, not only the most extreme ones. The is now clarified in this respect.

> *230: " to give unrealistic large values, which is why length and width have been truncated at 1.5 times the largest observed values," Why the specific value of 1.5 times the largest observed value?*

It turns out that these distances occur in events that are of a size comparable to South Africa, the domain of interest. Hence the cutoff has little practical implications. Other than that, there is no specific reason to choose this particular value. We added an explanation on this.

> *230: "In addition, the fraction f of the event area (the area of the ellipse spanned by major and minor axis of lengths l and w," Remember to write the last l and w in italics.*

This is now corrected.

> *235: "Table 1 lists the distributions and parameters for these event properties." Which method was used to estimate the parameters of the distributions?*

We used the standard matlab mle function to obtain the maximum likelihood estimate of the distribution parameters and included an explanation in the caption.

> *240: "We find that most frequently, events have an orientation of around 100°, i.e., propagate eastward to southeastward (Fig. 11f)." This is for the whole country. But it may be misleading as this is not the typical orientation for a high hail fall region like Gauteng where storms normally originate in Johannesburg and move north-easterly to Pretoria. As seen from the discussion in the next paragraph.*

This is right; we have added this specification to avoid any confusion.

> *280: "sets of random numbers for each property from a uniform distribution and determine ranks. Then, for each property, we draw values from the actual distribution, sort them, and attribute to events using the pre-determined ranks. " How? Does this again refer to a previously defined or described methodology?*

The methodology has been specifically designed for this application and has been described by Punge et al (2014) regarding the European hail event set; we added the reference to the text.

> *285: "could be expected, smaller regions show relatively higher variability, but there is strong correlation between the two. This" Which 2? Smaller regions and the country as a whole?*

Yes; we have included a statement that there is a strong correlation between all regions.

> *295: This section describes the South African domain in terms of latitude and longitude degrees, subregions etc. Should this description not be done earlier in the paper to set the scene – perhaps where Figure 9 is defined*

We have added a paragraph near Fig. 9 to discuss the importance of these regions for hail hazard.

> *305: "50 hail days per year), while in an equivalent sample of subsets from the stochastic event set, the event count ranges from 1 883 to 2 162 events on 671 to 703 days" Perhaps add the equivalent hail events and days per year for comparison.*

The numbers were added.

> *305-310: " In the Highveld region, there were 74 days per year..." These numbers are averages per year. The averages per year for the years defined by the authors and the years defined by Smith et al 1998 are different and it should be considered that several climate changes occurred in the years in between. This includes periods of severe drought in the several regions in the country especially between 2010 -2020*

We have added a comment highlighting the difference in methodology and the possible impact of climate change and variability.

> *310: Can the numbers given in this paragraph be added in a table for easier reference?*

We included a new table (Table 3) to cover the numbers.

> *310: "However, severe hail (>31 mm in diameter" Why 31 mm and not a round number like 30mm?*

The size threshold was chosen to match that of Smith et al. But we realise they use >= 31mm and hailstone diameters are given in 1mm classes. We therefore changed to >30mm instead, which is indeed more intuitive and should not differ too much from Smith's approach.

Author comments to reviewer comments RC2:

Review for NHESS-2021-342
Characteristics of hail hazard in South Africa based on satellite detection of convective storms
Recommendation: Reject

The authors have assembled a novel methodology for estimating hailfall in South Africa, a region with frequent hailfall but sparse observations. Such a hailfall climatology in South Africa is clearly needed, particularly with the possibility of shifting or increasing hailfall frequency with a changing climate. From this hail event climatology, they have additionally assembled a statistical model that includes estimations of hail size, hail swath shape and orientation, and frequency of occurrence over a much longer period. The creation of all these products is ambitious, but I feel the authors have overreached what is scientifically defensible though the necessary chain of assumptions. I don't doubt that there is a strong operational need for extended hail climatology products like these in this region, but if choosing to publish the work the assumptions must be reasonably defended. In sum, I recommend rejection in the paper's current form, but would welcome reviewing a resubmission on a narrow, better-grounded portion of the work.

