

# JGR Atmospheres

## RESEARCH ARTICLE

10.1029/2020JD034359

This article is a companion to Cooney et al. (2021), <https://doi.org/10.1029/2020JD034319>.

### Key Points:

- A new OT detection method combines tropopause-relative infrared (IR) temperature, anvil-relative prominence, anvil area and its spatial uniformity
- Overshooting cloud tops (OT) probability derived from spatial cloud analyses is validated with human-identified OTs, revealing improvements over previous methods
- Differing imagery spatial resolution necessitates optimization of sensitivity curves to account for warmer observed OTs from coarser imagery

### Correspondence to:

K. V. Khlopenkov,  
[konstantin.khlopenkov@nasa.gov](mailto:konstantin.khlopenkov@nasa.gov)

### Citation:

Khlopenkov, K. V., Bedka, K. M., Cooney, J. W., & Itterly, K. (2021). Recent advances in detection of overshooting cloud tops from longwave infrared satellite imagery. *Journal of Geophysical Research: Atmospheres*, 126, e2020JD034359. <https://doi.org/10.1029/2020JD034359>

Received 4 DEC 2020

Accepted 13 MAY 2021

### Author Contributions:

**Conceptualization:** Konstantin V. Khlopenkov, Kristopher M. Bedka  
**Data curation:** Kristopher M. Bedka, John W. Cooney  
**Formal analysis:** Konstantin V. Khlopenkov, Kristopher M. Bedka  
**Funding acquisition:** Kristopher M. Bedka  
**Investigation:** Konstantin V. Khlopenkov, Kristopher M. Bedka  
**Methodology:** Konstantin V. Khlopenkov, Kristopher M. Bedka

© 2021. American Geophysical Union. All Rights Reserved. This article has been contributed to by US Government employees and their work is in the public domain in the USA.

## Recent Advances in Detection of Overshooting Cloud Tops From Longwave Infrared Satellite Imagery

Konstantin V. Khlopenkov<sup>1</sup> , Kristopher M. Bedka<sup>2</sup> , John W. Cooney<sup>3</sup> , and Kyle Itterly<sup>1</sup> 

<sup>1</sup>Science Systems and Applications Inc., Hampton, VA, USA, <sup>2</sup>NASA Langley Research Center, Hampton, VA, USA, <sup>3</sup>NASA Postdoctoral Fellowship Program at NASA Langley Research Center, Hampton, VA, USA

**Abstract** This paper describes an updated method for automated detection of overshooting cloud tops (OT) using a combination of spatial infrared (IR) brightness temperature patterns and modeled tropopause temperature. IR temperatures are normalized to the tropopause, which serves as a stable reference that modulates how cold a convective cloud should become within a given region. Anvil clouds are identified using histogram analysis and cold spots embedded within anvils serve as OT candidate regions. OT candidates are then assigned an OT probability, which can be interpreted as a metric of storm intensity and an estimate of confidence in a detection for a particular pixel. It is produced using an original mathematical composition of four factors: Tropopause-normalized temperature, prominence relative to the surrounding anvil, surrounding anvil area, and spatial uniformity of anvil temperature, which are calculated from empirically derived sensitivity curves. The shape of the curves is supported by independent analysis of a large sample of matched IR and radar-derived OT regions. An optimal sensitivity for each factor was determined by maximizing correlation between the OT probability and a set of human-identified OT regions. Coarser spatial resolution of GOES-13 data cause OTs to be less prominent compared to GOES-16, necessitating different sensitivities for each satellite. Detection performance is quantified for each satellite based on human OT identifications and as a function of how prominent the OT appeared in visible and IR imagery. Based on analyses of human-identified OTs, OT detection accuracy, defined by the area under a receiver operating characteristic curve, is determined to be 0.94 for GOES-16 and 0.78 for GOES-13.

## 1. Introduction

Intense convective updrafts often cause cloud tops to penetrate through the surrounding cirrus anvil and into the upper troposphere and lower stratosphere (UTLS). These “overshooting cloud tops” (OTs) and resulting outflow impact UTLS composition (Smith et al., 2017), and storms with intense updrafts frequently produce a variety of severe and aviation weather hazards (Bedka & Khlopenkov, 2016; Reynolds 1980; Yost et al., 2018). Analysis of geostationary (GEO) satellite imagery collected at 30-s to 1-min intervals by Geostationary Operational Environmental Satellites (GOES) shows that deep convection cloud tops and OTs evolve very rapidly, and many OT regions can occur simultaneously (Bedka et al., 2015; Bedka & Khlopenkov, 2016), making it almost impossible for human analysts to quantify trends in this detailed data across many simultaneous storms. Automated OT detection algorithms are therefore required to determine when and where OTs occur, to better understand how convection impacts the UTLS, and to understand processes associated with hazardous weather conditions. GEO imagers generally observe the tops of deep convection, while other remote sensors such as cloud and precipitation radars (Cooney et al., 2021; Takahashi & Luo, 2014; Zisper et al., 2006), passive microwave imagers (Liu et al., 2020), and lightning sensors (Rudlosky et al., 2019) observe processes occurring within convective clouds. Automated OT detection permits accumulation of large samples of co-located remote sensing data to better understand relationships between cloud top and in-cloud processes (e.g., Bluestein et al., 2019).

Simple approaches for deep convection and OT detection, such as single-pixel processing based on infrared (IR) temperature thresholds and/or multi-spectral information (e.g., based on 6–7  $\mu\text{m}$  water vapor absorption or 12  $\mu\text{m}$  channels, or brightness temperature differences (BTD)), are not effective in all cases because 1) a convective pixel may have an identical spectral signature to cold but ordinary cirrus clouds, and 2) pixels with cold IR brightness temperature (BT) are widespread throughout an anvil, but no single BT or BTD threshold is effective for convection detection across a range of latitudes and large-scale environmental

**Project Administration:** Kristopher M. Bedka

**Resources:** Kristopher M. Bedka

**Software:** Konstantin V. Khlopenkov

**Supervision:** Konstantin V.

Khlopenkov, Kristopher M. Bedka

**Validation:** Konstantin V. Khlopenkov,

John W. Cooney, Kyle Itterly

**Visualization:** Kristopher M. Bedka

**Writing – original draft:** Konstantin

V. Khlopenkov, Kristopher M. Bedka

**Writing – review & editing:**

Konstantin V. Khlopenkov, Kristopher

M. Bedka, John W. Cooney, Kyle Itterly

conditions (Bedka et al., 2010). Deep convection is a dynamic process occurring at the mesoscale with characteristic patterns in imagery that require spatial analysis. Human perception, by its nature, is designed to recognize visual patterns within spatially defined objects. With some training, human analysts can easily recognize distinctive features that indicate the presence of convective cells and OTs. To a computer algorithm, however, this presents a particular challenge, as it requires a rigorous formal description of a process before it can be analyzed and quantified digitally.

In recent years, there has been increased interest in the development of more advanced satellite-derived IR-based OT detection algorithms and application of their output for weather and climate analysis. Previous studies have made it clear that OTs, identified via unique texture in visible imagery, have small and distinct IR BT minima with temperatures near to or colder than the tropopause that are embedded within convective anvil cirrus clouds. Key differences in these algorithms are (a) how BT minima (or “cold spots”) are defined, (b) if/how the temperature difference between a cold spot and the surrounding anvil is computed, and (c) the methods used to define a detection and to assign detection confidence. Detection methods have been based on single-pixel processing using multi-spectral temperature thresholding (Mikus & Mahovic 2013; Schmetz et al., 1997), spatial analyses to capture sharp BT gradients typically associated with OTs (Bedka et al., 2010; Bedka & Khlopenkov, 2016), and machine learning (e.g., Berendes et al., 2008; Cintineo et al., 2020; Kim et al., 2017). The GOES-R Aviation Algorithm Working Group (AWG) was tasked with developing an objective OT and enhanced-V signature detection algorithm (Bedka et al., 2010, 2011) for use with GOES-R Advanced Baseline Imager (ABI) data (Schmit et al., 2017). This GOES-R AWG algorithm focused on identification of cold spots colder than the tropopause and used very simple spatial analysis to compute anvil-relative temperature difference. Improvements to this approach were recently developed at NASA Langley Research Center (LaRC) based on product feedback from the operational forecasting community and researchers, where it was suggested that too many true OTs were missed due to strict detection criteria, and a simple yes/no binary detection mask was undesirable. These updated algorithms use more advanced spatial analysis to detect and characterize embedded BT minima and adjacent anvil clouds (Bedka & Khlopenkov, 2016). The term “characterize” refers to how cold and prominent a BT minimum is and how likely it is to be an OT. Such characterizations enable an improved capability to filter detections to include only those that are most intense and/or confident for weather and climate studies, which is an improvement over a yes/no binary OT product (e.g., Apke et al., 2018; Clapp et al., 2019).

Despite recent improvements and science applications with the Bedka and Khlopenkov (2016) method (e.g., Apke et al., 2018; Sandmael et al., 2019), the launch of GOES-16 and -17 satellites and application of this method to 1-min ABI data (Schmit et al., 2017) revealed some inconsistencies in detection performance from image to image, suggesting that further advancements were needed. This paper provides an enhanced technical description and describes recent advancements to IR-based deep convection and OT detection algorithms since the version described by Bedka and Khlopenkov (2016). We have sought to (a) make all BT-based factors and thresholds operate relative to the local tropopause temperature, (b) improve detection of anvil clouds, (c) improve the temporal stability of OT detection when dealing with 30-s to 1-min super rapid scan observations from the GOES-16/17 Advanced Baseline Imager, (d) improve detection consistency across historical and current imagers, and (e) improve the overall quality of OT detection through statistical analysis and validation against gridded U.S. Next-Generation Radar (NEXRAD) network data (see Homeyer & Bowman, 2017 and references therein). This paper is the first of a two-part paper, focusing on algorithm description and quantitative analyses of detection products relative to human-identified OT regions, while a companion paper, Cooney et al. (2021), focuses on quantitative comparisons of GOES-13/16 detection products with NEXRAD GridRad OT detections.

## 2. Satellite Data

The input required by the NASA LaRC OT algorithms is comprised of images from a visible band with  $\sim 0.65 \mu\text{m}$  central wavelength and an infrared (IR) window band with varied central wavelengths from 10.3 to 11.2  $\mu\text{m}$ , satellite dependent. This paper will only describe IR-based algorithms, while the brief description of visible processing here will be elaborated upon in a future paper. The original geolocation (longitude and latitude) data are also required for each satellite pixel and scan line. This imagery is acquired using the Man computer Interactive Data Access System (McIDAS) software (Lazzara et al., 1999) in this software's

proprietary “AREA format,” but in practice these algorithms can be applied to data in other formats, provided the required VIS, IR, and geolocation information is available.

As the OT detection is not based on per-pixel analysis but uses spatial context and patterns, it is important that the cloud features preserve their natural shape and appearance. For this reason, the satellite-viewed images have to be reprojected to any conformal (i.e., preserving object shapes rather than scales) map projection. We selected a simple equirectangular projection, which uses equally spaced latitude and longitude increments and is quite conformal at low- and mid-latitudes where most convective storms occur. The equirectangular grid also simplifies any subsequent data aggregation and cross matching against other meteorological products. Details on the reprojection and processing of data near the edges of valid satellite data are provided in Appendix A1.

The output spatial resolution of the reprojected data is selected to approximately match the nominal resolution of the input imagery, and the output resolution for VIS data is always made four times higher than for IR data, in order to streamline the subsequent processing independent of the input data source. For example, when processing data from the GOES-16/17 ABI or Himawari-8/9 Advanced Himawari Imager (AHI), which collect VIS data at 0.5 km/pixel and IR data at 2.0 km/pixel at their sub-satellite coordinate, the output resolution for VIS data is 224 pixel/degree (which is about 0.5 km/pixel) and so for IR data it is 56 pixel/degree. For other satellites, the output resolution is selected to be two times lower. The input imagery from the Meteosat Second Generation Spinning Enhanced Visible and Infrared Imager (SEVIRI) instrument, which provides IR data at 3.0 km/pixel, is also reprojected to 112 pixel/degree for VIS data and 28 pixel/degree for IR data.

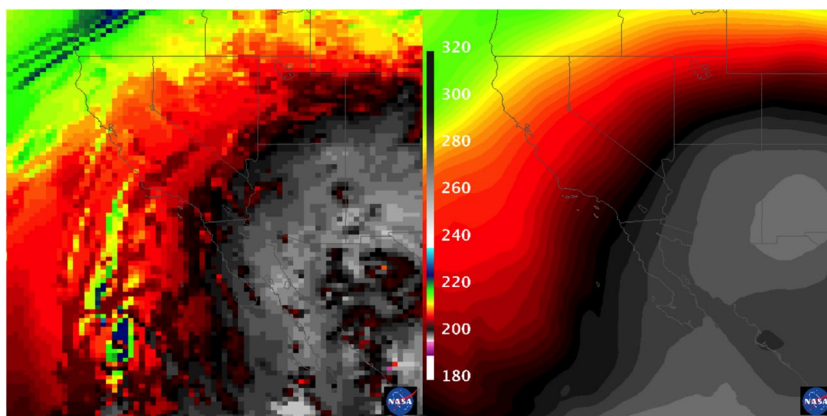
### 3. Overshooting Cloud Top Detection Processing

A set of processing steps are used to identify OTs in infrared imagery. IR BT is first converted to a tropopause-relative temperature. Anvil clouds are identified, and then cold spots embedded within the anvils are detected. A set of spatial analyses are performed in the anvil regions surrounding cold spots to quantify the tropopause-relative IR BT, prominence of the cold spot relative the surrounding anvil, anvil area, and anvil temperature spatial coherence, which are combined to derive the likelihood that a cold spot is an OT. The following sections detail these processing steps.

#### 3.1. IR Image Normalization by Tropopause Temperature

Anvil clouds are typically located somewhere in the altitude range between the level of neutral buoyancy (LNB, also commonly referred to as “equilibrium level”) and the tropopause, though some anvil clouds are colder than the tropopause (Bedka & Khlopenkov, 2016). The LNB is a challenging parameter to calculate accurately because it depends on temperature and moisture profiles in the planetary boundary layer, which are not well known, especially over data poor regions. Though Bedka and Khlopenkov (2016) show that anvil cloud temperature is correlated with the LNB, spatial inconsistencies or inaccuracies in LNB data, when compared with observed IR temperature patterns, could yield misleading analyses of storm intensity and OT detection error.

