We thank referee#1 for the helpful comments. Point-by-point responses to reviewer's comments are listed below (in blue).

Review of "**Near Global Distributions of Overshooting Tops Derived from Terra and Aqua MODIS Observations**" by Yulan Hong, Robert J. Trapp, Stephen W. Nesbitt and Larry Di Girolam

This is an interesting paper on providing insight about the frequency and size distribution of overshooting cloud tops into the low stratosphere for deep convective cloud systems on a global basis. This study represents the first such study using MODIS IR measurements where both TERRA (descending, ~10:30 am/pm LT equator crossing) and Aqua (ascending, ~13:30 pm/am equator crossing) MODIS measurements were employed. With about 20 years of MODIS daytime and nighttime data from both TERRA and Aqua they are able to derive global seasonal patterns, land-sea contrast, and information on diurnal variability of overshooting cloud tops. The paper is well written describing current and previous algorithms for detecting overshooting cloud tops. I would recommend publication after generally minor changes.

* Figure 1: Please separate (a) and (b) parts of the figure more – in the pdf version the labeling for these two parts overlap.

Thank you for pointing this problem. It is now fixed.

* Line 236: "cirrus anvil"

Corrected. Thank you!

* Table 1: If possible, please include uncertainties for the coefficients in this table. The table caption lists significance at 99% but it is not certain what the 99% is relative to (all coefficients assumed to be zero under the null hypothesis?). The table caption also says regression coefficients which the text indicates are the bi (i=1,2,3) constants, but instead xi (i=1,2,3) numbers are referred to in the table. The xi's if understood correctly represent the three temperature differences (input training data) to derive the bi's via regression. Please help clarify the table and numbers listed for the readers.

We now add the uncertainties for the coefficients in this table. The significant level is under the null hypothesis. Now p values are mentioned.

The table is now adjusted for a better read with explanation below the table. In Lines 308-310 in updated version, we add:

"** The Logistic Regression is based on python module – Statsmodels. The p values for the regressed coefficients are smaller than 0.008 based on z-test. The uncertainties for each coefficient are represented by standard errors.*"

* Figure 4: In the tropics, the tropopause height is consistently about 16-17 km for most any definition of tropopause, dynamical or chemical, but the tropopause height in Figure 4 for the tropics shows about 12 km. Is this because the tropical band in the figure extends to +/- 25 degrees latitude and there is some higher-latitude subtropical influence? Or perhaps the vertical axis numbers are

mislabeled for the tropics? Your tropopause height comes from combining MERRA-2 lapse-rate plus PV surface, so in the tropics it should be coming from lapse rate which is about 16-17 km.

Thank you for pointing out this problem. We recalculated the mean tropopause height. The average is about 15.7 km within -25 – 25 degrees, and about 11.5 km between 25 – 60 degrees S/N latitudes. Since we do not mention the mean tropopause information, we just remove it from Figure 4.

* Line 414: Do you mean "variety" rather than "verity"?

Corrected. Thank you!

* Line 613: "times"

Corrected.

* Line 636: "area" – also in this sentence, could the future work include taking advantage of the long ~20 year records from TERRA/Aqua MODIS IR measurements to investigate climate-related decadal changes/trends in OT events?

Yes, the trend analysis of OT events is now under investigation and will be part of a separate manuscript.

* In the Summary is it possible to expand a bit to describe some general science implications originating from these new results?

In conclusion section (Lines 657-663 of updated version), we add some descriptions of general science implications.

"This study has displayed a comprehensive analysis of OT occurrences near globally for the first time using MODIS data. As MODIS has a fine spatial resolution (1 km) and provides about a two-decade stable climate record, results in this study are an important complement to the current OT climatology in the literature derived from GPM, GOES and AMSU-B (Bedka et al., 2018; Hong et al., 2008; Liu et al., 2020). This study also lays a foundation to understand the near global climatological distributions of hazardous thunderstorms, leading to valuable insights into intense updraft size distributions in deep convection over the globe."

* Do you have an OT dataset derived from this study that can be listed in the "Data availability" section?

We provide two datasets attached as Supplementary 2 and Supplementary 3. The first dataset was used to train and cross validate the Logistic Regression. The second dataset was used to manually validate the Logistic Regression. The datasets include the following information: longitude, latitude, time, OT flag, Tb11, Tb67 and tropopause temperature.

In Lines 672-674, we added:

"The dataset used for training and cross-validating Logistic Regression is available in Supplementary 2.

The dataset used for manually validating Logistic Regression (plot figure 6) is available in Supplementary 3."

