Author comment 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 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 succint 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 diameters.

> *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 and the respective information will be added to the publication. Table 1 summarizes the results, which we suggest to put in the appendix.

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). This shows that through the use of OTprob > 0.5 to refine detections to those we are most confident in, we lose about 45% of the probably hail-producing storms. In other words, many severe hail storms 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). The fact that there can be some time 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 really prominent 3 min before the time of a report/MESH, but we only have just the one GOES image to match. We see the agreement with microwave may be lower than with MESH or reports because of parallax shifts in the storm positions in microwave data, especially those close to the limb of the overpass. We will put a comment on that in the revised manuscript.

Table 1: Detection counts and fractions of CONUS 2013-2016 GOES-13 derived parameters matched within 28 km^2 and 15 minutes of hail reports with various GOES OT Probability conditions applied.

	Count	Count with otProb	Fraction with otProb	Count with otProb>=50	Fraction with otProb>=50
MESH95 >=40 cm	72048	49866	0.69	27268	0.38
MESH75 >=25 cm	92066	61340	0.67	32222	0.35
SPC Hail >= 25 cm	35049	24478	0.70	14055	0.40
MWR P_hail >=50	1740	892	0.51	535	0.31

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.

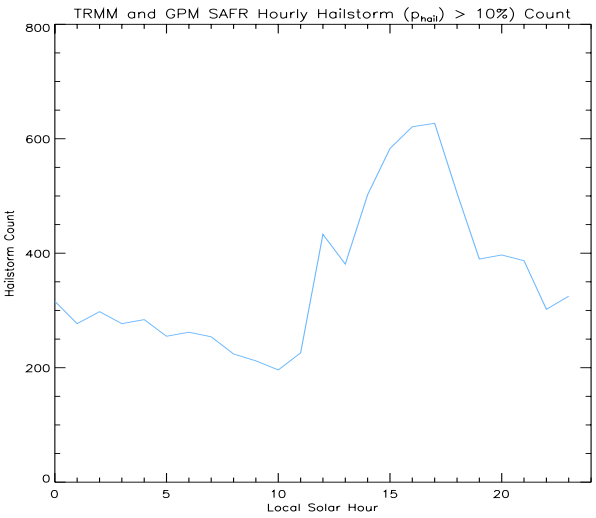
As TRMM and GPM were/are satellites in inclined orbits, their sampling is not continuous, like a geostationary satellite. It is therefore not reasonable to use the absolute counts from these satellites as indication of the true frequency of hailstorms. When assessing the gridded climatologies as in Bang 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 will add to the discussion of differences with other publications on hail in the region.