The tropopause temperature can be estimated more easily across the globe because the UTLS temperature is better observed by satellites, radiosondes, and commercial aircraft, and therefore provides a more reliable reference to identify anvil clouds and embedded OT regions based upon their IR temperature. Solomon et al. (2016) and Xian and Homeyer (2019) show that reanalyses such as ERA-Interim, JRA-55, MERRA-2, and CFSR identify tropopause altitudes with small bias (typically less than  $\pm 150$  m) and error comparable to the model vertical resolution when compared with tropopause altitudes derived from radiosondes. Storms in the mid-latitudes, where the tropopause is lower and warmer, will have warmer cloud top temperatures than those in the tropics. For example, an IR temperature of 215 K may indicate a very severe storm in northern mid-latitudes but would correspond to an unremarkable storm in the tropics. Therefore, to enable accurate convection detection globally, IR BT is converted to a tropopause-relative temperature using the hourly Modern-Era Retrospective analysis for Research and Applications Version 2 (Bosilovich et al., 2016) tropopause analysis, contained in the 2days, 1-Hourly, Time-Averaged, Single-Level, Assimilation, Single-Level Diagnostics V5.12.4 (“MERRA2\_400\_tavg1\_2d\_slv\_Nx”) collection. MERRA-2 data,



**Figure 1.** An example of the original (left) and filtered (right) GEOS-5 tropopause temperature over the Southwest U.S., Mexico, and Eastern Pacific Ocean.

originally at  $0.5^\circ \times 0.625^\circ$  spatial resolution, are interpolated spatially to our equirectangular grid (described in Section 2) using Lanczos filtering, and also in time using linear interpolation between the two nearest available time samples. The MERRA-2 “TROPT” parameters is used in this analysis that represents a blended estimate of the tropopause temperature based on a combination of the WMO definition of the primary lapse-rate tropopause (WMO, 1957) and equivalent potential vorticity. Such model analyses will be referred to here generically as numerical weather prediction (NWP) fields. In practice, these methods could be applied to any NWP tropopause field.

NWP tropopause temperature fields can depict spatial variations that may not be realistic relative to the actual conditions around ongoing storms. We feel that this is primarily due to the limited vertical resolution of the analyses compared to radiosondes. At one location the tropopause detection algorithm may successfully identify the primary tropopause, while at a nearby location it may fail to detect the primary tropopause, which causes a higher level to be incorrectly labeled as the primary tropopause. This leads to substantial overestimates of the tropopause altitude. This is most common near jet streams, where there is commonly a sharp jump in the tropopause altitude. In addition, the tropopause has details that, when compared to satellite IR BT, would result in a noisy product not amenable to storm detection and characterization. An example of this can be seen in Figure 1 where the left panel shows the tropopause temperature obtained from GEOS-5, which exhibits extensive high frequency spatial variability.

In order to obtain a reliable and stable reference temperature threshold, the following spatial filtering is applied to the tropopause temperature,  $T_{tp}$ . Within a circle of 500 km diameter around the current pixel (which is allowed to be clipped at the boundary of the user-selected domain), the area mean  $\overline{T_{tp}}$  and standard deviation  $\langle T_{tp} \rangle$  of the tropopause temperature are calculated and the filtered temperature is obtained as:

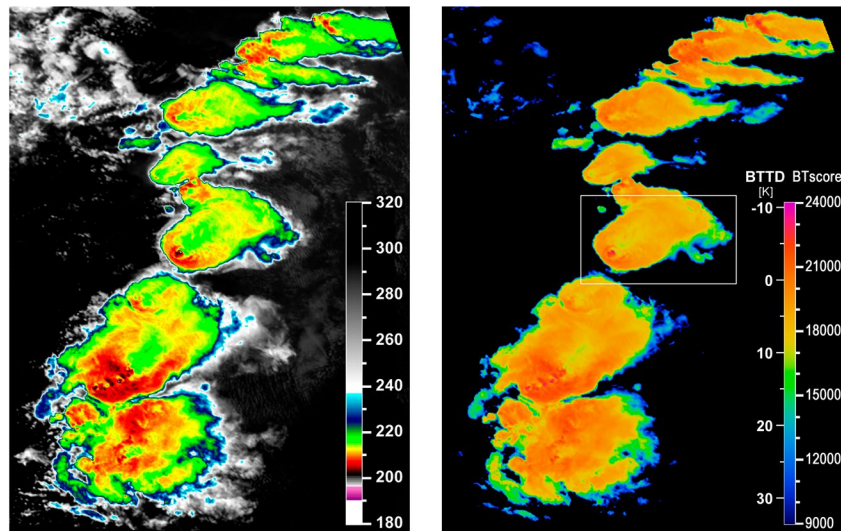
$$T_{tp\text{filtered}} = \overline{T_{tp}} - 0.6 \langle T_{tp} \rangle \quad (1)$$

This filtering produces a much smoother tropopause temperature field, which is designed to be biased slightly colder than the original tropopause in areas of strong spatial gradients. Such biasing accounts for situations where the updraft region of a storm may be rooted on the colder side of a gradient, but the anvil extends into the warm side. A discontinuity in the anvil and OT detection would occur here if the sharp tropopause gradient were preserved, so the right-side term of this equation extends cold temperatures to the warm side of a tropopause gradient. An example of a filtered tropopause analysis is shown in the right panel of Figure 1, which is much more amenable to IR-based convective storm detection and analysis.

IR temperature normalization by the tropopause is accomplished through a parameter referred to as the “BT-score.” It is obtained from the filtered tropopause temperature as follows:

$$BT \text{ score} = (60 - BTDD) \cdot 340, BTDD = BT_p - T_{tp\text{filtered}} \quad (2)$$





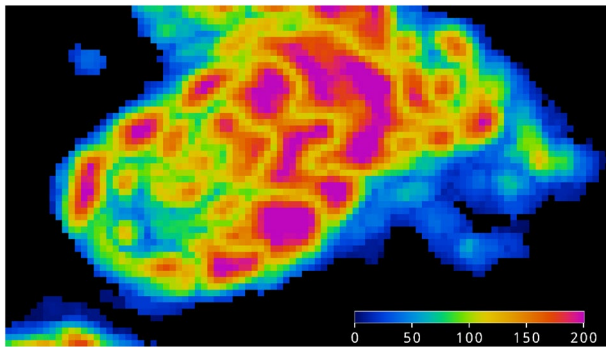
**Figure 2.** GOES-16 10.3  $\mu\text{m}$  brightness temperature on May 5, 2019 at 23:00 UTC across the U.S. Southern Plains (left) and the corresponding image of BT-score (right). A scale of brightness temperature (BT) minus tropopause difference (BTDD) is also provided for reference.

where  $BTDD$  is BT minus Tropopause temperature difference,  $BT_p$  is the brightness temperature of the current pixel, and the scale (340) and the offset (60) are designed to fit the resulting BT-score to the range of a 2-byte integer. Thus, extremely strong OT events of 20 K colder than the tropopause would yield a BT-score of 27,200. This BT-score parameter is similar to S-score used by Bedka and Khlopenkov (2016), which relied on a difference of the local temperature from a regional mean temperature and a difference from a reference IR temperature of 255 K. The new BT-score, however, presents an improvement over the S-score in that it allows all subsequent analysis to be independent of absolute temperature thresholds by instead making it relative to the local tropopause temperature. Figure 2 (left panel) shows a GOES-16 10.3  $\mu\text{m}$  BT scene with numerous OT-producing and severe storms across the U.S. southern Great Plains. The corresponding BT-score image is shown in the right panel of Figure 2. The rectangular frame highlights a strong supercell storm that will be used to demonstrate subsequent stages of the detection process.

### 3.2. IR Anvil Mask

As shown in Figure 2, anvil clouds appear as a spatially continuous area of cold pixels. By quantifying the “coldness” and spatial continuity it is possible to derive an anvil mask. The anvil mask is a 1-byte rating that indicates a confidence in anvil detection, with values over 10 roughly corresponding to human perception of anvil cloud extent while values over 100 indicate a very high level of confidence. This anvil rating is crucial for successful OT detection as it is used during several processing phases described below. The new anvil mask algorithm has been improved significantly over the version from Bedka and Khlopenkov (2016), which was based on 16-vertex polygons constructed around each OT candidate. That method required a distinct OT to be present for an anvil to be detected. Also, it was not always possible to capture curved anvil boundaries using simple polygon shapes. In this new method, calculation of the anvil rating/mask can be roughly described in the following three steps: Histogram analysis, expansion to encompass the anvil spatial extent, and final refinements.

Here and further in the description of the algorithm, we present several key algorithm parameters, which may appear arbitrary/empirical but, in fact, have been derived after years of extensive tests on large volumes of real satellite data and have been found to work best in various convection scenarios. Some specific values are helpful in improving the computational performance, which is especially critical in large volume data processing. The current version of the algorithm can process a typical GOES-16 full disk scene in 15–30 s when executed on a typical 3 GHz CPU in a single thread mode.



**Figure 3.** Initial anvil rating based on the dominant peak on the brightness temperature (BT)-score histogram calculated for GOES-16 case of May 5, 2019 23:00 UTC. The image boundaries correspond to the rectangular region in Figure 2 right panel.

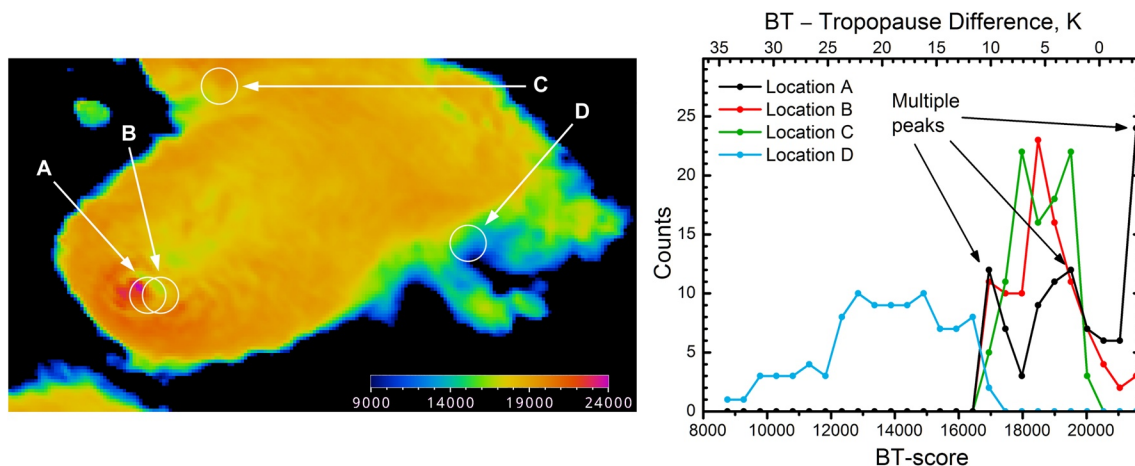
The input BT-score image is processed in subsets taken by a circular 22 km diameter window at every other pixel and every other line. The local distribution of BT-score within each subset is analyzed by constructing a histogram  $H$  having  $N = 32$  bins, covering the BT-score range of 8,500–24,884. A value of 8,500 corresponds to clouds at a level of 35 K warmer than the tropopause, which is an extremely low tropopause-relative bound for anvil clouds. The upper limit of 24,884 indicates a level 13 K colder than the tropopause, which only occurs in updraft regions or adjacent anvil outflow. Pixels colder than this threshold are accumulated in the last bin.

As noted previously, BTs within anvils are, in most cases, cold and spatially uniform, which would yield a sharply peaked histogram of BT-score. Thus, the anvil rating should be made proportional to the peak's height ( $H_i$ , which is the number of counts in the  $i$ -th bin) and to that bin's index  $i$ , because a higher bin corresponds to a colder region. After extensive testing, Equation 3 is found to be able to describe the dependence of anvil rating,  $r_{\text{anvil}}$ , on index  $i$  reasonably well:

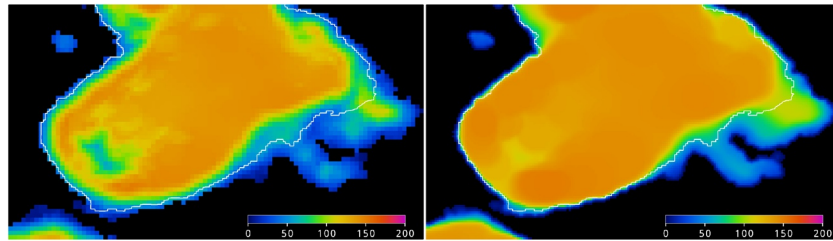
$$r_{\text{anvil}} = C_H \cdot H_i \cdot i \cdot (2N + 8 - i) \quad (3)$$

Here,  $C_H$  is a normalization coefficient and the term in parentheses flattens the dependence out at high values of  $i$  in order to simulate the natural saturation of confidence (similar to any probability in general) in anvil detection when the anvil BT is extremely low. Figure 3 shows  $r_{\text{anvil}}$  for the outlined region in the middle of Figure 2 (right). Here,  $C_H$  is equal  $0.35/D^2$  where  $D$  is the diameter of the histogram window in pixels. This normalizes the result by the total number of pixels in the histogram to account for varying satellite pixel size.

One can immediately see in Figure 3 that the obtained anvil rating field is very non-uniform and consists of large spots of higher rating separated by smaller, low rating areas. The main anvil area has the rating ranging from about 50 to 200 units. The lower rating is produced not only outside the main anvil but also very close to the primary OT area in the lower left part of the anvil (see Figure 4 (left) for a BT-score image corresponding to the same region). The OT and nearby vicinity exhibit a highly non-uniform BT field, which translates into a broad histogram and a smaller dominant peak resulting in a lower anvil rating calculated from Equation 3. This becomes a problem for all subsequent OT detection steps because the OT area itself may happen to be excluded from or penalized in further processing due to the low anvil rating.



**Figure 4.** A subset of the BT-score image (left) with white circles showing four locations where histogram's peaks are analyzed. BT-score histogram (right) calculated at four selected location: Around the main Overshooting cloud tops (OT) (locations A and B), at the anvil's upper extent (location C), and at the anvil's boundary (location D). Brightness temperature (BT) minus tropopause temperature difference scale is provided at the top axis for reference.



**Figure 5.** Preliminary anvil rating (left) obtained as a summed contribution from three highest peaks in BT-score histogram. The white contour selects the main anvil cloud at BT-score level of 16,000. Final anvil rating (right) after spatial expansion and refinement.

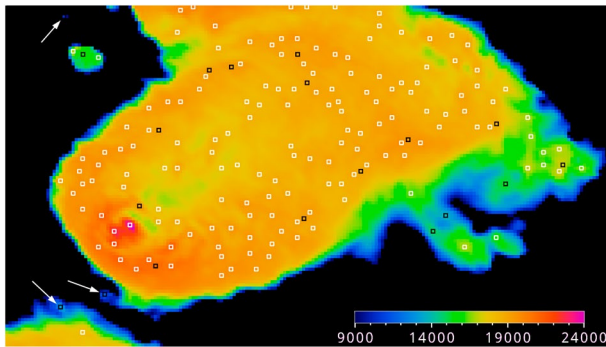
To highlight this problem, examples of the obtained histogram are analyzed at four selected locations shown with white circles in Figure 4 (left). The circle diameters represent the actual size of the window where the histogram is probed. Although in the middle of the anvil the histogram's peak reaches well over 30 counts, it falls below 25 in selected anvil locations A, B, and C as demonstrated by Figure 4 (right). Location A includes many extremely cold pixels from the OT core that are counted in the last bin of the histogram making it the highest peak in this case (black curve). The histogram also reveals two minor peaks at around 17,000 and 19,500 highlighted by the arrows in the figure. On the other hand, location B (which is just 8 km eastward from A) produces a completely different picture. The peak in the last bin disappears and the other two peaks become merged into a single peak at 18,500 (red curve). Location C, though located in a seemingly uniform anvil region far from the OT area, still develops two distinct peaks of the same height (green curve). Lastly, location D at the edge of the anvil produces a broad histogram distribution that is shifted much more towards lower BT-scores. The latter logically yields a low anvil rating because of the smaller peak and the low  $i$ -index. The first three locations, however, can produce a completely volatile result because the height and position of the peak irregularly fluctuate.