We would like to thank referee 2's insightful comments, which greatly help improve this paper. A list of our responses and the marked-up manuscript are given below, highlighted in Blue.

Reviewer's comments on "Near global distributions of overshooting tops derived from Terra and Aqua MODIS observations" by Hong et al.

General comments

This paper used MODIS IR brightness temperatures to identify overshooting tops (OTs) and compiled a two-decade record of OTs. The authors utilized a Logistic Regression algorithm to come up with the OT probability for each MODIS IR pixel. Those with OT probability higher than 0.9 are classified as OTs. The Logistic Regression was trained and validated with about 300 hand-picked cases with the collocated CloudSat vertical profile as the ground truth. Once the OT dataset is created, the authors continue to study season distributions of OT occurrences, their diurnal cycle (based on 4 samples on a 24-hr time scale from Aqua and Terra), and land-ocean contrasts. Results are largely consistent with previous studies.

Using a long data record to compile OTs should contribute to study of the climatology and variabilities of OTs. Initial results all look reasonable. Therefore, the paper is publishable. There are, however, a few concerns which need to be addressed in revision.

First, using only 287 hand-picked cases to build the Logistic Regression algorithm is probably not robust enough. Asking the authors to increase the cases to 2000 or 20,000 would probably be unrealistic since the hand-picking procedure is very time consuming. I'd suggest the authors find another 300 different cases, build another Logistic Regression relationship, and use it to re-process all data. Then, compare some benchmark statistics between the two OT datasets.

Thank you for this question. There is no consensus about how many samples are sufficient to reach a robust performance of a model. Usually, more samples are required for a model when it depends on more predictors (variables). Here in our paper, the Logistic Regression is dependent on three variables. We have 287 cases, whose size is 95 times larger than the number of variables. The sample size is enough for training our model based on papers in the literature that examine the minimum sample size for developing a predicted model (e.g. Riley et al., 2018).

Also, the complexity of the dataset plays a key role in obtaining a robust model as well. We selected the samples over four seasons and in different locations on Earth (Figure 1). This makes the dataset as abundant as possible, which help improving the performance of our model.

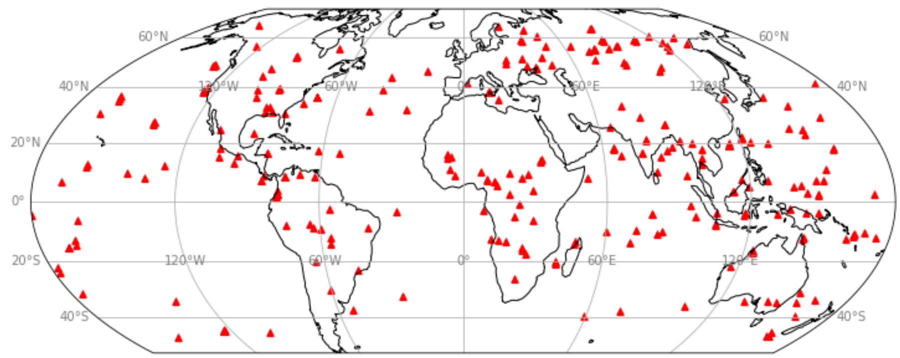

Figure 1. The locations of selected OTs for training the Logistic Regression.

Another way to check if the amount of data is enough is to look at the learning curve, which shows the model performance as function of training sample size (e.g. Beleites et al., 2012). Using our dataset to check the performance of Logistic Regression, Figure 2 displays that our model reaches to a stable and acceptable accuracy (~ 0.84) when sample size is larger than 150. This demonstrates that our sample size near 300 OTs should be sufficient to ensure a robust performance of our model.

