The Illumination of Thunderclouds by Lightning:

Part 3: Retrieving Optical Source Altitude

Michael Peterson[1], Tracy E. L. Light[1], Douglas Mach[2]

[1] ISR-2, Los Alamos National Laboratory, Los Alamos, New Mexico

[2]Science and Technology Institute, Universities Space Research Association, Huntsville, AL, USA

Corresponding author: Michael Peterson (mpeterson@lanl.gov), B241, P.O. Box 1663 Los Alamos, NM, 87545

Key Points:

- Machine learning is employed to predict the source altitude for GLM groups from attributes of their spatial optical energy distributions
- GLM altitude models predict the matched LMA mean source altitudes with a median absolute error of < 1.5 km
- The models capture changes in vertical LMA source distributions from convective invigoration or maturation and resolve vertical flash extent

LA-UR-21-23257

Abstract

Optical space-based lightning sensors such as the Geostationary Lightning Mapper (GLM) detect and geolocate lightning by recording rapid changes in cloud-top illumination. While lightning locations can be determined to within a pixel on the GLM imaging array, these instruments are not individually able to natively report lightning altitude. It has previously been shown that thunderclouds are illuminated differently based on the altitude of the optical source. In this study, we examine how altitude information can be extracted from the spatial distributions of GLM energy recorded from each optical pulse. We match GLM "groups" with LMA source data that accurately report the 3-D positions of coincident Radio-Frequency (RF) emitters. We then use machine learning methods to predict the mean LMA source altitudes matched to GLM groups using metrics from the optical data that describe the amplitude, breadth, and texture of the group spatial energy distribution. The resulting model can predict the LMA mean source altitude from GLM group data with a median absolute error of < 1.5 km, which is sufficient to determine the location of the charge layer where the optical energy originated. This model is able to capture changes to the source altitude distribution following convective invigoration or maturation, and the GLM predictions can reveal the vertical structure of individual flashes - enabling 3-D flash geolocation with GLM for the first time. Additional work is required to account for differences in thunderstorm charge / precipitation structures and viewing angle across the GLM Field of View.

47

48 **Plain Language Summary**

49 Lightning is detected from space by monitoring the Earth for rapid changes in clou-top
50 illumination. We can determine where the lightning occurred from the location of the pixel that
51 was triggered. However, since we're looking down at the Earth from above the cloud tops, there
52 is no simple way to determine the altitude of the lightning flash with this kind of instrument, and
53 this is a significant limitation of sensors like the Geostationary Lightning Mapper (GLM).
54 This study uses machine learning methods to attempt to predict lightning altitude from
55 the spatial distribution of energy across the cloud illuminated by each optical pulse. We find that
56 it is possible to predict source altitude well enough to determine which charge layer an optical
57 pulse originated from, and also identify changes in storm structure over time and the vertical
58 development of individual flashes. While these results are still preliminary and come from a
59 single thunderstorm, they demonstrate that altitude prediction is possible with GLM and
60 additional work could result in a general prediction model for all observations by GLM and
61 legacy instruments.

1 Introduction

The optical lightning imagers that have been operated in Low Earth Orbit (LEO) by NASA and geostationary orbit (GEO) by NOAA record rapid changes in cloud-top illumination caused by lightning within the cloud medium (Christian et al., 2000). As these instruments are pixelated, the horizontal extent of lightning can be determined by projecting the footprint of each pixel on the imaging array to an ellipsoid above the Earth's surface. The chosen ellipsoid should correspond to the upper boundary of the cloud that the optical emissions transmit through, otherwise parallax will be introduced into the GLM measurements (Virts and Koshak, 2020). However, these optical measurements are only a composite two-dimensional view of lightning that describes its geospatial distribution across the Earth (Christian et al., 2003; Cecil et al., 2014; Albrecht et al., 2016) and the horizontal extent of individual flashes (Peterson et al., 2018; Lyons et al., 2020; Peterson et al., 2020). The third dimension – source altitude – is not resolved natively by these instruments, and this is considered one of their primary shortcomings compared to certain ground-based lightning measurements.

Lightning source altitude is an important parameter because it provides unique insights into the intensity of convective systems and how thunderstorm kinematics organize charge regions within the thunderstorm (Williams, 1989; Smith et al., 2004; Carey et al., 2005; Ely et al., 2008; Stolzenburg and Marshall, 2008; Bruning et al., 2010;). Non-Inductive Charging (NIC: Reynolds et al., 1957; Takahashi, 1978; Jayaratne et al., 1983; Saunders et al., 1991; Saunders and Peck, 1998; Takahashi and Miyawaki, 2002; Mansell et al., 2005; Bruning et al., 2014) is considered to be a primary mechanism for creating the charge separation in thunderstorms that leads to lightning activity. Under the NIC model, collisions between different species of ice particles within the updraft cause a net transfer of charge (usually from small ice particles

depositing electrons on larger graupel pellets rimed with supercooled liquid water). These ice particles are then sorted according to their masses with the smaller ice particles lofted by the updraft towards the cloud top while the heavier graupel remains in the mid-levels of the storm. Over time, accumulation of charged ice particles at different altitudes produces a strong electric field that can overcome the electrical impedance of the air to generate a lightning discharge.