The problem of high volatility in the anvil rating can be solved by combining the contributions from several peaks:

$$r_{\text{anvil}} = C_H \sum_i H_i i (2N + 8 - i) \quad (4)$$

Here the summation is carried out over the three highest histogram bins and the bins are allowed to be adjacent. This allows a single large peak to be counted together with its neighboring bins, and if the peak splits into two or more separated bins (as is the case with location C above), the total contribution remains comparable, which should help the overall stability of the resulting rating across the entire anvil region. This is confirmed by Figure 5 (left) which depicts the preliminary anvil rating obtained from Equation 4. The normalization coefficient  $C_H$  used here is smaller,  $0.22/D^2$ , to compensate for the summation of several bins. The primary difference from Figure 3 is that the anvil rating field is now much more homogeneous within most of the anvil area, except for the main OT and downwind from the OT where warm temperatures are present due to an above anvil cirrus plume. The white contour in this figure shows the extent of the main anvil using a BT-score threshold of 16,000, or 13 K warmer than the tropopause. One can see that the uniform orange region nearly reaches the anvil boundary outlined in white. There is, however, an apparent margin along the contour where the anvil rating drops significantly even though the BT-score level is high. This is due to how the histogram is derived: When the circular window reaches the anvil boundary, the number of cold pixels within the window declines, thereby reducing the contribution to the major cold peaks. This reduction is further corrected by additional processing steps that involve a spatial expansion and refinement of the anvil rating, as described in Appendix A2.

The outcome of this refinement is the final anvil rating shown in Figure 5 (right), which demonstrates a very much improved depiction of the anvil extent and a more uniform anvil rating field. The OT vicinity is now covered by a higher anvil rating and the transition of the rating from a high level to background is now smooth at the anvil boundaries. This is important for cases when OT centers are situated at the boundary, due to strong advection of anvil away from the OT. Validation of the IR anvil mask is described by Scarino et al. (2020) where it was demonstrated to have accuracy comparable to several other anvil identification



**Figure 6.** White squares denote the refined list of brightness temperature (BT) minima after optimization, overlaid on the BT-score image. Discarded candidates are shown with black squares and the arrows point to a few less noticeable spots.

approaches. Scarino et al. (2020) noted that while true anvil clouds are identified well with this method, other spatially coherent areas of cold cloud with temperatures near to or colder than the tropopause could also be identified. For example, non-convective cold cloud shields along fronts within mid-latitude cyclones could be mis-identified as anvil, so Scarino et al. (2020) propose use of environmental parameters such as Convective Available Potential Energy (CAPE) to filter out false anvil detection where deep convection is extremely unlikely. For the purpose of OT detection, anvil cloud identification simply provides a search region for identification of OT candidates. Additional algorithm logic would mitigate false OT detection in such frontal cloud patterns.

### 3.3. Identification of Initial OT Candidates

OTs appear as localized cold spots with diameters typically less than 15 km embedded within a surrounding anvil cloud (Bedka et al., 2010).

Therefore, the starting phase for OT detection involves a search to identify cold spots followed by a series of distance tests. This is to ensure that detected OT candidates have appropriate spacing and thus two pixels from the same OT are not classified as two distinct OT areas. The required spacing gradually reduces for candidates with higher BT-score. This allows colder OT cores, more likely to truly be OTs, to form a denser distribution, thus increasing their chances to pass all the detection tests.

Unlike the method used by Bedka and Khlopenkov (2016), the current algorithm identifying local maxima in a pixel image is an original one, developed specifically for OT candidate selection. It is also quite general, so it can be used for any other image data, such as visible reflectance, temperature, and others. The process begins with analyzing  $3 \times 3$  pixel subsets of the BT-score image in order to identify pixels surrounded by all lower scored nearest neighbor pixels, which guarantees that the pixel is a true local maximum. As a result, the initial local maxima are spaced apart by at least two pixels. Then, for each identified maximum, its  $11 \times 11$  pixel proximity is checked for the presence of an even higher maximum. If a better candidate is found in this proximity, then the current candidate is discarded. A “better candidate” refers to a higher nearby maximum that was not yet discarded. Thus, this process continues recursively and thins out the initial list of candidates leaving the most prominent local maxima in the BT-score field. This proximity check uses an effective distance  $D_{\text{eff}}$ , and if the geometrical pixel distance between the two candidates is less than  $D_{\text{eff}}$ , then the weaker one is discarded.  $D_{\text{eff}}$  is based on the minimal desired distance  $L$  (its default value is 4 km), which is corrected with two factors:

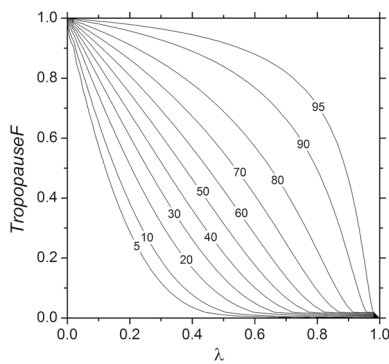
$$D_{\text{eff}} = L \cdot \left( 1 + Z \left( 10 \sqrt{\frac{|A - B|}{A + B}} - 1 \right) + Z \left( \frac{17000 - \min(A, B)}{170} \right) \right) \quad (5)$$

where  $A$  and  $B$  are the two BT-scores of the candidates in the pair,  $Z(x)$  equals  $x$  for positive arguments or zero otherwise (i.e., the ramp function), and the other constants here are derived from empirical testing. The first term in the outer brackets increases the effective distance if the candidates' BT-scores are too different, meaning that a weaker candidate can survive only if the stronger one is further away than the default distance  $L$ . The second term increases the effective distance if either of the candidates is too weak, which allows for a more dense population of colder candidates and aids detection of smaller OTs inside a large cluster of cold pixels. The result of this optimization is demonstrated in Figure 6, where the discarded candidates are shown with black squares.

### 3.4. Calculation of OT Probability

Ideal OT areas appear “cold” based on their tropopause-relative IR temperature and comparisons with the surrounding anvil, and are located within broad and cold, spatially uniform anvils. Such a conventional description can be quantified by means of an IR OT rating, which has previously been obtained as the sum





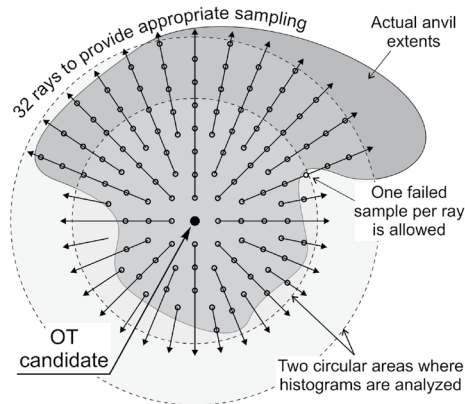
**Figure 7.** Infrared (IR) Overshooting cloud tops (OT) probability (shown in numbered contours) as a function of the tropopause factor and other factors combined in  $\lambda$ .

factors together helps the overall robustness of the detection scheme and ensures that a high OT probability corresponds to a true OT with strong convection. Contributions from these four factors have to be combined in such a way to properly account for the role of each contributor. For example, although the *TropopauseF* is the most significant factor among the four, it alone cannot define the resulting OT probability because an OT candidate pixel may be very cold relative to the tropopause but can, in fact, be just a small perturbation in a broad uniformly cold anvil cloud (Cooney et al., 2021). Therefore, the IR OT probability *OTprob* is constructed as follows:

$$OTprob = 100 TropopauseF^{0.6(1/\lambda-1)} \quad (6)$$

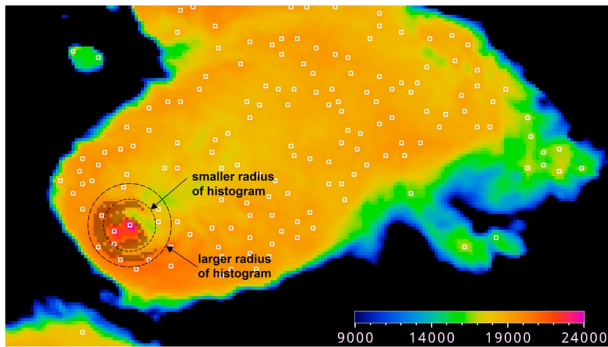
where 100 is a scale factor that scales the resulting OT probability into the range of 50–100 for the strongest OT detections.  $\lambda$  is a function proportional to the remaining three factors (more details are provided below) and ranges from 0 to 1. The main factor *TropopauseF* here is raised to the power of  $1/\lambda-1$ , which makes the result close to zero when  $\lambda \rightarrow 0$  and close to 1 when  $\lambda \rightarrow 1$ . The power of 0.6 is needed here to control the growth rate of *OTprob* with the four factors and was found to achieve the best agreement in validation (described in the next section). Overall, this formula makes the *OTprob* grow with the *TropopauseF* while the growth rate is suppressed, if the other factors are insignificant, or inflated when  $\lambda$  is high (Figure 7).

When all of the factors comprising  $\lambda$  are high (e.g.,  $\lambda > 0.7$ ) the resulting *OTprob* reaches 70 and higher even at low level of 0.3 of *TropopauseF*. For a low  $\lambda$  of 0.2, the same *OTprob* requires a high *TropopauseF* of 0.9. On the other hand, embedded cold spots in broken cirrus can achieve anomalously high temperature differences between an OT candidate and the surrounding warmer semi-transparent cloud, though this scenario is uncommon because broken cirrus is rarely identified as anvil. Thus, it is important to reduce the influence of *ProminenceF* at warmer temperatures, and therefore a low *TropopauseF* in Equation 6 causes *OTprob* to remain small for virtually any  $\lambda$ , as  $\lambda$  never reaches 1.0 in reality.



**Figure 8.** A concept drawing showing two circular areas where the histograms are probed and the sampling pattern of the nearby anvil to derive anvil mean parameters: Temperature, rating, and effective area.

The three factors comprising  $\lambda$  are derived from an anvil region in the OT candidate vicinity, so we first need to describe how that area is evaluated. The goal here is to calculate the anvil mean parameters averaged over a set of pixels that represent the most homogeneous area around a candidate. In the first step, two histograms of brightness temperature are computed within circles of radius,  $R_H$ , of 16 and 24 km around a candidate (see Figure 8), which allows for adequate capturing of both small- and large-scale anvils. A small area around the candidate is excluded here, which is taken as  $3 \times 3$  pixels in the center (for 2 km or finer spatial resolution) or just the central pixel with its four immediate neighbors (for



**Figure 9.** Extents of the anvil cloud (dark transparent shading) derived for the strongest OT, overlaid on the brightness temperature (BT) score image for GOES-16 case of May 5, 2019 23:00 UTC.

coarser resolution). The histograms have 40 bins covering the temperature range from  $BT_p$  (the central pixel's BT) to  $BT_p + 25K$ . Within each histogram, the two highest bins are selected, and their peak positions are calculated as:

$$X_{\text{peak}} = \frac{\sum_{n=i-1}^{i+1} n H_n}{\sum_{n=i-1}^{i+1} H_n} \quad (7)$$

where  $i$  is the index of the selected bin. This is converted to the corresponding temperature  $BT_{\text{peak}}$ , which gives us four temperature values from the two histograms.

In the second step, 32 rays are cast off from the center with their initial offset from the center varying in the pattern depicted by Figure 8. The offset pattern is needed to reduce spatial oversampling around the center and to achieve a reasonably uniform sampling coverage in the OT candidate vicinity. These offsets can be obtained by taking each ray's ordinal

number in binary form (e.g., 0000, 0001, 0010, 0011, 0100, etc.) and using the number of trailing zero bits  $z$  (4, 0, 1, 0, 2, respectively) to scale down the number 8 (the pattern period) by means of the bitwise 'shift-right' operation as  $8 \gg z$  (to get 0, 8, 4, 8, 2, respectively).

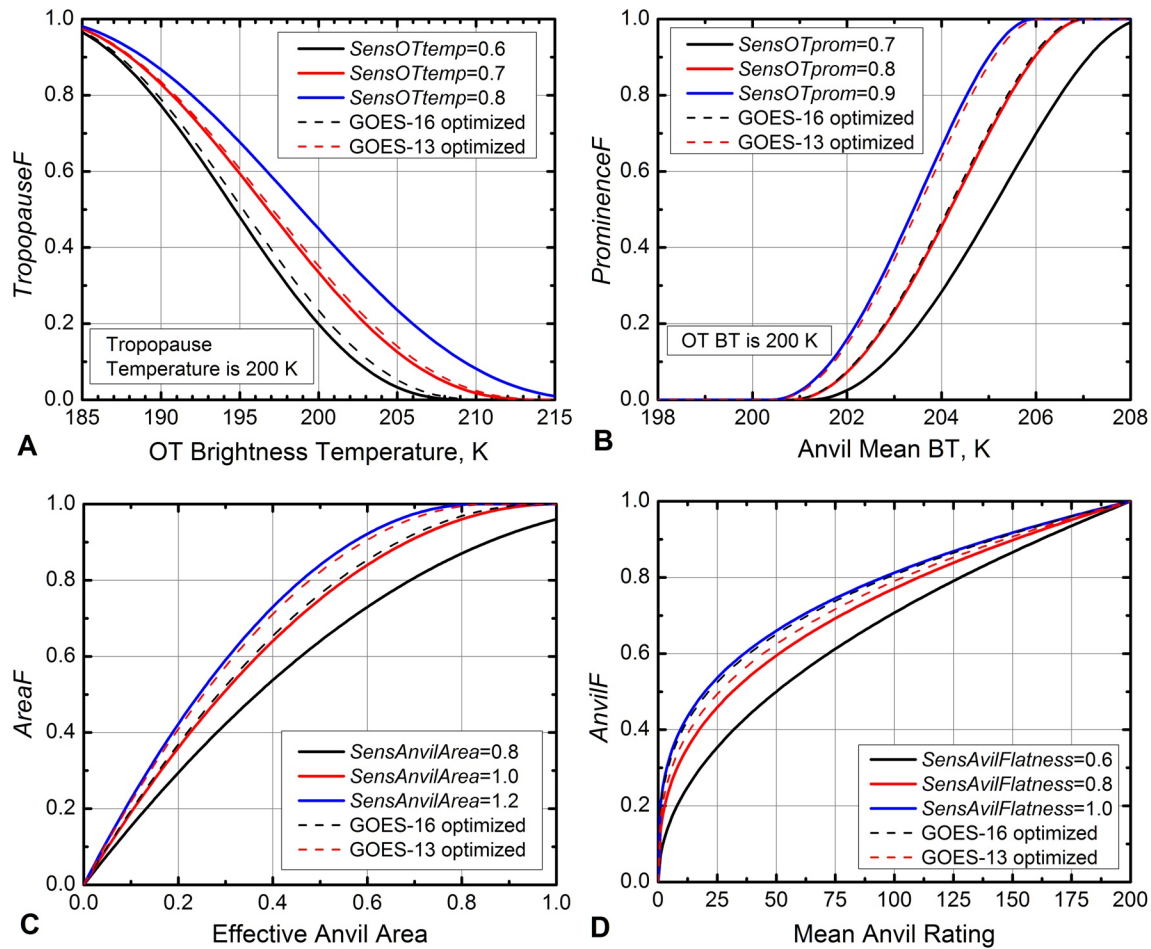
Samples of the BT and the anvil rating are acquired along each ray (by using the Lanczos interpolation) as long as local BT stays within the range of allowed temperatures defined as  $BT_{\text{peak}} \pm 1.3 K$ , a threshold based on empirical testing. The processing of each ray stops when the distance from the center exceeds the current radius,  $R_{\text{th}}$ , or more than one out-of-range samples are encountered, which effectively means that the anvil boundary has been reached. As a result, the mean temperature  $WinAvgBT$  and the mean anvil rating  $WinAvgAnvil$  are computed by averaging all of the samples acquired along the rays. The number of used samples is divided by the maximum possible number of pixels along all rays for the current radius, and this ratio defines an effective anvil area  $AnvilArea$ . As a last step, the four cases of  $WinAvgBT$ ,  $WinAvgAnvil$ , and  $AnvilArea$  obtained for the two major peaks of the two histograms are weighted-averaged with the weights equal to their corresponding effective anvil areas. For the strongest OT region embedded in the anvil cloud shown in Figure 6, the derived anvil extents satisfying the  $BT_{\text{peak}} \pm 1.3 K$  condition are shown in Figure 9. Accordingly, the following anvil mean parameters are found:  $WinAvgBT = 209.55 K$ ,  $WinAvgAnvil = 127.6$ , and  $AnvilArea = 0.2377$ , with the OT's lowest BT of 196.76 K, and the tropopause temperature of 208.24 K.