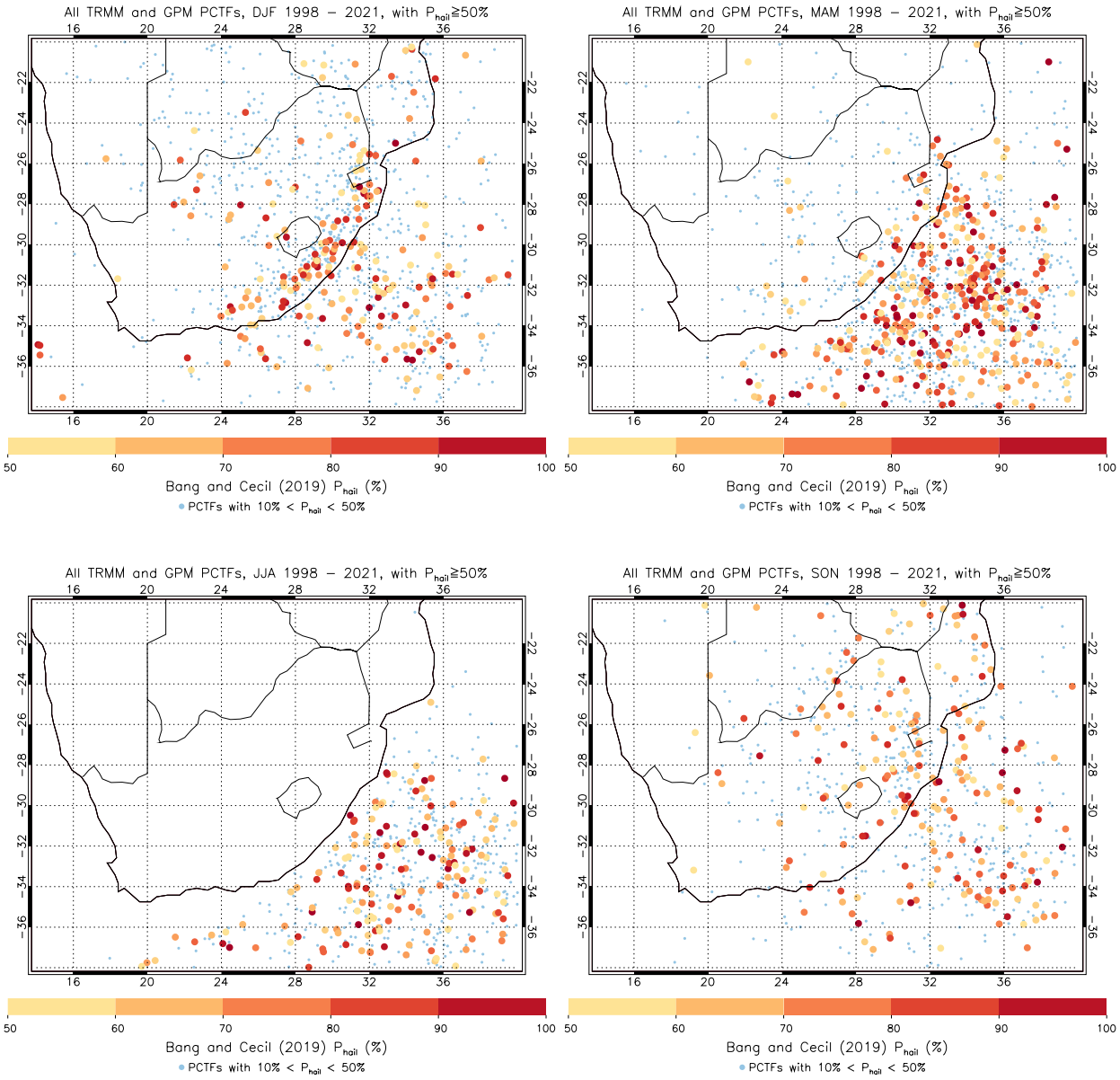
> *3. The hail grouping methodology into events reasonably represents hail swaths from a single storm system. While the description of the methodology (lines 176-177) is intuitive and simple, the results of the grouping methodology in Fig. 7 don't seem to follow that description. Why are there multiple 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 will comment on this in the text and may add an example with radar for the US to the appendix

> *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.*



Below there are the four seasonal frequencies of hailstorms as well as the diurnal cycle as seen by TRMM and GPM combined.
We will provide a direct comparison of these seasonal climatologies to the OT-based estimate 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 sited. 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 will be included. In addition, the presentation of the methodology will be 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 will clarify the use of the OT product as a proxy for hail 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 and the model's event catalogue are drawn from hail size distributions in reports. Relations between event properties are used only for the problem of matching those hail sizes to the other event characteristics in the stochastic events.

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. Still we find and will present in the appendix of the revised work a relation between hail size parameters based on radar – where the spatial match can be extected 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.

We will clarify our statement on that. The performance of the OT algorithm is constantly improving, and recent evaluations against different hail size metrics do show a certain relationship.

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.

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 will rephrase all text mentioning hail size into equivalent hail size 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 the light of the further explanations and material presented and would welcome her/his continued guidance in the review process.