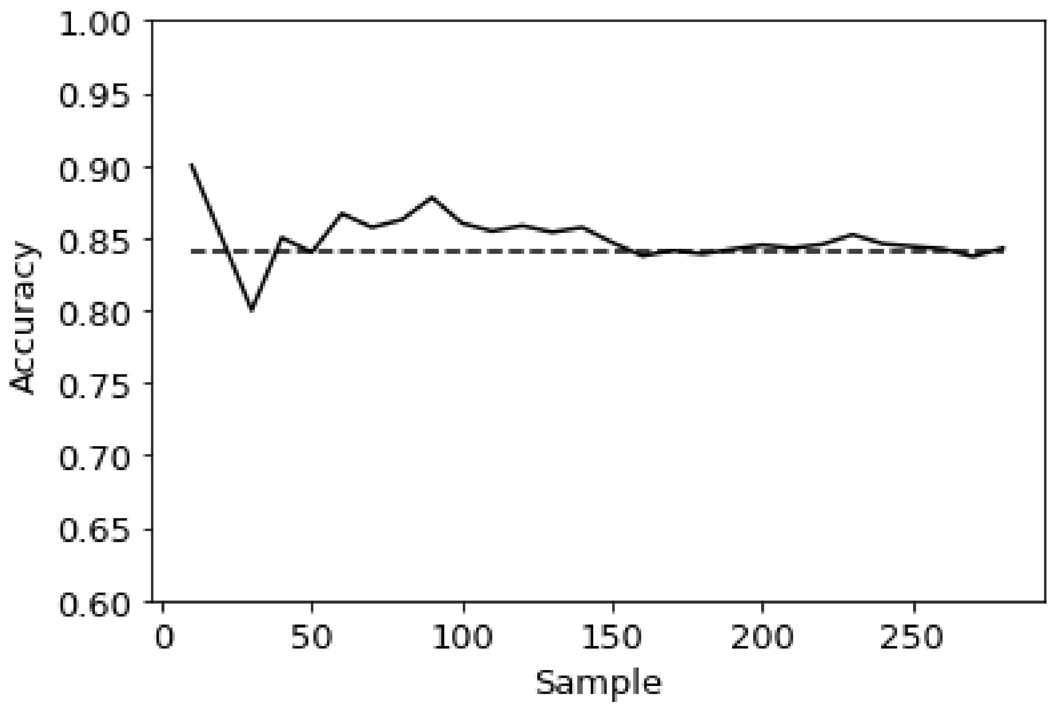

Figure 2. the accuracy of Logistic Regression with training datasets in different sample sizes.

Therefore, we keep our analysis without searching for another 300 OTs for benchmark. we now add more statements to clarify this question in the paper.

In Lines 238-239, we add:

"The total 287 samples were randomly distributed over four seasons and in different locations on Earth. The data are available in Supplementary 2. "

In Lines 303-305, we add:

"*We have tested the accuracy as a function of sample size and found a stable model accuracy of ~84% when total sample size is larger than about 150, indicating that total samples of 287 should be sufficient for training a robust model in our study.* "

Reference

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: Part I–Continuous outcomes. Statistics in medicine, 38(7), 1262-1275.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. Analytica chimica acta, 760, 25-33.

Second, due to parallax shift, MODIS pixels for deep convection are about 5 km away from the CloudSat footprint. See Wang et al. (2011 https://doi.org/10.1029/2011JD016097). The authors cited this paper but didn't really incorporate the parallax correction in their methodology. Mismatch between MODIS pixels and OTs could lead to biases in their Logistic Regression algorithm.

Thank you for this comment. We did check the parallax bias in the colocated Tb from MODIS to CloudSat track using the method from Wang et al. (2011) during the early phase of our analysis. After a careful examination of the parallax bias, we decided not to incorporate the parallax correction for three reasons.

First, the collocated Tb is very similar in deep convective clouds before and after parallax correction (Figure 3), i.e. parallax bias in Tb is relatively small for our study. This is also consistent with the statement in Wang et al. 2011: The effect of parallax correction is small if the cloud object occupies a large enough area and is relatively homogeneous over the range comparable to the parallax correction.

Second, the parallax correction method relies on cloud top height (CTH). Using different CTH product will lead to different collocated Tb. For example, left panel in Figure 3 is collocated Tb derived with the CTH from the 2B-CLDCLASS-LIDAR product, which includes thin cirrus cloud top, and right panel in Figure 3 is the collocated Tb derived using CTH determined by the topmost height of -25 dBZ, which misses the cloud top information of thin cirrus. Differences in the parallax corrected Tb using different CTH products can be observed. This indicates that additional uncertainty in parallax correction induced by biases in CTH retrievals which will propagate to the parallax correction.

Third, there are more bumpiness (spikes) in the parallax-corrected Tb owing to noisiness in the CTH. In our methodology, we have tried to smooth the curve for better deriving the derivatives along the Tb curves. Therefore, the bumpiness induced by parallax correction is not helpful in our algorithm.