Our extensive tests have found that the described two-peak approach is imperative in situations when the OT vicinity is comprised of complex non-uniform cloud structures, especially in cases when multiple OTs are clustered together. Having the histograms sampled over differently sized areas and blending the results from two histogram peaks ensures a smooth change in the resulting anvil metrics when the histogram's maximum transitions from one peak to another due to a natural time evolution of a convective area. This ultimately helps the overall temporal stability of the OT probability, in particular when dealing with 30-s to 1-min Mesoscale Domain Sector observations from the GOES-16/17 imager.

Once the anvil mean parameters are retrieved it is possible to describe how the factors comprising the OT probability are constructed. The  $TropopauseF$  quantifies the effect of OT temperature on the resulting OT probability, as a colder OT implies a higher detection confidence. This confidence flattens out at extremely low temperatures as any probability derivation in general saturates when approaching 100%. Therefore, the  $TropopauseF$  is calculated as follows:

$$TropopauseF = Z \left( 1 - Z \left( \left( BT_p / T_{tp} - 0.91 \right) \cdot 4.3 / SensOTtemp \right)^2 \right)^3 \quad (8)$$

where  $SensOTtemp$  is a "Sensitivity to OT temperature" parameter, typically  $0.65 \pm 0.10$ . Figure 10a shows the  $TropopauseF$  curves calculated for a tropopause temperature of 200 K for different values of  $SensOTtemp$ . The shape of the curve is based on empirical analyses of tropopause-relative BT characteristics of OTs by the authors, supported by NEXRAD-based OT analyses described below. The formula is built around the



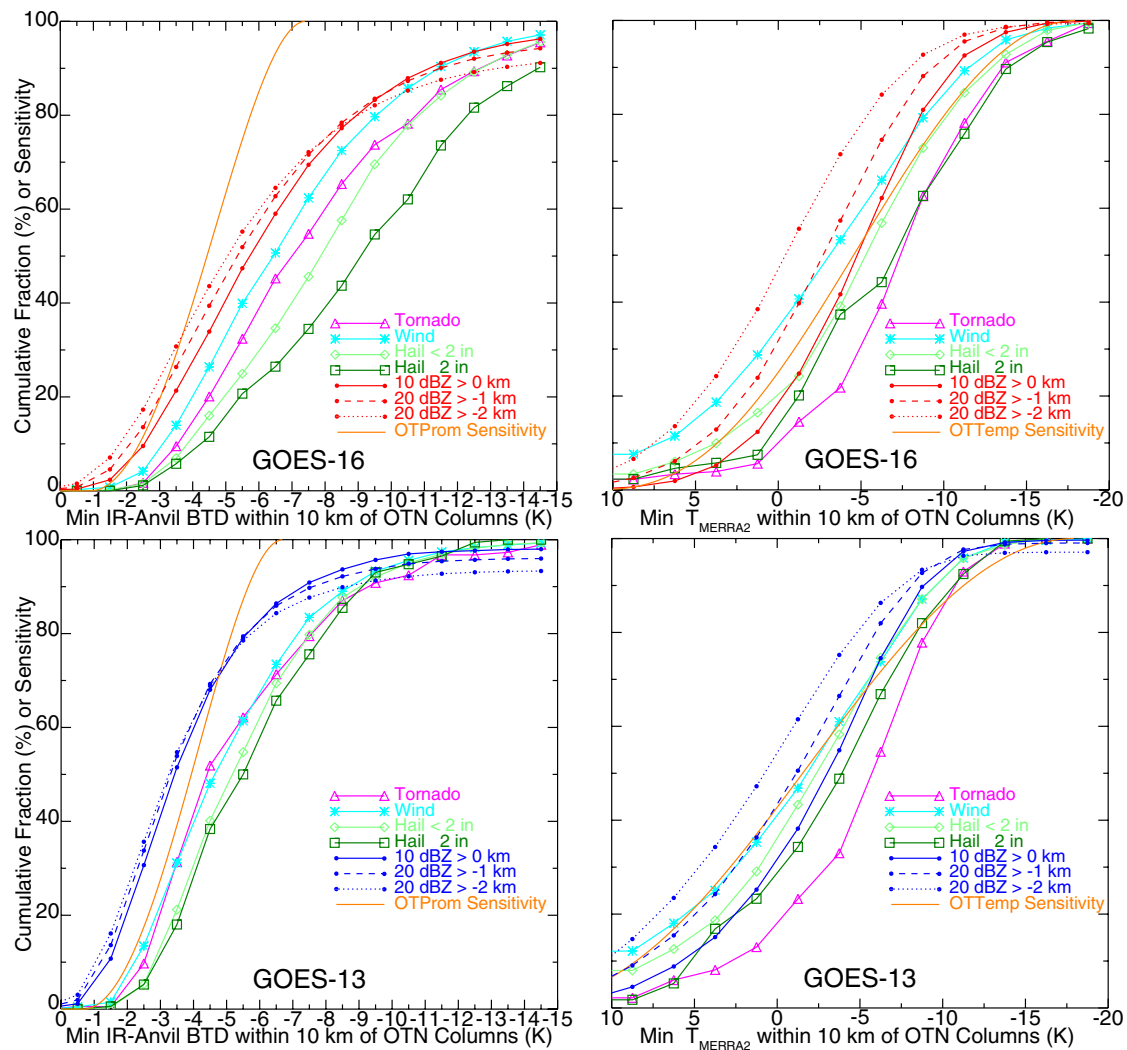
**Figure 10.** Dependences of (a) the tropopause factor on the OT temperature calculated for the tropopause temperature of 200K (b) the prominence factor on the anvil mean temperature calculated at the fixed OT BT of 200K (c) the area factor on the effective anvil area, and (d) the anvil factor on the mean anvil rating. Sensitivities associated with optimal GOES-13 and -16 detection of human-identified OT regions are depicted by dashed curves.

ratio of the OT brightness temperature,  $BT_p$ , to the tropopause temperature,  $T_{tp}$ , and thus works relative to the changing tropopause temperature and eliminates fixed temperature thresholds used in previous OT detection algorithms (e.g., Bedka et al., 2010 and references therein). The square power in Equation 8 controls how the curves approach the saturation level of 1.0, while the third power is responsible for a smoother decay towards zero. One can see from Figure 10,  $TropopauseF$  decreases slower with higher values of  $SensOTtemp$ , implying that a higher sensitivity makes the  $TropopauseF$  larger for the same ratio of  $BT_p$  to  $T_{tp}$ .

The  $ProminenceF$  describes the effect of OT prominence (in terms of temperature) relative to surrounding anvil, and so it is based on the ratio of  $WinAvgBT$  to the  $BT_p$  and is calculated as follows:

$$ProminenceF = 1 - Z \left( 1 - Z \left( \left( WinAvgBT / BT_p - 1.02 + 0.02 \cdot SensOTprom \right) \cdot 40 \cdot SensOTprom \right)^2 \right)^2 \quad (9)$$

where  $SensOTprom$  is a “Sensitivity to OT prominence” parameter, typically  $0.80 \pm 0.10$ . Figure 10b shows three sample curves (solid lines) of the  $ProminenceF$  obtained for different values of  $SensOTprom$  at a fixed  $BT_p$  of 200 K. The ratio of  $WinAvgBT$  to  $BT_p$  is effectively scaled by  $SensOTprom$  to enhance the steepness of the curve and, similar to the  $TropopauseF$  factor, a higher sensitivity  $SensOTprom$  yields a larger  $ProminenceF$ . The formula also includes a small offset of  $0.02 \cdot SensOTprom$ , which shifts the curve to the right with lower  $SensOTprom$  and delays the rise of the  $ProminenceF$  until the difference from the anvil mean

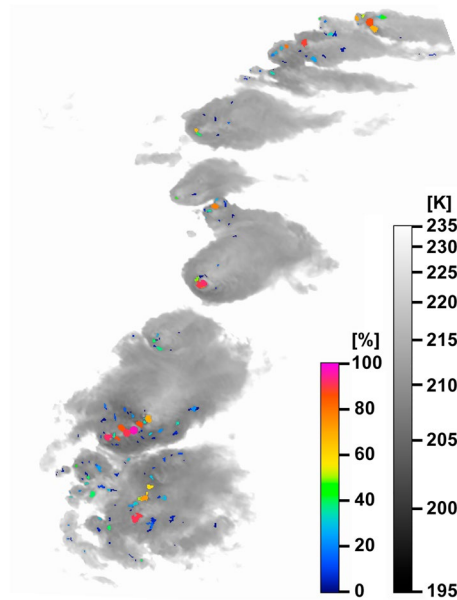


**Figure 11.** Cumulative frequency diagrams of the coldest GOES-16 (top) and GOES-13 (bottom) infrared (IR) minus anvil brightness temperature (BT) difference (i.e., prominence) and tropopause-relative IR BT values within 10 km of GridRad radar echoes at varying tropopause-relative heights, 20 dBZ echo top 2 km below tropopause and 1 km below tropopause, and 10 dBZ top at or above the tropopause defined by MERRA-2. Distributions of the same GOES IR parameters within 2.5 min (GOES-16) or 7.5 min (GOES-13) and ~30 km of severe weather reports are also shown. The database is dominated by wind and 1–2 inch (2.5–5 cm) hail events, but includes 189 tornado and 173 2+ inch diameter hail events. The sensitivity curves found to be optimal based on analysis of GOES data within human-identified OT regions are also shown.

temperature becomes sufficient. This prevents false detection of minor dips in the temperature field of non-perfectly uniform anvils.

The profiles of the two sensitivity curves above are supported by cumulative frequency diagrams (CFD) derived using a large database of GridRad 20 dBZ echo tops (59,248 samples) above the level of tropopause altitude minus 2 km (see Cooney et al., 2021) and reported severe weather events from the NOAA National Weather Service Storm Prediction Center database (3,967 samples). This database of high-altitude echo tops was compiled at 5-min intervals over 15 days in the year 2017 across the U.S. when both GOES-16 and GOES-13 (Hillger & Schmit, 2007) were observing the same storms in periods of widespread intense convection. Figure 11 shows the cumulative frequency as a function of tropopause-relative IR BT (see *BTTD* in Equation 2) and IR minus anvil BT difference (i.e., prominence). The bulk of the distribution ranges from +10 to −15 K for *BTTD* and from −1 to −14 K for prominence. These frequency distributions suggest that the corresponding *TropopauseF* and *ProminenceF* curves should stay inside these limits, because virtually no OTs are actually observed outside these ranges. Extending the sensitivity curves beyond these limits is





**Figure 12.** Final product of OT probability after the spatial expansion of OT cores (colored pixels) overlaid on a grayscale image of GOES-16 10.3  $\mu\text{m}$  brightness temperature on May 5, 2019 at 23:00 UTC.

not desired, as this will make them less steep and will also reduce the operating range of *TropopauseF* and *ProminenceF* factors. As seen from Figure 11, coarser resolution of GOES-13 reduces the CFD ranges, which suggests that the steepness of the sensitivity curves need to be increased by adjusting their corresponding sensitivities. The effect of coarser sensor's resolution on the OT detection efficiency is discussed in Section 5.

The third factor contributing to the OT probability is the area factor, *AreaF*. Experience with analysis of OT-producing storms has revealed that storms with larger anvils are more likely to generate OTs than storms with little to no anvil area. Thus, *AreaF* is designed to lower the OT probability when an OT candidate has insufficient anvil area nearby:

$$\text{AreaF} = 1 - Z(1 - \text{SensAnvilArea} \cdot \text{AnvilArea})^2 \quad (10)$$

where the *SensAnvilArea* is a configurable parameter (typical value is  $1.0 \pm 0.2$ ) that controls the influence of the area factor on the OT probability, as shown in Figure 10c.

The last component of the OT probability is the anvil factor, *AnvilF*, which is designed to assign higher OT probability to OTs located in areas with a higher anvil rating, which in turn describes mostly the anvil's uniformity and coldness in a tropopause-relative sense of its BT field (see Equation 4). It is more likely that a pixel is truly an OT if it is embedded within a region of uniformly cold anvil cloud, in contrast to a cold spot surrounded by broken cold pixels that are not as likely to comprise a true

anvil cloud. Thus, it is made proportional to the mean anvil rating *WinAvgAnvil* obtained by sampling with the ray casting above:

$$\text{AnvilF} = (\text{WinAvgAnvil} / 200)^{0.3/\text{SensAnvilFlatness}} \quad (11)$$

Here the *SensAnvilFlatness* is a configurable parameter (typical value is  $0.8 \pm 0.2$ ) that, being in the power of the exponent, controls the steepness of the curve and thereby the sensitivity to the anvil rating as shown in Figure 10d.

These three factors—the prominence, area, and anvil—describe an OT's prominence relative to a wide, uniform anvil cloud. If any of these conditions are not met, then the OT candidate is not likely to correspond to a true OT and thus should be assigned a lower detection probability. After extensive tests and comparisons of various formulations, we decided to combine all three factors symmetrically in one  $\lambda$ -function as follows:

$$\lambda = \sqrt{\text{ProminenceF} \cdot \text{AreaF} \cdot \text{AnvilF}} \quad (12)$$

Using all three factors in one product requires all of them to be high concurrently in order to produce a significant level of  $\lambda$ . The square root here corrects for too-low levels of the three-component product and helps to balance the distribution of contour lines in Figure 7.