The authors appreciate the reviewer's thorough assessment of our manuscript. As the reviewer points out, to cover the entire modeling process from satellite detection of storms to the stochastic footprints can appear ambitious. However, for each of the steps, we can build on existing publications where similar assumptions had to be made, and focus on improving the methodology in the best possible way on the basis of the available data. Presenting all those steps in a succinct way and in one article will benefit other authors pursuing similar objectives or trying to test the accuracy of our results and therefore benefit scientific exchange as a whole. Naturally, the supporting evidence is clearer and the assumptions to be made are weaker in the portions of the work directly dealing with observations compared to the modelling part; this is however a common situation in atmospheric science. To address the reviewers concerns, during the revision of our work for final publication, we will particularly stress the assumptions made and caution needed in interpreting the results. As an example, the use of hail diameters in the model portion may be susceptible to over-interpretation, which is why we will instead center on the aspect of hail severity and avoid the use of hail diameters in the hazard part of the model.

However, hail size as a measure of intensity is needed in the stochastic model to estimate hail risk for a given portfolio. These sizes do not derived from OT intensity estimates, but are modelled stochastically based on observed hail size spectra. This approach is the only way to provide hail risk assessments. It is also implemented more or less in all catastrophe (risk) models.

Major comments/fatal flaws: *The work performed here was obviously extensive, and I appreciate the effort to scientifically ground an operational product. I've broken down my view of the chain of reasoning presented in the paper, along with my opinion of how well each step is grounded in the article.*

1. *Hail occurrence can be estimated using the Khlopenkov et al. (2021) OT detections in GOES data over CONUS. This step is well-grounded, given Khlopenkov et al. and Cooney et al. (2021) results discussed in the introduction, although a quick sentence or two discussing the skill level of that algorithm with the severe hail report database used in those studies would be useful to add.*

The skill level of the algorithm has been assessed against severe hail reports, MESH radar and satellite products. We have summarized the most important findings in Section 2.1, but most of the respective information and details were added to the appendix (including a Figure and the Table shown below).

"To better emulate the present study methodology, MESH cell objects exceeding 2 pixels in area (10 km^2) and spaced by at least 28-km are derived using watershed segmentation applied to the hourly 10 mm+ MESH95 climatology (Bowman and Homeyer, 2017) over CONUS between 2013-2017 using the open-source Tracking and Object Based Analysis of Clouds (tobac v1.2; Heikenfeld et al. 2019) Python package. Further, following Murillo et al. (2021), we have applied Linear Discriminant Analysis (LDA) using their coefficients to combine precipitable water and 0-6-km shear to filter out likely false alarms.

Cooney et al. (2021) showed that we can detect updrafts near to or above the tropopause with about a 60% success rate using data from GOES-13 a proxy for Meteosat (Cooney et al. 2021). Table 1 compares the frequency of GOES-13, GOES-16, and MSG embedded cold spot (ECS) detections, e.g. areas that appear distinctly colder than the surrounding anvil and are considered to be OT candidates, and OT detections (OT probability >= 0.5) matching various hail detections from radar cells, ground spotter reported hail size, and MWR hail detections. Requiring OT probability >= 0.5 to refine severe hail detections to those we are most confident in, we lose 56% (46%) of the severe hail-producing storms exceeding 40 cm MESH95 maxima for GOES-13 (GOES-16). In other words, many severe hailstorms can look quite "boring" from a satellite infrared perspective, but the boring ones are hard to differentiate from false OT detections in anvils (i.e. detections in cold outflow near to real OTs). Uncertainty between report time or the time a radar scanned a storm vs the time of OT detections may also influence our results. For example, an OT may have been prominent several minutes before the time of a hail detection, but we only have a single GOES snapshot to match. By relaxing the matching criterion to ECS detections, we lose only 30 (17%) of likely severe hail producing cells for GOES-13 (GOES-16).