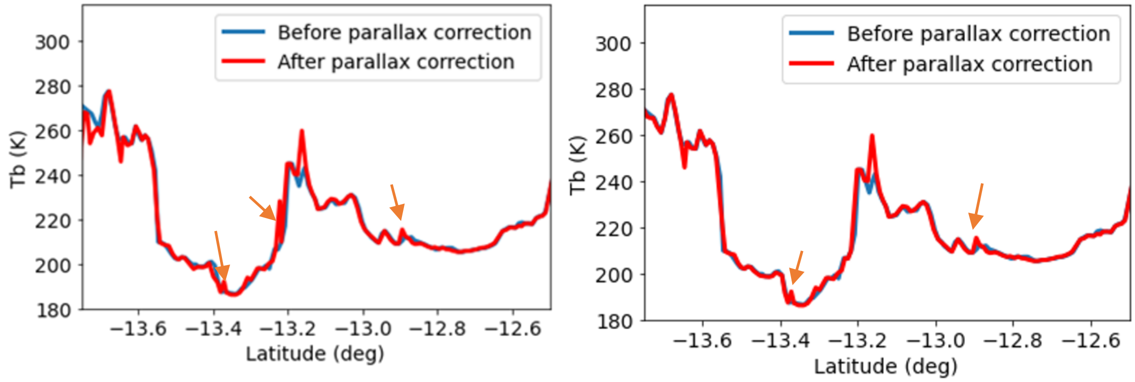

Figure 3 collocated Tb before (blue) and after (red) parallax correction: left using CTH from the 2B-CLDCLASS-LIDAR product and right using topmost height of -25 dBZ as CTH.

Overall, considering that parallax bias is small in deep convective cloud region, accuracy of CTH product affects the accuracy of parallax correction, and parallax correction induces more spikes, we did not apply parallax correction for our algorithm.

In Lines 214-221, we add:

"Parallax correction was examined, but not employed in this work. We found that parallax correction produced nearly identical T_b11 values as those without parallax correction. This is consistent with Wang et al., (2011), who stated that the effect of parallax correction will be small if the cloud object occupies a large enough area and is relatively homogeneous over the range comparable to the parallax correction, which is certainly the case for the OTs being studied here. Additionally, the parallax correction introduces a small amount of noise in the along-track T_b11 values due to noise in the cloud top height used in the parallax correction. The noise is sufficient at times to create artifacts in our OT algorithm."

Overall, I'd like to recommend the paper be accepted after these issues are addressed.

Specific comments:

(Line 112) "...large spatial resolutions": sounds a little awkward. "Coarse spatial resolution" may sound better.

Thank you. Corrected as suggested.

(Lines 224-226) A typical non-OTs (NOTs) example should be shown. "Tb11 is cold": how cold is considered cold enough for NOTs? Any threshold?

Thank you for this question. Manual selection of OT and NOT varies from case to case. However, the Tb of NOT should be close to the tropopause temperature. We now add additional OT and NOT cases in supplementary 1.

Also, to clarify the OT and NOT selection, in Lines 223-224, we add:

"*The OT selection basically followed four principles: T_{b11} colder than T_p, cloud top height above tropopause height, T_{b11} smaller than $T_{b6.7}$, and an obvious convective core* ."

In Lines 234-237, we modify the statement:

"*The NOTs share very similar characteristics with OTs, i.e. T_{b11} is cold (close or colder than T_p) and has a local minimum, but no obvious convective core is observed from the visualization. Supplementary 1 displays four OT and three NOT cases.*"

(Line 250) "Figure 1b indicates the cirrus anvil in cyan": the right-side cirrus anvil ends before Tb11 > 260K. I wonder why it doesn't extend all the way to where Tb11 > 260K, the threshold set for cirrus anvil.

Our algorithm needs the brightness temperature difference between OT and OT's surrounding region. Cirrus anvil in this study is defined as a small area within 20 km from the OT center, which is sufficient to check Tb difference between OT and its surrounding region because OT's size is usually smaller than 15 km. Although Tb11 > 260 K could also have cirrus, but it is too far away from the OT center.

In Lines 268-272, we add:

"Note that cirrus anvil is only defined as a small area within 20 km from the OT center, which is sufficient to check Tb differences between the OT and its surrounding region because the OT's size is usually smaller than 15 km (Bedka and Khlopenkov, 2016). Pixels outside the 20-km radii can also contain cirrus, but do not contribute to our calculation of cirrus T_b."

(Line 264) I am a little puzzled to see that only 16% of NOTs are warmer than the tropopause temperature. For the remaining 84%, the cold centers are colder than the tropopause but no overshooting feature? I guess showing a few examples of NOTs will help.

Yes, we now add three NOTs examples in Supplementary 1. Also, the training dataset includes Tb and tropopause information, which is now included as Supplementary 2.

In Line 280, we add:

"*The selected OTs and NOTs are available in Supplementary 2.*"

(Line 360) "...considering the tropopause height variability": in the presence of overshooting tops, local temperature profile often shows double tropopause: one is associated with the background tropopause and another is caused by the overshoots which are extremely cold. Some examples of the double tropopause can be found near hurricane eyewalls, e.g.,Fig. 2 in https://doi.org/10.1002/2013JD020934 . This kind of overshoot-generated tropopause height variability should be noted.