Thus far, an OT candidate represents a single pixel corresponding to a local maximum of BT-score, however, an overshooting updraft core typically extends beyond one IR pixel. The process for expanding the updraft core to encompass all pixels within a given OT is described in Appendix A3. Once an OT's extent is defined, its pixels are filled with the derived IR OT probability making up a region that spatially matches the input reprojected BT. This is demonstrated in Figure 12 where the resulting OT probability is overlaid on a grayscale BT image. The OT probability is presented in color scale from 0 to 100 percent. One can see that all strong OT cores (see Figure 2) are covered with regions of OT probability reaching 80% and higher.

In summary, we have provided a detailed technical background on a new OT detection method, that offers several significant improvements relative to the previous Bedka and Khlopenkov (2016) method:

- (1) IR BT is now normalized to the tropopause, whereas the previous version used an areal mean BT as a reference.
- (2) IR anvil mask is now a completely new product and presents a per-pixel anvil rating, whereas the previous version used polygons to identify anvil cloud only in the near vicinity to an OT candidate that excluded the outer periphery of cold anvil regions.
- (3) Identification of OT candidates has been fully redesigned.
- (4) OT probability is derived using a completely new approach based on combination of four factors that better captures the spatial characteristics of the OT candidate and anvil region.
- (5) The current algorithm is using a more sophisticated method for estimating anvil mean parameters around an OT candidate.
- (6) OT spatial extent is defined more accurately.
- (7) Key sensitivities are derived from an extensive statistical analysis based on manual OT identification (see Section 4 below).
- (8) The current algorithm automatically adjusts the sensitivities according to the spatial resolution of the input BT image in order to achieve, on average, the same OT probability independent of the satellite (see Section 5 below)

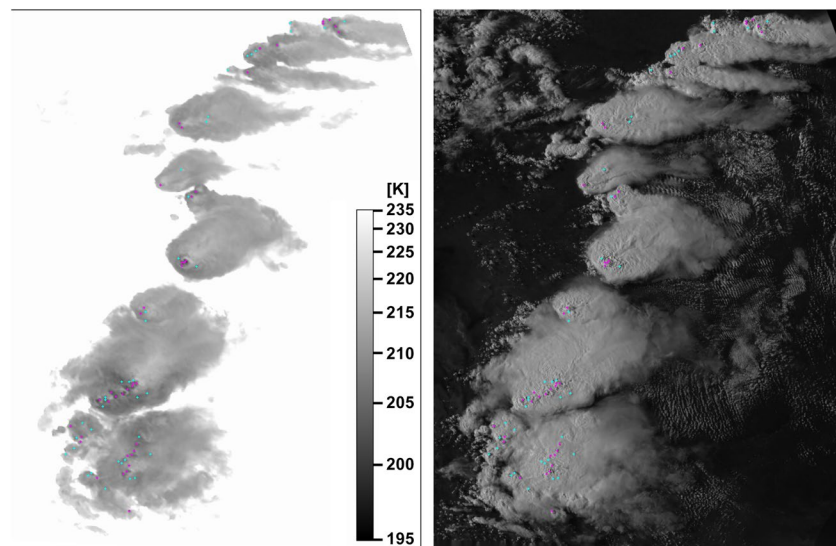
#### 4. Optimization of OT Probability

During development and improvement of the OT detection algorithm, expert feedback from scientists has been constantly utilized to fine-tune the key parameters and thresholds used in the equations above. This way, the human perception can serve as the first direct validation of the correct performance of our algorithm provided that this analysis is rigorously formulated and quantified. Detection performance relative to OT detections from precipitation radars such as NEXRAD, a quantitative estimate for OT locations, is demonstrated by Cooney et al. (2021). The current algorithm is designed to assign high confidence to features that are clearly OTs to human analysts, and thus can only detect features that are distinct in the imagery. Therefore, a pixel mask of OT locations was produced by a human analyst based primarily on the VIS image but supported by IR data from the same case as Figure 12. The higher resolution VIS image is particularly helpful for manual identification of OT locations and was used as an independent verification since the OT detection scheme described above is based solely on IR-based data.

Three classes are included in the human identified mask, that is *no-OT*, *weak OT*, and *strong OT*. Compared to a binary (yes/no) mask, the *weak OT* class is helpful in that it provides more flexibility in the statistical comparison between a continuous variable (the OT probability from 0 to 100) and a discrete human OT identification. The exact pixel locations of OT cores in the human identified mask were bound to the local BT minima in the IR image, so as to exclude any misdetections due to spatial mismatch and to focus only on the accuracy of detection probability. In total, the human mask shown in Figure 13 contains 40 *strong OT* locations and 39 *weak OT* ones. We should note that this analysis excludes pixel pairs where the human mask shows *no-OT* and the OT detection probability is below 0.5%, which are the majority of pixels in the image and correspond to “true negative” detections.

With these considerations, the human mask is matched against the algorithm-generated OT detection probability with the following major goals (a) find the optimal set of algorithm's parameters (i.e., sensitivities *SensOTtemp*, *SensOTprom*, *SensAnvilArea*, and *SensAnvilFlatness*): That yields the best agreement with the human identified mask; and (b) assess the statistical accuracy of OT detection, that is achieving the highest probability of detection (POD) while keeping the false alarm ratio (FAR) low.

As a measure of agreement between the OT probability and the human mask, the Spearman correlation is appropriate, because it is designed to analyze the statistical dependence between continuous and rank (i.e., discrete) variables. Therefore, the Spearman correlation coefficient,  $\rho$ , can be used to optimize the four sensitivities so as to achieve the highest correlation with the human mask. This can be implemented by means of the Powell's conjugate direction method (Powell, 1964), which iteratively converges to a local minimum of a function (this can be, for example,  $1-\rho$ ) in multidimensional space while varying the input sensitivities. Due to the high dimensionality of our problem, the iterative routine may be trapped around a local minimum if the initial guess values happen to be far from the actual absolute minimum. To avoid

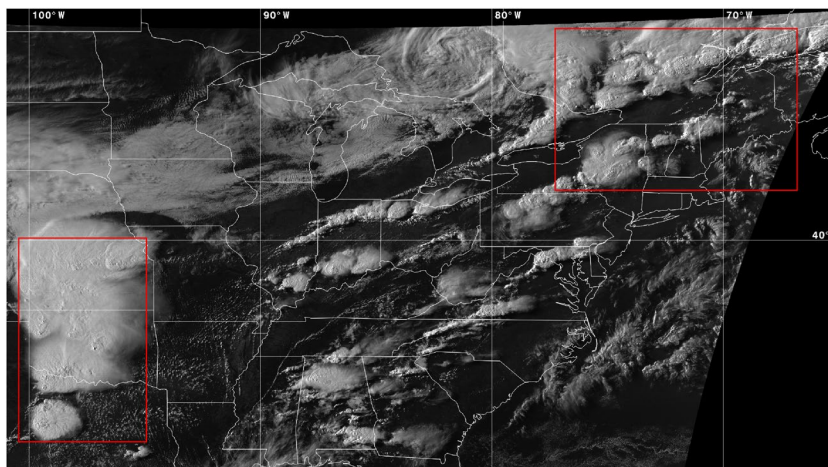


**Figure 13.** Human identified mask overlaid on a GOES-16 infrared (IR) brightness temperature (BT) image (left) and VIS image (right) on May 5, 2019 at 23:00 UTC. The most confident OT locations are shown with magenta (“strong OT” class) and less confident overshooting cloud tops (OTs) are colored in cyan (“weak OT” class).

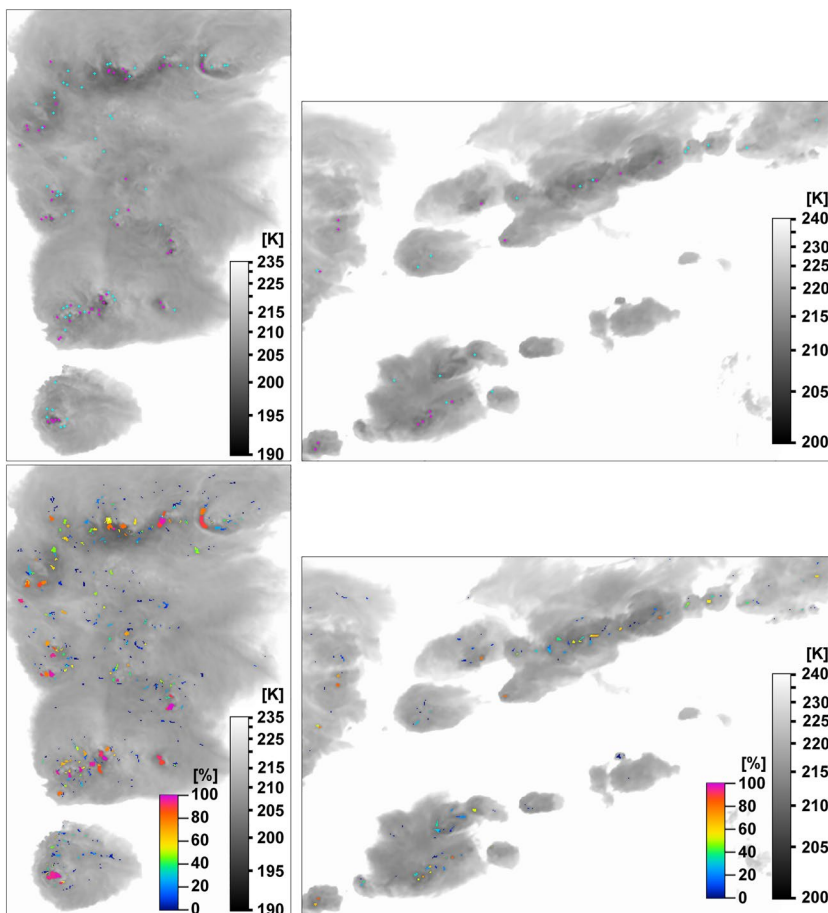
that, the correlation was first computed over the following pre-selected ranges of the four sensitivities:  $SensOT_{temp} = 0.65 \pm 0.10$ ,  $SensOT_{prom} = 0.80 \pm 0.10$ ,  $SensAnvilArea = 1.00 \pm 0.10$ , and  $SensAnvilFlatness = 0.90 \pm 0.10$ . The 0.10 variance in the analysis of  $SensOT_{temp}$  and  $SensOT_{prom}$  allows the sensitivities to encompass the dynamic range of GOES-13 and GOES-16 tropopause-relative BT and anvil prominence shown in Figure 11. Each of the sensitivities is varied by a 0.05 step, which translates into a total of  $5 \times 5 \times 5 \times 5 = 625$  calculations of the correlation. Among those, the two best sets of sensitivities are selected and then used as the initial guess values in the Powell’s minimization. Thus, the iterative routine is executed twice and the highest correlation among the two results is selected, yielding the optimal set of the four sensitivities. This approach significantly improves the overall stability of the convergence and helps the iterative routine to converge to the true absolute minimum. As a result of this optimization, the correlation  $\rho$  of 0.7914 is achieved with the following sensitivities:  $SensOT_{temp} = 0.6313$ ,  $SensOT_{prom} = 0.8275$ ,  $SensAnvilArea = 0.9020$ , and  $SensAnvilFlatness = 0.7502$ .

Once the key sensitivities are defined, the OT detection algorithm is ready to be applied to any IR satellite image with 2–4 km spacing per pixel at nadir, which includes nearly all current and historical geostationary imagers from recent decades. In order to validate the sensitivities derived from this single case, we have selected three additional cases of GOES-16 data from May 18, 2017 at 22:47 UTC (see Figure 14) August 13, 2017 23:02 UTC, and April 29, 2017 23:00 UTC for analysis. Figure 14 image and the other two cases (not shown) provide a wide variety of convective regions with varying intensity extending across a broad geographic domain. If the OT detection algorithm is capable of detecting OTs correctly from these cases, then this builds confidence that it will operate reliably afterwards. To demonstrate the qualitative agreement between OT probability and human OT identifications, two zoomed in areas of the mask are shown in Figure 15, which correspond to the red frames in Figure 14. The large anvil cloud in the left frame exhibits much lower temperature down to 190 K, while the convective cells in the right frame are mostly warmer with OT cores at 205–210 K. The colder anvil area in Figure 15-upper left is identified to have a higher density of OT locations compared to the convective cells in Figure 15-upper right. The human mask for this case is also using three classes for OT identification: *no-OT*, *weak OT* (cyan color), and *strong OT* (magenta color). For all of the three cases in total, 452 *strong OT* and 985 *weak OT* locations are identified, which presents a much larger statistical sample than the case shown in Figure 13.

The OT probability data obtained from the three additional cases are added to those from the initial May 5, 2019 case and their corresponding human masks are also combined into one statistical sample for subsequent optimization of the algorithm’s sensitivities. Table 1 presents an example of how the correlation



**Figure 14.** GOES-16 0.65  $\mu\text{m}$  visible reflectance image on May 18, 2017 22:47 UTC. The two frames in red show the regions where the human overshooting cloud tops (OT) mask will be zoomed in for closer viewing.



**Figure 15.** Two subset images of the human identified mask of OT locations (using the same color scheme as in Figure 13) overlaid on BT image (grayscale) for GOES-16 case on May 18, 2017 at 22:47 UTC (Upper panels). The same BT images overlaid with the overshooting cloud tops (OT) probability (colored spots) computed using the optimized set of sensitivities (Lower panels).