The frequency of geostationary updraft detections that are co-located with microwave hail detections is comparable to MESH95 and ground spotter severe hail reports despite added uncertainty due to parallax shifts in the storm positions in microwave data, especially those close to the limb of the overpass. Enabled by the global coverage of MWRs, Table 1 shows that 2005-2018 MSG SEVIRI ECS and OT detections over South Africa match with likely severe MW hail detections with frequencies similar to GOES-13 over CONUS. Although the total number of matches is relatively low over South Africa, this suggests that MSG IR-based updraft detections agree with independent hail detections; thus, supporting the use of MSG SEVIRI to detect hail cores over South Africa."

2. The Khlopenkov et al. OT algorithm can be applied to MSG SEVIRI data over S. Africa with similar success as GOES data over CONUS, with the additional environmental filtering applied. This claim is generally supported by the results in the paper (c.f., Figs. 3 and 4), but needs a fuller explanation. The geographic hotspots are similar in Figs. 3 and 4, but is the frequency of potential hail occurrences reasonable? Comparison of OTs, GPM/TRMM detections, and radar-based detections over CONUS could confirm the relative change in frequency between OT and GPM/TRMM detections over S. Africa is reasonable. Comparisons should also be made to climatologies made over the region from other methods, such as those discussed in the introduction (Admirat et al. 1985; Prein and Holland 2018; Kunz et al. 2020; Dyson et al. 2020).

There is an inherent difficulty in comparing hail frequency estimates based on different methods, in particular when it is not possible to directly compare each occurrence of hail (see, e.g., the review of Punge and Kunz (2016) on hail frequency estimates in Europe). The nature of the model-based studies cited here and the satellite-based approaches presented in this study are such that such a direct comparison is hardly possible for South Africa.

As TRMM and GPM were/are satellites in inclined orbits, their sampling is not continuous, like a geostationary satellite. It is therefore not reasonable to use the absolute counts from these satellites as indication of the true frequency of hailstorms. When assessing the gridded climatologies as in Bang and Cecil (2019; their figure 7), the values are scaled to account for the sampling. Comparing the CONUS hail events/year to the radar methodology of Cintineo et al. (2012; their figure 9), we see a GPM climatology over the central US that ranges from ~6-13 events per year, while the radar-derived climatology estimates about ~4-12 hail days (note events versus days). A US climatology using MESH to estimate the presence of hail from Murillo et al. (2021) shows a frequency of 3-7 hail days per year over the central US. This substantiates our confidence in the passive-microwave hail retrievals. We have added a statement about that to the discussion of differences with other publications on hail in the region.

3. The hail grouping methodology into events reasonably represents hail swaths from a single storm system. While the description of the methodology (lines 176-177) is intuitive and simple, the results of the grouping methodology in Fig. 7 don't seem to follow that description. Why are there multiple events occurring at a single place and time? Once the methodology itself is cleaned up, a few example applications of this methodology in an area with radar data would show its value in establishing hail events and their duration and speed. Right now, the results of the methodology are only briefly compared in text to two other radar-based studies of severe convective storms (not limited to hailstorms) in the literature.

In the given example, there are indeed overlapping events. Those simply are too distanced in time and space (in particular time) to be grouped into the same event. The algorithm is designed to follow storms related to a common conditions or trigger, such as a propagating front triggering storms along its way. Hence, the temporal aspect will receive stronger emphasis in the revised article. We have commented on this in the text.

4. The created hail event climatology shows reasonable distributions of hail event frequency by time of year and time of day. No comparison of these distributions is made to the observational or GPM/TRMM datasets. While they are admittedly sparse, they should at least be able to confirm general seasonality. Comparisons should also be made to the other climatology datasets mentioned in point 2 above.

In the Appendix, we have included four seasonal frequencies of hailstorm detections based on overshooting top activity (2004-2018) and passive microwave hail retrievals (1998-2018) for austral summer, autumn, winter, and spring. Additionally, we have included two Figures showing the diurnal cycle from OTs and as seen by TRMM and GPM combined.