Thank you for sharing this interesting paper. We add the reference in Lines 379-381:

"Previous studies also adopted a level below the tropopause as the OT reference considering the tropopause height variability (Sun et al., 2019; Zhuge et al., 2015) such as the noted double tropopause observed in deep convection (Vergados et al., 2014)."

(Lines 430-441) OTs associated with shallow convection are interesting. How tall are these OTs and what is the mean tropopause height in this region?

Thank you for this comment. We do not have statistical results for those OTs associated with shallow convection. However, based on the limited cases that we observed, these OTs likely have cloud top height between 5- 8 km. Tropopause height is about 5 – 6 km.

Here is an example of OT associated with shallow convection where tropopause is 1-2 km lower than the tropopause climatology, allowing these shallow convection overshoots the tropopause. There is another case shown by Geerts et al. 2022 (https://journals.ametsoc.org/view/journals/bams/103/5/BAMS-D-21-0044.1.xml), displaying that OTs are below 5 km during the COMBLE Campaign in Arctic Cold-Air Outbreaks (their fig. 6a).

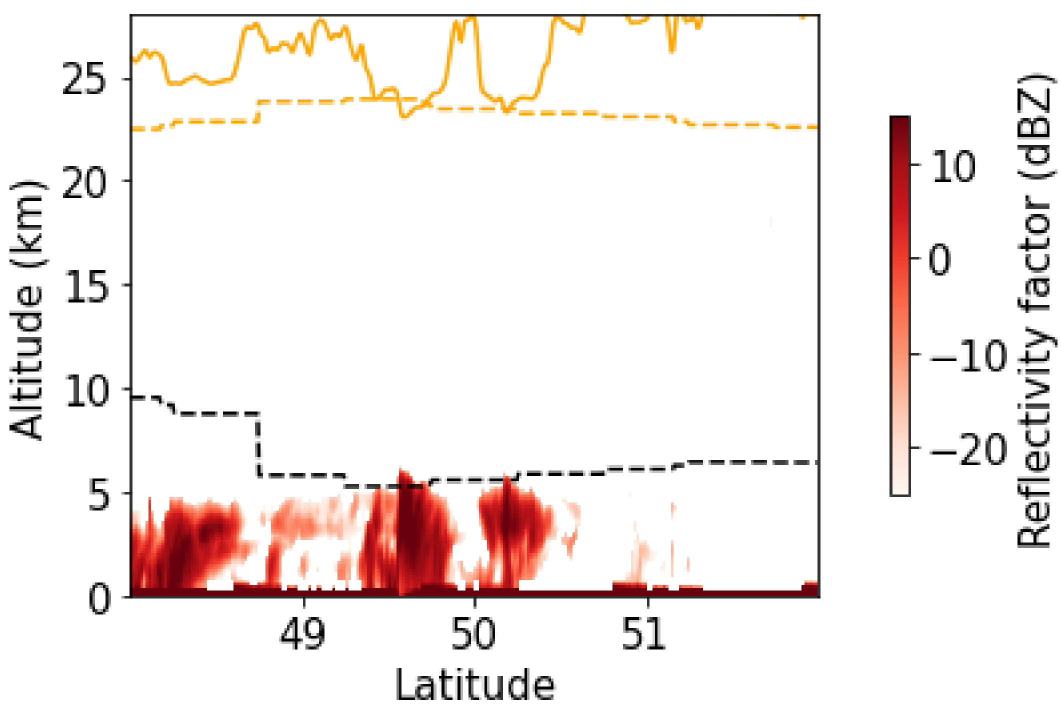

Fig. an overshooting case associated with shallow convection from Jan. 19th, 2007 Over North Atlantic Ocean.

In our paper (Lines 452-457), we also mention that:

"Cold air outbreaks can produce shallow convection when cold air blows from frozen surfaces to warmer ocean. The Cold-Air Outbreaks in the Marine Boundary Layer Experiment

(COMBLE) found that these convective clouds are commonly lower than 5 km associated with updrafts of 4-5 m s^{-1} (Geerts et al., 2022). In the cold air outbreaks, the tropopause is low, which is often at a level below 500 hPa (Papritz et al., 2019; Terpstra et al., 2021), compared to the mid-latitude tropopause climatology of 200-300 hPa (Wilcox et al., 2012)."

(Line 445) "Without no strong convective cores": should be "without strong convective cores".

Corrected. Thank you!

(Line 626) Change "accept" to "except".

Corrected. Thank you!