**Table 1**

Dependence of the Correlation Coefficient  $\rho$  on a Range of “Sensitivity to OT Temperature” and “Sensitivity to Peak Prominence” for a Combined Statistical Sample From GOES-16 Cases

Sens. anvil area = 1.00		Sensitivity to peak prominence				
Sens. anvil flatness = 0.95		0.70	0.75	0.80	0.85	0.90
Sensitivity to OT Temperature	0.55	0.7812	0.8342	0.8693	0.8824	0.8728
	0.60	0.8301	0.8720	0.8912	0.8876	0.8627
	0.65	0.8581	0.8874	<b>0.8920</b>	0.8749	0.8385
	0.70	0.8702	0.8873	0.8803	0.8527	0.8079
	0.75	0.8727	0.8785	0.8617	0.8257	0.7756

coefficient,  $\rho$ , changes while varying the input sensitivities. To simplify the presentation, the “Sensitivity to Anvil Area” and “Sensitivity to Anvil Flatness” are fixed here at 1.00 and 0.95, respectively. The best correlation  $\rho$  of 0.8920 is observed at  $SensOTtemp = 0.65$  and  $SensOTprom = 0.80$ . These values are then used as the initial guess values for the iterative minimization, which finally yields  $\rho$  of 0.8941 with  $SensOTtemp = 0.6252$ ,  $SensOTprom = 0.8052$ ,  $SensAnvilArea = 1.0284$ , and  $SensAnvilFlatness = 0.9676$ . The following mean and standard deviation values of OT probability have been achieved:  $8.97 \pm 10.04$  for *no-OT* class,  $38.93 \pm 14.81$  for *weak OT* class, and  $77.90 \pm 10.04$  for *strong OT* class. The obtained sensitivities are similar to those derived for the May 5, 2019 case alone, except with slightly higher  $SensAnvilArea$  and  $SensAnvilFlatness$ . This can be explained by a higher variability in sizes and homogeneity of anvil clouds in the three additional cases due to a much larger sample, which causes the optimization to adjust to the higher sen-

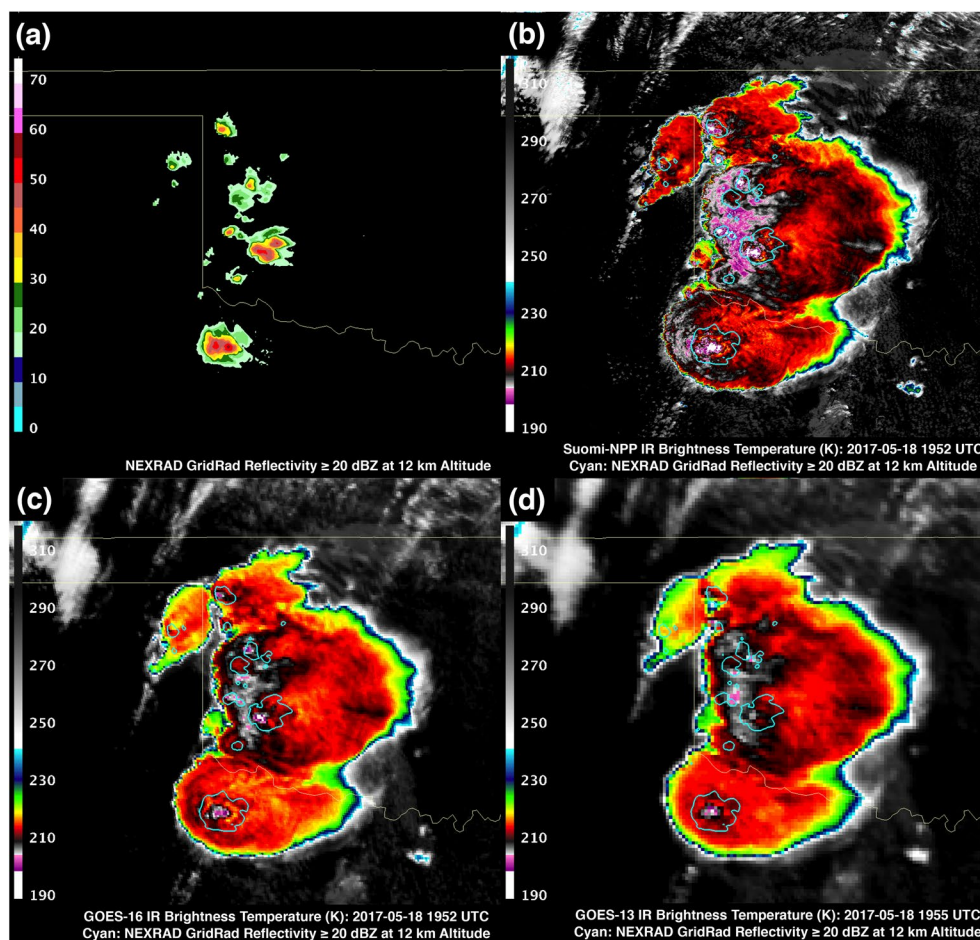
sitivities  $SensAnvilArea$  and  $SensAnvilFlatness$ . The corresponding OT probability computed by using the optimized set of sensitivities is shown in the bottom panels of Figure 15.

Sensitivity curves associated with these parameters are shown in Figure 11 (orange lines). They align very well with the distribution for echo tops above the height 2 km beneath the tropopause. As echo tops increase in height, GOES tropopause-relative BT and to a lesser extent, prominence, indicate a more intense storm. Severe storms, especially those that produce tornadoes and 2+ inch (5+ cm) diameter hail, are colder and have greater prominence compared with non-severe storms that dominate the NEXRAD OT database. The sensitivity curves would allow most severe storms to be detected, except ~5%–10% of the severe storm population where  $BTTD > 10$  K.

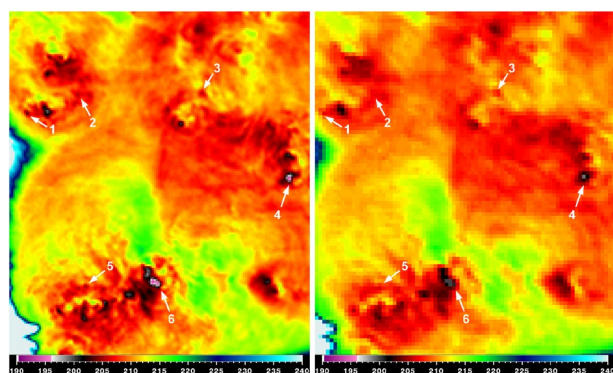
## 5. Impact of Image Resolution on OT Detection

For effective climate analyses, detection algorithms must be able to operate consistently across transitions in satellites, such as from the GOES-8 to –15 era to GOES-16/17, where imager spatial sampling, image navigation, and sensor noise characteristics improve throughout this 25+ year data record. Higher resolution, better-quality imagery causes OTs to appear more prominently, which would bias a time series if such image improvements are not accounted for in detection algorithm formulation. Figure 16 shows an example of how the appearance of cloud tops differs between imagery from different satellite sensors. The tropopause height, derived from MERRA-2 where convection is present, ranges from 12–13 km. Thus, reflectivity exceeding 20 dBZ derived from the NEXRAD GridRad data set at an altitude of 12 km provides a good proxy for OTs (Figure 16a). The Suomi-NPP Visible Infrared Imaging Radiometer Suite (VIIRS) collected 11.45  $\mu\text{m}$  channel IR BT imagery of these OT-producing storms with 375 m pixel spacing at nadir (Figure 16b). This imagery depicts a very complex BT pattern throughout the anvil cloud regions. NEXRAD OT regions overlaid upon the VIIRS image (cyan contours) are usually co-located with distinct IR BT minima, but the coldest BTs associated with the OTs vary considerably and many cold areas (magenta) occur outside the NEXRAD OTs. GOES-16 ABI viewed these storms within 1 min of the VIIRS image at a resolution of ~2.5–3 km/pixel and shows much warmer cloud tops in the 10.3  $\mu\text{m}$  data (Figure 16c). Differences in OT region BTs exceed 10 K in nearly all instances which is mostly attributed to a factor of seven or greater difference in pixel size. The GOES-13 Imager viewed these storms ~3 min after the ABI and VIIRS with 5–6 km pixel size (Figure 16d). Though this image looks quite similar to the ABI image, close inspection of OT regions reveals BT differences exceeding 2 K for nearly all OTs. This example is consistent with differences across a large OT database described in this paper and further reinforces that no single IR BT threshold will provide consistent OT detection accuracy, even across a relatively small region, thus detection algorithms should be flexible to account for variations in image detail and quality.

Assuming the derived sensitivities are the optimal parameters for reliable OT detection, it is important to verify the algorithm’s performance when applied to IR BT observed by older satellites, such as GOES-13 with a coarser pixel resolution of 4 km/pixel at nadir. The three additional GOES-16 scenes listed above were observed very close in time by GOES-13 (time difference ~2 min). Thus, the spatial mismatch between



**Figure 16.** (a) NEXRAD GridRad reflectivity exceeding 20 dBZ at 12 km altitude, compiled using data collected from 19:50–19:55 UTC on 18 May 2017 over western Oklahoma and northern Texas. (b) Suomi-NPP VIIRS 11.45  $\mu\text{m}$  infrared (IR) brightness temperature (BT) at 19:52 UTC. The image was shifted to account for parallax and align with the GOES-16 and GOES-13 data. (c) Parallax-corrected GOES-16 10.3  $\mu\text{m}$  IR BT at 19:52 UTC. (d) Parallax-corrected GOES-13 10.7  $\mu\text{m}$  IR BT at ~19:56 UTC. Cyan contours on the VIIRS, GOES-16, and GOES-13 images correspond to the radar echoes in the upper-left panel.



**Figure 17.** Comparison of BT images from GOES-16 on May 18, 2017 at 22:47 UTC at 2 km/pixel (left panel) and GOES-13 on May 18, 2017 at 22:45 UTC at 4 km/pixel (right panel) over the same area in Oklahoma corresponding to the middle of the left frame of Figure 14.

the cloud features is minimal in this case, which allows us to focus on the influence of the imager's spatial resolution on the derived OT probability.

Figure 17 provides a zoomed view of IR BT images recorded by the two different imagers for one of the three scenes used to validate both GOES-13 and -16. The image from GOES-13 is naturally less sharp and does not resolve many of the compact convective cells (indicated by white arrows in the Figure), which are more evident in the GOES-16 ABI data. This not only reduces the OT prominence relative to anvil, but also causes some local BT minima (used as initial OT candidates) to disappear completely. As a result, the OT probability product from GOES-13 can be biased lower, and this needs to be compensated by adjusting the algorithm's parameters in order to arrive at another set of sensitivities optimized for a coarser resolution input.

To ensure as equivalent as possible operation with GOES-13 imagery, it is important to use a human identified OT mask from the same scene (as shown in Figure 14) with the same number of *weak OT* and *strong OT*

locations. To achieve that, the OT locations in the GOES-16 human OT mask are first displaced according to motion vectors computed from correlation matching between IR images from GOES-16 and GOES-13, and then bound to the nearest local minima in GOES-13 BT. As a result of this process, 86 locations happened to have no matching local minimum due to the lower sharpness of the GOES-13 image. Among other locations, 36 are found to have zero OT probability, which can be corrected by increasing the sensitivities, and so the human mask was assigned to the *weak OT* class at such locations. This refined mask was finally used in the iterative minimization, which resulted in the correlation  $\rho$  of 0.7518 with the following optimal sensitivities:  $SensOTtemp = 0.7135$ ,  $SensOTprom = 0.8881$ ,  $SensAnvilArea = 1.1558$ , and  $SensAnvilFlatness = 0.8829$ . Sensitivity curves from these parameters are depicted by the red dashed curves in Figure 10. The following mean and standard deviation values of OT probability have been achieved in this case:  $11.83 \pm 13.79$  for *no-OT* class,  $39.66 \pm 20.05$  for *weak OT* class, and  $68.70 \pm 23.83$  for *strong OT* class.

As expected, all of these GOES-13 sensitivities are higher than those derived for GOES-16, except for the lower  $SensAnvilFlatness$ . This is because the effect of the anvil rating controlled by  $SensAnvilFlatness$  is quite stable and uniform across most anvils (see the uniform field in Figure 5 right panel), and thus variation of  $SensAnvilFlatness$  merely scales up or down the overall OT probability. Therefore, the increase in the first three sensitivities (caused by the reduced sharpness of GOES-13 imagery) is automatically compensated by the optimization routine with the lower level of  $SensAnvilFlatness$ . These adjusted sensitivities should help to offset resolution-induced bias, but some weak OTs will still be difficult to detect with coarser imagery like that from GOES-13. Cooney et al. (2021) quantifies the accuracy and consistency of OT detections relative to NEXRAD GridRad, but an initial estimate of accuracy can also be gained from analysis of human-identified OTs.

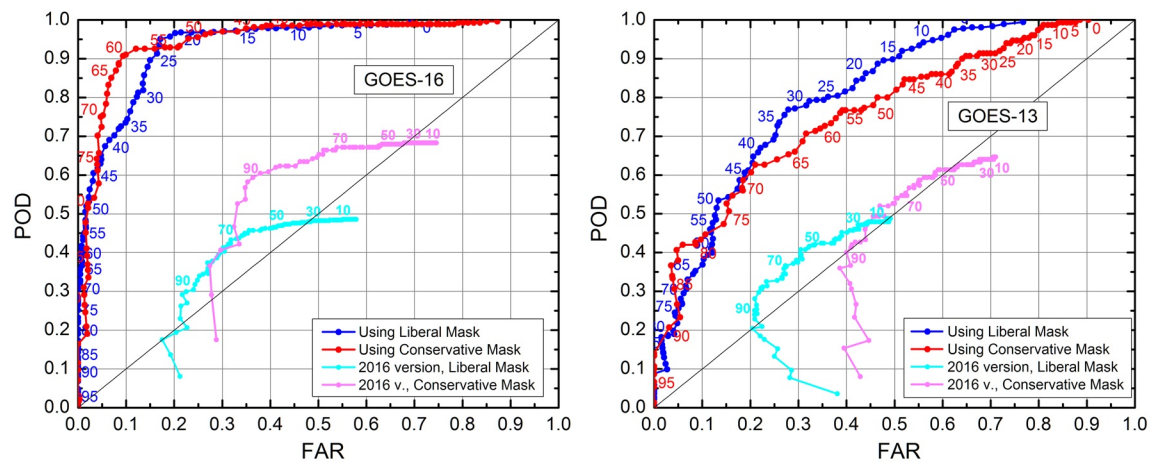
## 6. Statistical Analysis

The OT probability product provides a flexible rating that estimates the confidence in a detection from 0 to 100. This flexibility, however, does not answer what threshold best separates positive detections from negative. For the purpose of a dichotomous (yes/no) forecasting, the OT probability has to be converted to either positive (OT is detected) or negative (no detection) events using a certain probability threshold (PT). On the other side, the human OT mask can also be reduced to a two-class variable by one of two ways, by treating only the *strong OT* class as true OTs and the rest as no OTs (conservative mask) or using both *weak OT* and *strong OT* classes (liberal mask). Thus, there are four possible combinations of forecasts (OT detection) and observations (the human mask): True positive (TP), false positive (FP), true negative (TN), and false negative (FN). Probability of detection (POD), defined as  $POD = TP / (TP + FN)$ , and the false alarm ratio (FAR), defined as  $FAR = FP / (TP + FP)$ , are the well-known statistical metrics used to assess the quality of a forecast product. Their main advantage is that they are not based on the TN events, which have no practical use with OT detection given that the vast majority of the imagery has clear sky or low cloud, which is always correctly identified as TN by our algorithm.

Dependence of POD on FAR is analyzed in the form of receiver operating characteristic (ROC) curves shown in Figure 18, where the red curves use the conservative OT masks and the blue curves use the liberal masks. The numerical labels next to the data points show the probability threshold (PT) associated with a given POD and FAR. For GOES-16 (left panel) and a PT of 50%, 95% of the strong OTs were detected, but 24% of the GOES detections were not co-located with a strong OT. Using the liberal mask, 51% of OTs were detected for PT of 50. The reduction is caused by the fact that subtle textured areas in the anvil may have very minimal prominence in IR imagery and are harder to detect. But the false detection rate decreases to 4% when weak OTs are included because a substantial fraction of the GOES detections happen to be co-located with weak OTs. Area under the ROC curve (AUC) are both near 0.94. The greatest POD minus FAR difference for conservative (liberal) is from a PT of 61 (23) with POD of 0.9067 (0.9504) and FAR of 0.0933 (0.1716).