It turned out that daily cycle of OT activity is more pronounced compared to that for the microwave detection, and its maximum sets in almost an hour earlier. Overall the agreement between the two satellite climatologies, however, is very good, but the OT algorithm may be slightly too sensitive for weaker convection. Concerning seasonality, it turned out that both methods indicate widespread hail activity quite well, with some discrepancies depending on the season. All these points are briefly discussed in the Appendix.

5. The statistical method established in lines 202-213 can be used to produce similar hail event daily and seasonal hail event variations established by points 1-4 above (assuming points 1-4 are successful at representing actual hailfall). The annual and daily distributions produced by the model do appear similar – I'd prefer a difference plot instead of a side-by-side comparison, given the relatively large magnitudes involved. However, the description of the statistical method is not clear, and only one reference is cited. How common are methods like these? The steps involved in its description are very specific, making one wonder if the model is being over-fit to its underlying dataset.
How similar is the methodology used here to Punge et al. (2014, unfortunately behind a paywall), what changes were made, and why?

A difference plot daily and seasonal hail event variation is included, now Figure 10c. In addition, the presentation of the methodology was revised. The methods are a recoded version of the 2014 article. In the redesign, the use of a von-Mises distribution was considered for modeling the periodic variables, at the cost of losing information on the shape of the distributions.

6. The statistical method in lines 225-238 can be used to produce similar hail event length, width, area, and orientation as the event climatology produced in point 3 above (again, assuming point 3 is valid). These results do seem reasonable as presented in Fig. 11, but no point of comparison is provided. How well do other statistical methods perform? What is expected behavior?

Indeed, a host of different options exist to model such relationships, including machine learning models, which may perform better, but the metrics and parameters will have to be chosen carefully to consider all event properties adequately and avoid overfitting and other issues. We instead opted to build on the methodology developed for the Punge et al. 2014 article, which is explainable, reproducible, and uses a rather small set of parameters.

7. Hail size can be estimated using the OT climatology product produced in point 2 (I don't think the event climatology from point 4 is being used here, but text isn't clear). This claim is (currently) indefensible.

To avoid a possible misunderstanding, we clarified the use of the OT product as a proxy for hail intensity (but not size). We do appreciate that a significant amount of uncertainty remains on the exact relation of OT strength and maximum reported hail size.
For the sake of modelling, we do not require this relation, and just need to assume that the size-extent relation holds on average, thus the strongest storms in terms of updraft occur in the largest systems as determined per event detection procedure. Maximum hail sizes in the stochastic event set and the model's event catalogue are drawn from hail size distributions in reports (ESWD or other). Relations between event properties are used only for the problem of matching those hail sizes to the other event characteristics in the stochastic events. We have included some statements to make that more clear.

Marion et al. (2019) suggested a relationship between OT area, not strength, with updraft width and hence potential tornadic intensity. That's a not insignificant difference. Hail size, particularly as one reaches larger hail sizes, is more related *to updraft width than updraft strength (e.g., Nelson 1983, Foote 1984; Kumjian et al. 2021). I am concerned that by relating hail size to an updraft strength metric, an erroneous hail size distribution will be produced.*

We agree that the focus and findings of Marion et al. are different from our study. While tornadic activity may indeed rather be related to the size of a storm system, it is still a reasonable assumption that hail size is related to updraft strength rather than size. There is no doubt that besides updraft strength, other factors like the buoyancy in the UTLS region and timing of the imagery relative to peak intensity also impact the measured cloud top temperature differences. In the manuscript, we have deleted statements about hail sizes derived from OT detections. But we briefly present in the appendix a relation between hail size parameters based on radar – where the spatial match can be expected to be better than with hail reports – and the OT- anvil temperature. The inherent assumption is thus that other factors will average out when the sample size is big enough, which we are confident is the case.

Khlopenkov et al. (2021) connected OT detection probability with hail occurrence and did not try to distinguish among hail sizes.

Actually Figure 11 of Khlopenkov et al. (2021) showed that increasing embedded cold spot (ECS) IR-anvil brightness temperature difference and decreasing IR-tropopause temperature difference, both metrics of storm intensity, were more extreme for 2+ inch hail reports than hail < 2 inch in diameter. But again, we have deleted the part of hail size estimation using OT data.