For GOES-13 and a PT of 50, the POD (0.80) is lower than for GOES-16 (0.95) with the conservative mask because OTs especially evident in visible imagery are not always especially cold or prominent in coarser resolution GOES-13 data. The 15% reduction in detection rate is consistent with the GOES-16 versus -13 results from Cooney et al. (2021) based on NEXRAD OT regions. For the liberal mask, GOES-13 POD is





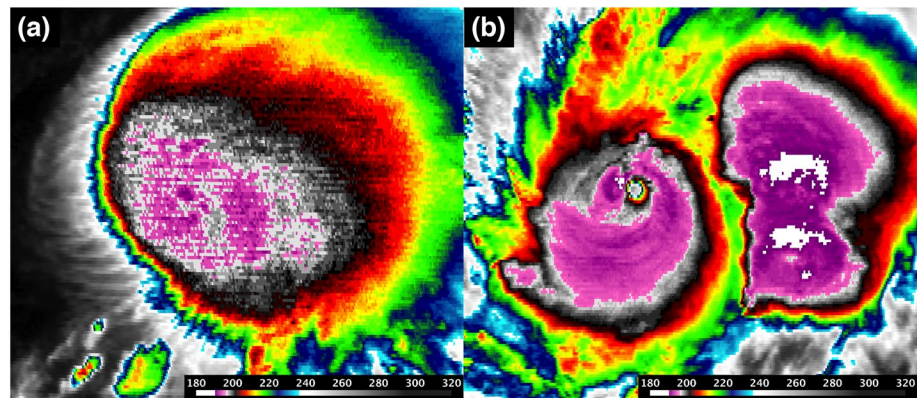
**Figure 18.** Receiver operating characteristic (ROC) curves showing probability of detection (POD) versus false alarm ratio (FAR) low FAR accumulated using data from four GOES-16 images (left) and three GOES-13 images (right) for the current detection method (red and blue lines, see legend) and the Bedka and Khlopenkov (2016) method (magenta and cyan lines). GOES-13 images were collected within two minutes of the corresponding GOES-16 images.

nearly equal to GOES-16, but false alarm rate increased by 10%. This is likely caused by the fact that higher sensitivities required for optimal detection (Figure 10) allow warmer and less prominent cold spots to be detected. Our experience indicates that anvils routinely have random spatial temperature variations of 3 K or more, so trying to detect weak OT feature that too has a 3 K prominence increases likelihood of false detection. AUC is 0.765 (0.803) for the liberal (conservative) mask. The greatest difference in POD minus FAR for conservative (liberal) is from a PT of 68 (32), very similar to GOES-16, with POD of 0.6267 (0.7686) and FAR of 0.2101 (0.2791). Despite differences in POD and FAR between the two satellites, Cooney et al. (2021) shows that the GridRad echo top distribution is nearly the same between GOES-13 and GOES-16 for the same PT, which gives us confidence that the optimal sensitivity values found in this paper provide as consistent as possible detection capability between the two satellites.

ROC curves for the Bedka and Khlopenkov (2016) method are also shown in Figure 18 (magenta and cyan). It is clear that detection performance is significantly poorer than the method described in this paper, and relative to the results derived from MODIS cases shown in Bedka and Khlopenkov (2016). We attribute a reduced POD to how OT candidates were filtered in the previous method, where metrics of anvil area, spatial uniformity, and OT shape were previously used to include/exclude an OT candidate from further processing, which prohibited detection of some human-identified OTs here. We attribute increased FAR, which most often occurred in random temperature variations in cold non-OT anvil cloud, to less precise derivation of anvil cloud properties. Overall performance is worse for these cases than the MODIS-based results because of (a) reduced GOES spatial resolution and OTs being less prominent and (b) differences in validation criteria where MODIS detections were formerly considered a hit if they were up to 10 km away from a human-identified OT which was reduced to 5 km for this study. In addition, continued experience working with NWP and reanalysis-derived convective equilibrium level temperature, used in the Bedka and Khlopenkov method, has indicated that this field, even after spatial smoothing, can be quite noisy and in disagreement with IR BT observations, especially over relatively data poor regions where deep convection is frequent such as Africa, South America, South Asia, and ocean in general. The approach described in this paper relies on only tropopause temperature, which is much more spatially coherent than equilibrium level, after some smoothing is applied.

One might assume that OT detection performance metrics for GOES-13 would be the same as those from GOES-8 to -12 imager, considering that the pixel sizes and imager optics for these satellites are essentially identical. Figure 19 shows that GOES-12 imagery of Tropical Storm Alberto in June 2006 suffers from increased intra-scanline noise and “striping” compared with GOES-13 imagery of Hurricane Matthew in 2016. During the GOES-13 Science Test in 2006, Hillger and Schmit (2007) found that “the GOES-13 Imager striping is less than that on GOES-12, possibly due to the longer black-body look.” A longer black-body look would improve imager calibration and reduce striping. Some evidence of striping can be seen in GOES-13





**Figure 19.** (a) GOES-12 10.7  $\mu\text{m}$  IR brightness temperature image of Tropical Storm Alberto on June 12, 2006 at 11:02 UTC. (b) GOES-13 10.7  $\mu\text{m}$  IR brightness temperature image of Hurricane Matthew on October 1, 2016 at 04:45 UTC.

data as well, but the intra-line BT variations were only 0.5–1.0 K, compared with 1.0–3.0 K from GOES-12. GOES-12 striping is substantial enough to trigger OT candidate detection, and the 1–3 K magnitude is an appreciable fraction of the dynamic range of the GOES-13 anvil prominence that ranges from about  $-1$  to  $-14$  K (Figure 11). Therefore, it is critical to remove these stripes before generation of GOES climate data records of OT detections. Methods for de-striping have been developed at NASA LaRC and will be described in a future paper.

In general, a satellite imager can be affected by sensor thermal noise, spectral band crosstalk, inaccurate black body measurements, and other factors. For the IRW band of the Advanced Himawari Imager (AHI) instrument, sensor noise is reported to be about 0.45 K for a 200 K target (Ai et al., 2017). This estimation of the brightness temperature bias can be used to analyze our algorithm's uncertainty as a function of the input BT error. We selected three OT probability ranges, 25, 50, and 75, and used a simulated convection region, similar to the one shown in Figure 4, to estimate a variation in the OT probability, as calculated by Equation 6, in response to an artificial increase in the input BT by 0.5 K. As a result of this simulation, OT probability is decreased from 25.0 to 22.3, from 50.0 to 47.1, and from 75.0 to 73.0, respectively, which is considered to be relatively negligible.

## 7. Summary

This paper provides extensive theoretical background of an updated method for automated detection of overshooting cloud tops using a combination of spatial IR BT patterns that have been quantified in a variety of ways and NWP tropopause temperature. IR temperatures are converted to a tropopause-relative temperature, which serves as a stable reference that modulates how cold a convective cloud should become within a given region. Anvil clouds are identified using histogram analysis to generate an anvil rating that is used in subsequent phases of processing. Cold spots embedded within anvils serve as OT candidate regions. OT candidates are then assigned an OT probability and the spatial extent of OT cores is found.

The OT probability can be interpreted as a metric of storm intensity and an estimate of confidence in a detection for a particular pixel. It is produced using an original mathematical composition of four factors: Tropopause-normalized temperature, prominence relative to the surrounding anvil, surrounding anvil area, and spatial uniformity of anvil temperature, which are calculated from empirically derived sensitivity curves. Such a design that aggregates the four factors together, coupled with analysis of surrounding anvil properties, provides for a higher reliability and accuracy of the OT probability product, and also helps its temporal stability when dealing with 30-s to 1-min Mesoscale Domain Sector observations from the GOES-16/17 imager or other high-temporal resolution GEO imagers. The shape of the sensitivity curves is supported by independent analysis of a large sample of matched IR and NEXRAD-observed OT regions. An optimal sensitivity for each factor was determined by maximizing correlation between the OT probability and a set of human-identified OT regions. Coarser spatial resolution of GOES-13 data cause OTs to be less prominent compared to GOES-16, necessitating different sensitivities for each satellite. The statistical accuracy of OT

detection was also assessed by analyzing the probability of detection (POD) and false alarm ratio (FAR), which revealed notable improvements over the Bedka and Khlopenkov (2016) method.

Though deep learning pattern recognition methods have shown promise for OT detection (e.g., Cintineo et al., 2020; Kim et al., 2017), the approach described here, developed through years of experience and empirical testing, identifies and quantifies OT features at the individual satellite pixel scale similar to how a human analyst would perform such identification. Cintineo et al. (2020) seeks to identify the “probability of intense convection” within a  $64 \times 64$  km region using a combination of GOES-R series visible (when available), IR, and geostationary lightning mapping imagery. While their method can perform quite well, their detections encompass a spatial scale larger than typical OT regions and their scheme requires at least two of the imagery inputs above and cannot yet operate on IR imagery alone. Bedka and Khlopenkov (2016) demonstrated that filtering of IR-based detections using visible wavelength texture detection can further improve accuracy. Nevertheless, well-performing techniques based on IR imagery alone like the one described in this paper are extremely valuable for defining the climatology of intense convection at high temporal frequency throughout the diurnal cycle and studying their climate impacts.

Though our method seeks to identify features evident to a human in the imagery, there is no guarantee that precipitation echo tops that reach altitudes near to the tropopause will generate detectable BT perturbations. Cooney et al. (2021) quantifies the detectability of echo tops near to and above the tropopause, based on NEXRAD processing methods described in Cooney et al. (2018). In addition, consistency in detections between the GOES-13 Imager and GOES-16 ABI data is also quantified. Such consistency is critical to establish when one seeks to develop climate data records from the modern-era, 25+ year duration GOES satellite data record beginning with GOES-8 and continuing to the present with GOES-16 and -17, assuming sensor noise and calibration described above can be sufficiently addressed.

## Appendix A1: Image Reprojection and Boundary Data Processing

Each of the input images is remapped to the output projection by means of a so-called inverse mapping. First, the valid data boundary in the input image (satellite perspective) is constructed as a polygon defined by a set of vertices (typically 500), which are then remapped to the output projection. Then for each output location within the remapped boundary, its latitude and longitude are sought in the input latitude/longitude images. This is implemented by means of the concurrent gradient search (Khlopenkov & Trishchenko, 2008). This search yields a fractional position in terms of the input image coordinates, which is then used to interpolate the adjacent input values by means of a  $6 \times 6$  pixel resampling function. At the image boundaries, the missing contents of the  $6 \times 6$  pixel window is padded by replicating the edge pixel values. Here and further in these algorithms, image resampling operations are implemented as Lanczos filtering (Duchon 1979) extended to the 2D case with the parameter  $a = 3$ . This interpolation method is based on the sinc filter, which is known to be an optimal reconstruction filter for band-limited signals, e.g., digital imagery. The interpolated value is finally stored at the current location in the output image. This inverse mapping process has two main advantages: (a) it ensures that all output pixels are filled with valid data, and (b) the interpolation is performed in the input pixel/line space where samples are aligned in a regular grid, which allows for straightforward resampling even with higher order polynomials. Similar to the described reprojection from the satellite perspective to a geographic map, the algorithm also provides the capability to remap storm detection and characterization products back to the original satellite perspective, which can be useful for some applications.

As the algorithms below use various window-based filters, it is important to ensure that those operate correctly on pixels close to the boundary of valid data that may not necessarily be rectangular. This is achieved by first creating a per-pixel mask of valid data in the reprojected image. Then the valid data are extrapolated spatially by about 36 km beyond the boundary, which is implemented by replacing the out-of-boundary fill values with a window mean calculated from the nearby valid pixels. The window mean is obtained as a weighted average with the weights defined by a radial basis Gaussian function having  $\sigma = 3.2$  km. The fraction of valid pixels within the averaging window can be adjusted from 10% to 50%. Using a larger fraction results in higher output quality but requires more processing passes before the requested length of extrapolation is achieved. Smaller fractions allow for fewer passes but may result in noticeable striping

in the extrapolated areas. Once the spatial extents of the valid input data are expanded, all the subsequent processing is carried out using fixed-size spatial filters as if the image contained no fill values. This may produce false detections and other artifacts near the new (expanded) edges of the valid data, but those would occur in the extrapolated areas only and do not impact the valid pixels. At the end of processing, all the output images are screened out by the valid pixel mask obtained above (before the extrapolation) and so any near-boundary artifacts are replaced by fill values.

## Appendix A2: Spatial Expansion of the Anvil Mask

The anvil mask obtained from Equation 4 has to be expanded in order to ensure that it includes all pixels inside the circular window that have a high enough BT-score. First, a minimum threshold *MinAnvilScore* is introduced as:

$$\text{MinAnvilScore} = 8500 + 512(X_{\text{peak}} - 0.5) - 32r_{\text{anvil}} \quad (\text{A1})$$

where  $X_{\text{peak}}$  is the effective horizontal coordinate of the peak in the histogram:

$$X_{\text{peak}} = \frac{\sum_i i H_i}{\sum_i H_i} \quad (\text{A2})$$

Here, the index  $i$  again denotes the summation over the three highest histogram bins. The  $32r_{\text{anvil}}$  term subtracted in Equation A2 makes the *MinAnvilScore* somewhat smaller than the peak-equivalent BT-score in order to include pixels with BT-score lower than the effective peak's position. This allowance is made proportional to the anvil rating, which helps to include the anvil boundary pixels adjacent to very cold OT cores when the peak may be shifted to higher bins.

Once the *MinAnvilScore* is calculated, the anvil rating mask can be expanded by raising the anvil rating for pixels inside the circular window. Specifically, for pixels with BT-score higher than *MinAnvilScore* their anvil rating is increased to reach that of the current pixel. This effectively fills out all lower rated pixels around the current one. In addition, the anvil mask has to be refined in order to absorb some pixels (such as inside curved cloud edges) surrounded by many neighbors with high anvil rating. To achieve this, for any pixel inside the circular window having BT-score of at least 2/3 of the *MinAnvilScore* threshold, a dedicated counter associated with that pixel is incremented to indicate a high rating neighbor nearby. The amount of the increment equals the spatial resolution (km/pixel) squared, which corrects for lower pixel count at coarser resolution. In this way, the neighbor counters are incremented while the whole image undergoes the anvil mask expansion. On the second pass, if a pixel's  $r_{\text{anvil}}$  is still under 115 but that pixel's neighbor counter exceeds 130 (or 80 but its *BTscore* is over 11,000) then a sum of anvil ratings  $S_{\text{AR}}$  is calculated over pixels having *BTscore* > 10,000 within a circle of 14 km diameter around the current pixel. The current pixel's anvil rating is then set to  $S_{\text{AR}}/(N+1)$  where  $N$  is the number of pixels in the sum, which helps to slightly reduce the calculated average when  $N$  is low. The obtained image of the anvil rating is finally filtered by a Gaussian blur with  $\sigma = 2$  pixels in order to smooth out minor pixel-size artifacts caused by integer counting, resulting in the final anvil mask.

## Appendix A3: Spatial Expansion of OT Cores

To derive the spatial extent of OT cores, a temperature threshold  $BT_{\text{max}}$  is introduced that defines the highest level of BT for a nearby pixel to be included into the OT:

$$BT_{\text{max}} = BT_p + Z(\text{WinAvgBT} - BT_p) \cdot \text{SensOTsize} \cdot \text{TropopauseF} \cdot (\lambda + 0.1) \quad (\text{A3})$$

Here, *SensOTsize* is a configurable parameter controlling the sensitivity of OT size expansion with the typical range of 0.7–1.0. This sensitivity is additionally factored by the  $\lambda$  function, which reduces the OT expansion for candidates with weak prominence or those located inside uneven anvil areas. Without this correction, the OT may expand to an unreasonably wide region in a very cold cloud (high *TropopauseF*)

but weak spatial gradient (low *ProminenceF*). On the other hand, uneven areas are also often encountered within large clusters of very cold OTs, and then the second correction using the tropopause factor is added to help to capture smaller but still cold OTs when a more significant OT is located nearby.