Figure 2 appears to represent original work from the authors (sentence is oddly phrased, making it seem like it is sourced from Murillo and Homeyer 2019). While I do appreciate the correlation shown, I am concerned the MESH95 dataset is being used, and not actual hail reports. Per Murillo and Homeyer, the MESH95 dataset has a significant large bias, with 40 mm being most skillful at determining 25 mm hail, and 64 mm being most skillful at determining 50 mm hail. That bias does not appear to be accounted for in Fig. 2. Further, while Murillo and Homeyer (2019) did not specifically examine the skill of tropospheric-OT temperature difference in differentiating among hail sizes, they did examine the distribution of minimum GOES IR Brightness and GOES OT Area (see their Figs. 6a, b, 8a, b), and did not find a strong relationship between those fields and observed hail size.

As indicated above, we do appreciate the uncertainty in relating of maximum reported hail size, average size of hailstones (which may be a better proxy for the damage), and hail metrics based on radar or satellite sources. There is significant need for further research in this area for improved and more reliable modeling, this aspect is stressed in the revised discussion. Further, we have deleted the original Figure 2, but provide another Figure about the relationship in the Appendix.

After the initial submission, we have made several methodology tweaks to the radar and satellite matching to remove view angle dependencies from the geostationary measurements and to ensure highly confident matches by requiring larger cell objects, which better emulates the present study’s methodology. Specifically, ECS-Anvil BTD is normalized by the effective grid resolution degradation compared to nadir and MESH cells under 10 km^2 are excluded. The positively biased MESH95 distribution does not change the strength of the correlation among the GOES parameters compared to the more realistic values of MESH75, which is the primary takeaway from Figure 2. We have noted the positive MESH95 bias in the text to aid the interpretability of the estimated hail size bins and included multiple thresholds of MESH95 (25 mm and 40 mm) in Table 1 to illustrate this bias relative to the ground reports.

Without removing view angle dependence, the prominence of an OT observed at low VZA (e.g., <40° in the Southeastern US) would be greater than had this same OT occurred at high VZA (e.g., >50° in the Northern Plains) due to differences in the effective pixel resolution. We derive the normalization factor based on the effective footprint area of GOES relative to the nadir footprint area (16 km2 for GOES-13) to account for the imager's reduced ability to observe the prominence of a colder pixel relative to the background anvil at higher view angles. The formulae to derive the x and y component of pixel resolution for are shown in the Appendix for GOES-13.
The normalization results in improved correlations between ECS-Anvil BTD and MESH95 for GOES-12/13 and GOES-16 shown in box and whisker plots (Fig. A1). We have summarized this discussion in the paper's appendix section to justify use of a satellite metric of updraft intensity to estimate storm severity.

We feel that Murillo and Homeyer did not find a correlation between satellite parameters and reported hail size because of uncertainties in hail size reporting. Though MESH has uncertainties as well, it is spatially and temporally consistent, unlike reporting and the relationships described above indicating a correlation between storm intensity and hail size are robust.

> *In my opinion, this claim cannot be supported given the current literature, and hail sizes should be removed from the database (or only provided to customers with a strong caution about their use, and not published in the literature).*

It is clear that improved data and methodology may yield more accurate results in terms of estimated hail size distributions, and the ones presented in our model may be proven wrong. Still, our work may serve as a reference to such studies. To account for the large uncertainty, we have rephrased all text mentioning hail size into hail intensity estimate. The uncertainties have always been communicated quite clearly with the model users, and results are generally not used directly for pricing insurance premiums but as a general indicator of risk. There are additional uncertainties in the exposure and vulnerability models that users have to deal with.

> *Given these issues above, I cannot recommend the article for acceptance. I would be happy to review an article focusing on points 1-4 above, after addressing the issues I've described. A companion paper focusing on points 5-6, after points 1-4 are successfully established, would also be interesting. I cannot support an article including point 7 at the current time.*

We encourage the reviewer to revise his/her decision in light of the further explanations and material presented and would welcome her/his continued guidance in the review process.