With the  $BT_{max}$  threshold defined, the pixels surrounding the initial OT candidate are tested to have the BT below that threshold. Pixels that pass the test are added to the OT extents by flagging them with the same OT identification number, which is unique for each OT in each scene. This test is carried out along 16 rays cast off from the OT candidate pixels similarly to the process shown in Figure 8 but within an 8 km radius from the center. As a result, this routine ensures that the included pixels form a contiguous OT shape.

## Data Availability Statement

The IR OT detection data and locations of the human-identified OT signatures are available through NASA Langley Research Center at <https://science-data.larc.nasa.gov/LaRC-SD-Publications/2021-04-26-001-KMB/>. NEXRAD GridRad data for data analyzed to create Figure 11 are available at: <https://science-data.larc.nasa.gov/LaRC-SD-Publications/2021-04-29-001-KMB/>. Level 1b radiances from NOAA's GOES-R series satellite used in this study are available through cloud infrastructures such as Amazon Web Service: <https://registry.opendata.aws/noaa-goes/> or Google Cloud Platform: <https://console.cloud.google.com/launcher/details/noaa-public/goes-16> and <https://console.cloud.google.com/launcher/details/noaa-public/goes-17> or through NOAA's online subsetter: <https://www.ncei.noaa.gov/access/search/data-search/goesr-abi-level-1b-radiances>. Level 1b radiances from NOAA's GOES-M and GOES-N series satellites are available through: [https://www.avl.class.noaa.gov/saa/products/search?sub\\_id=0&datatype\\_family=G-VAR\\_IMG](https://www.avl.class.noaa.gov/saa/products/search?sub_id=0&datatype_family=G-VAR_IMG). Level 1b radiances from NASA's Suomi NPP VIIRS satellite are available through: <https://ladsweb.modaps.eosdis.nasa.gov/search/order/1/VNP02IMG--5110>. Tropopause data are available through NASA GMAO: <https://doi.org/10.5067/3Z173KIE2TPD>. Severe storm reports are available through NOAA's Storm Prediction Center: <https://www.spc.noaa.gov/climo/reports/>. See also data references below: Bowman and Homeyer, (2017), Global Modeling and Assimilation Office GMAO (2015), GOES-R (2017), NOAA (1994), VIIRS (2016).

## Acknowledgments

This work was supported by the NASA Applied Sciences Disasters program project award 18-DISASTER18-0008 and work within the NASA Earth Venture Suborbital DCOTSS mission. The authors thank Benjamin Scarino and Douglas Spangenberg for their valuable feedback and numerous contributions to this algorithm development effort. The authors also thank the Data Center within the University of Wisconsin-Madison Space Science and Engineering Center for providing the GOES data used in this study and for continued development of the McIDAS-X software package.

## References

- Ai, Y., Li, J., Shi, W., Schmit, T. J., Cao, C., & Li, W. (2017). Deep convective cloud characterizations from both broadband imager and hyperspectral infrared sounder measurements. *Journal of Geophysical Research - D: Atmospheres*, 122, 1700–1712. <https://doi.org/10.1002/2016JD025408>
- Apke, J. M., Mecikalski, J. R., Bedka, K. M., McCaul, E. W., Homeyer, C. R., & Jewett, C. P. (2018). Relationships between deep convection updraft characteristics and satellite-based super rapid scan mesoscale atmospheric motion vector-derived flow. *Monthly Weather Review*, 146(10), 3461–3480. <https://doi.org/10.1175/mwr-d-18-0119.1>
- Bedka, K. M., Brunner, J., Dworak, R., Feltz, W., Otkin, J., & Greenwald, T. (2010). Objective satellite-based overshooting top detection using infrared window channel brightness temperature gradients. *Journal of Applied Meteorology and Climatology*, 49(2), 181–202. <https://doi.org/10.1175/2009JAMC2286.1>
- Bedka, K. M., Brunner, J., & Feltz, W. (2011). Overshooting top and enhanced-V detection: Algorithm theoretical basis document. Retrieved from [https://www.goes-r.gov/products/ATBDs/option2/Aviation\\_OvershootingTop\\_v1\\_no\\_color.pdf](https://www.goes-r.gov/products/ATBDs/option2/Aviation_OvershootingTop_v1_no_color.pdf)
- Bedka, K. M., & Khlopenkov, K. (2016). A probabilistic pattern recognition method for detection of overshooting cloud tops using satellite imager data. *Journal of Applied Meteorology and Climatology*, 55, 1983–2005. <https://doi.org/10.1175/JAMC-D-15-0249.1>
- Bedka, K. M., Wang, C., Rogers, R., Carey, L., Feltz, W., & Kanak, J. (2015). Examining deep convective cloud evolution using total lightning, WSR-88D, and GOES-14 super rapid scan datasets. *Weather and Forecasting*, 30(3), 571–590. <https://doi.org/10.1175/WAF-D-14-00062.1>
- Berendes, T. A., Mecikalski, J. R., MacKenzie, W. M., Bedka, K. M., & Nair, U. S. (2008). Convective cloud identification and classification in daytime satellite imagery using standard deviation limited adaptive clustering. *Journal of Geophysical Research*, 113(D20). <https://doi.org/10.1029/2008JD010287>
- Bluestein, H. B., Lindsey, D. T., Bikos, D., Reif, D. W., & Wienhoff, Z. B. (2019). The relationship between overshooting tops in a tornadic supercell and its radar-observed evolution. *Monthly Weather Review*, 147(11), 4151–4176. <https://doi.org/10.1175/MWR-D-19-0159.1>
- Bosilovich, M. G., Lucchesi, R., & Suarez, M. (2016). MERRA-2: File specification. GMAO Office Note No. 9 (Version 1.1) 73. Retrieved from [http://gmao.gsfc.nasa.gov/pubs/office\\_notes](http://gmao.gsfc.nasa.gov/pubs/office_notes)
- Bowman, K. P., & Homeyer, C. R. (2017). GridRad - Three-dimensional gridded NEXRAD WSR-88D Radar Data Mar. 2017-Aug. 2017, 50°N, 100°W; 23°N, 65°W. *Research data archive at the national center for atmospheric Research, computational and information systems laboratory*. Boulder, Colo. (Updated irregularly). <https://doi.org/10.5065/D6NK3CR7>. Accessed 2020-02-01.
- Cintineo, J. L., Pavolonis, M. J., Sieglaff, J. M., Wimmers, A., Brunner, J., & Bellon, W. (2020). A deep-learning model for automated detection of intense mid-latitude convection using geostationary satellite images. *Weather and Forecasting*, 57, 2567–2588. <https://doi.org/10.1175/WAF-D-20-0028.1>
- Clapp, C. E., Smith, J. B., Bedka, K. M., & Anderson, J. G. (2019). Identifying source regions and the distribution of cross-tropopause convective outflow over North America during the warm season. *Journal of Geophysical Research: Atmosphere*, 124, 13750–13762. <https://doi.org/10.1029/2019JD031382>



- Cooney, J. W., Bedka, K. M., Bowman, K. P., Khlopenkov, K., & Irtterly, K. (2021). Comparing tropopause-penetrating convection identifications derived from NEXRAD and GOES over the contiguous United States. *Journal of Geophysical Research Atmospheres*, e2020JD034319. <https://doi.org/10.1029/2020JD034319?af=R>
- Cooney, J. W., Bowman, K. P., Homeyer, C. R., & Fenske, T. M. (2018). Ten-year analysis of tropopause-overshooting convection using GridRad data. *Journal of Geophysical Research: Atmospheres*, 123, 329–343. <https://doi.org/10.1002/2017JD027718>
- Duchon, C. E. (1979). Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology*, 18(8), 1016–1022. [https://doi.org/10.1175/1520-0450\(1979\)018<1016:LFIOAT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1979)018<1016:LFIOAT>2.0.CO;2)
- Global Modeling and Assimilation Office (GMAO) (2015). *MERRA-2 tavg1\_2d\_slv\_Nx: 2d,1-Hourly, Time-Averaged, Single-Level, Assimilation, Single-Level Diagnostics V5.12.4, March 2017–August 2019, 50°N, 120°W; 15°N, 60°W*. Greenbelt, MD: Goddard Earth Sciences Data and Information Services Center (GES DISC). <https://doi.org/10.5067/3Z173KIE2TPD>
- GOES-R Calibration Working Group, & GOES-R Series Program (2017). *NOAA GOES-R Series Advanced Baseline Imager (ABI) Level 1b Radiances, March 2017–August 2019, 50°N, 120°W; 15°N, 60°W*. NOAA National Centers for Environmental Information. <https://doi.org/10.7289/V5BV7DSR>
- Hillger, D. W., & Schmit, T. J. (2007). *The GOES-13 science test: Imager and sounder radiance and product validations*. NOAA Technical Report. NESDIS 125. Retrieved from <https://repository.library.noaa.gov/view/noaa/1132>
- Homeyer, C. R., & Bowman, K. P. (2017). Algorithm description document for version 3.1 of the three-dimensional gridded NEXRAD WSR-88D radar (GridRad) dataset. Retrieved from <http://gridrad.org/pdf/GridRad-v3.1-Algorithm-Description.pdf>
- Khlopenkov, K. V., & Trishchenko, A. P. (2008). Implementation and evaluation of concurrent gradient search method for reprojection of MODIS level 1B imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 46(7), 2016–2027. <https://doi.org/10.1109/TGRS.2008.916633>
- Kim, M., Im, J., Park, H., Park, S., Lee, M.-I., & Ahn, M.-H. (2017). Detection of tropical overshooting cloud tops using Himawari-8 imagery. *Remote Sensing*, 9(7), 685. <https://doi.org/10.3390/rs9070685>
- Lazzara, M. A., Benson, J. M., Fox, R. J., Laitsch, D. J., Rueden, J. P., Santek, D. A., et al. (1999). The man computer interactive data access system: 25 years of interactive processing. *Bulletin of the American Meteorological Society*, 80(2), 271–284.
- Liu, N., Liu, C., & Hayden, L. (2020). Climatology and detection of overshooting convection from 4 years of GPM Precipitation Radar and passive microwave observations. *Journal of Geophysical Research: Atmosphere*, 125. <https://doi.org/10.1029/2019JD032003>
- Mikuš, P., & Mahović, N. S. (2013). Satellite-based overshooting top detection methods and an analysis of correlated weather conditions. *Atmospheric Research*, 123. <https://doi.org/10.1016/j.atmosres.2012.09.001>
- NOAA Office of Satellite and Product Operations. (1994). *NOAA Geostationary Operational Environmental Satellite (GOES) I-M and N-P Series Imager Data, March 2006–August 2017, 50°N, 120°W; 15°N, 60°W*. NOAA National Centers for Environmental Information. <https://doi.org/10.25921/Z9JQ-K976>. Accessed 2020-02-01.
- Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal*, 7(2), 155–162. <https://doi.org/10.1093/comjnl/7.2.155>
- Reynolds, D. W. (1980). Observations of damaging hailstorms from geosynchronous satellite digital data. *Monthly Weather Review*, 108(3), 337–348. [https://doi.org/10.1175/1520-0493\(1980\)108<0337:OODHFG>2.0.CO;2](https://doi.org/10.1175/1520-0493(1980)108<0337:OODHFG>2.0.CO;2)
- Rudlosky, S. D., Goodman, S. J., Virts, K. S., & Bruning, E. C. (2019). Initial geostationary lightning mapper observations. *Geophysical Research Letters*, 46(2), 1097–1104. <https://doi.org/10.1029/2018GL081052>
- Sandmæl, T. N., Homeyer, C. R., Bedka, K. M., Apke, J. M., Mecikalski, J. R., & Khlopenkov, K. V. (2019). Evaluating the ability of remote sensing observations to identify significantly severe and potentially tornadic storms. *Journal of Applied Meteorology and Climatology*, 58(12), 2569–2590. <https://doi.org/10.1175/jamc-d-18-0241.1>
- Scarino, B. R., Bedka, K., Bhatt, R., Khlopenkov, K. V., Doelling, D. R., & Smith, W. L. Jr. (2020). A kernel-driven BRDF model to inform satellite-derived visible anvil cloud detection. *Atmospheric Measurement Techniques*, 13(10), 5491–5511. <https://doi.org/10.5194/amt-13-5491-2020>
- Schmetz, J., Tjemkes, S. A., Gube, M., & van de Berg, L. (1997). Monitoring deep convection and convective overshooting with Meteosat. *Advances in Space Research*, 19(3), 433–441. [https://doi.org/10.1016/S0273-1177\(97\)00051-3](https://doi.org/10.1016/S0273-1177(97)00051-3)
- Schmit, T. J., Griffith, P., Gunshor, M. M., Daniels, J. M., Goodman, S. J., & Lebar, W. J. (2017). A closer look at the ABI on the GOES-R Series. *Bulletin of the American Meteorological Society*, 98(4), 681–698. <https://doi.org/10.1175/BAMS-D-15-00230.1>
- Smith, J. B., Wilmoth, D. M., Bedka, K. M., Bowman, K. P., Homeyer, C. R., Dykema, J. A., et al. (2017). A case study of convectively sourced water vapor observed in the overworld stratosphere over the United States. *Journal of Geophysical Research Atmospheres*, 122(17), 9529–9554. <https://doi.org/10.1002/2017JD026831>
- Solomon, D. L., Bowman, K. P., & Homeyer, C. R. (2016). Tropopause-penetrating convection from three-dimensional gridded NEXRAD data. *Journal of Applied Meteorology and Climatology*, 55, 465–478. <https://doi.org/10.1175/JAMC-D-15-0190.1>
- Takahashi, H., & Luo, Z. J. (2014). Characterizing tropical overshooting deep convection from joint analysis of CloudSat and geostationary satellite observations. *Journal of Geophysical Research - D: Atmospheres*, 119, 112–121. <https://doi.org/10.1002/2013JD020972>
- VIIRS Characterization Support Team. (2016). VIIRS/NPP imagery resolution 6-Min L1B swath 375m. *March 2017–August 2019, 50°N, 120°W; 15°N, 60°W. L1 and Atmosphere Archive and Distribution System (LAADS)*. <https://doi.org/10.5067/VIIRS/VNP02IMG.001>
- WMO. (1957). Meteorology - A three-dimensional science: Second session of the commission for aerology. *World Meteorological Organization Bulletin*, VI(4), 134–138.
- Xian, T., & Homeyer, C. R. (2019). Global tropopause altitudes in radiosondes and reanalyses. *Atmospheric Chemistry and Physics*, 19, 5661–5678. <https://doi.org/10.5194/acp-19-5661-2019>
- Yost, C. R., Bedka, K. M., Minnis, P., Nguyen, L., Strapp, J. W., Palikonda, R., et al. (2018). A prototype method for diagnosing high ice water content probability using satellite imager data. *Atmospheric Measurement Techniques*, 11, 1615–1637. <https://doi.org/10.5194/amt-11-1615-2018>
- Zisper, E., Cecil, D. J., Liu, C., Nesbitt, S. W., & Yorty, D. P. (2006). Where are the most intense thunderstorms on earth? *Bulletin of the American Meteorological Society*, 87(8), 1057–1072. <https://doi.org/10.1175/BAMS-87-8-1057>