If we can resolve the vertical profile of lightning sources, then we can determine the altitudes of these charge regions and track how they change over time. Presently, lightning is related to convective intensity and thunderstorm microphysics through lightning rates (Blyth et al., 2001; Cecil et al., 2005; Prigent at al., 2005; Takayabu et al. 2006; Xu et al. 2010; Liu et al. 2011; Peterson and Liu, 2011; Liu et al. 2012) because this information is widely available across broad geospatial domains. Altitude information is only reported on local or regional scales by dense networks of ground-based instruments that detect Radio-Frequency (RF) lightning emissions. The most accurate three-dimensional source information is provided by Lightning Mapping Arrays (LMAs: Rison et al., 1999) whose effective range is limited to just a few hundred kilometers. The only truly global lightning network that attempts to resolve altitude is the Earth Networks Global Lightning Network (ENGLN: Zhu et al., 2017), but their intracloud (IC) altitude parameter is not well refined, leading to highly-inaccurate results (Peterson et al., 2021a).

If accurate lightning altitudes could be provided across large swaths of the Earth, it would add a new dimension to discussions of the connection between lightning and impactful weather. Convective invigoration has been linked to the onset of severe weather (such as hail, tornadoes, derechos) (Schultz et al., 2009; Gatlin and Goodman, 2010), and is also considered important for hurricane Rapid Intensification (RI) (DeMaria et al., 2012; Jiang and Ramirez, 2013; Fierro et

al., 2018). These studies look for convective invigoration by tracking how flash rates change as the storm develops over time. Rapid increases in source altitude would provide an alternate means to identify strengthening updrafts that could either confirm the flash rate trend or potentially catch events that are missed due to poor instrument performance. Geostationary Lightning Mapper (GLM: Goodman et al., 2013; Rudlosky et al,. 2019) total flash rates are adversely affected by attenuation from optical sources transmitting through thick cloud layers, over-clustering in high flash rate compact thunderstorms, and artificial flash splitting in non-convective flashes. The first and third issues can also be amplified by a high instrument threshold , as we saw in Part 2 of this series (Peterson et al., 2021b). However, none of these issues would prevent the highest-altitude sources from being resolved from space.

We propose that altitude information can be extracted from GLM measurements of how the surrounding thunderclouds are illuminated by lightning. Our previous modelling work (Peterson, 2020a) demonstrated that low-altitude sources result in different spatial radiance patterns than high-altitude sources regardless of cloud geometry, and this was confirmed with GLM observations in and Part 1 of this series (Peterson et al., 2021a). Our discussion of "optical repeater" flashes in Peterson et al., 2021a and previous analyses of groups with complex spatial radiance distributions (Peterson 2020b) further showed that radiance patterns were consistent between subsequent illuminations of the same cloud layer. However, these pictures of cloud illumination would change if the flash moved into a different layer – for example, cases in Peterson et al., 2021a where the LMA sources developed vertically.

In this third part of our thundercloud illumination study, we investigate whether the link between source altitude and the spatial radiance patterns recorded by GLM is sufficiently robust that we might predict the altitudes of the optical sources responsible for arbitrary GLM groups

131 that consist of more than one event. To accomplish this, we will construct a new set of group
132 metrics that describe the spatial distribution of GLM-recorded energy and then use a random
133 forest generator to construct a machine learning model to predict the mean altitude of coincident
134 LMA sources associated with each group. These predictions will be analyzed to determine
135 whether GLM-retrieved altitudes can resolve the major features of the LMA source altitude
136 distribution from the thunderstorm and the vertical development of individual flashes mapped by
137 both GLM and the LMA. We limit our analysis to a single thunderstorm case (the Colombia case
138 from Peterson et al., 2021a and Peterson et al., 2021b) to demonstrate the feasibility of this
139 approach, and leave validation across multiple storm types for future work.
140

2 Data and Methodology

This third part of our thundercloud illumination study will leverage the combined Geostationary Operational Environmental Satellites (GOES)-16 GLM and ground-based Colombia LMA (COLLMA: Lopez et al., 2016; Aranguren et al., 2018) data generated in Part 1 (Peterson et al., 2021a) and the random forest regressor in the Python scikit-learn machine learning module (Pedregosa et al., 2011) to generate a random forest model for predicting the mean LMA source altitude associated with each GLM group from a thunderstorm of interest. Section 2.1 discusses the lightning measurements that we will consider. Section 2.2 describes how the feature and label data that will be input into the machine learning model are generated. Finally, Section 2.3 documents the random forest regression.

2.1 Combined LMA / GOES Measurements of a Colombia Thunderstorm

In the first two parts of this study (Peterson et al., 2021a,b), we examined a thunderstorm on 01 November 2019 that occurred in the vicinity of Barrancabermeja in central Colombia that was measured by both the COLLMA and GLM. This storm is noteworthy because it contained a diverse collection of convective and non-convective lightning, was located near the GOES-16 satellite subpoint, and was subject to particularly-low GLM instrument thresholds (~0.7 fJ) that allowed GLM to resolve more detail from its flashes and their illumination of the surrounding clouds than thunderstorms elsewhere in the GLM Field of View (FOV).

2.1.1 Colombia Lightning Mapping Array (COLLMA) Data

COLLMA is a 6-sensor LMA network that was moved to Barrancabermeja from Santa Marta in 2018. LMA sources collected by the COLLMA on 01 November 2019 were provided

by Lopez (2020, personal communication) over a 1.7º longitude (74.5º W – 72.8º W) by 1º degree latitude (6.5 º N – 7.5º N) box within the LMA domain for comparison with GLM. The source data were first processed by Lopez (2020, personal communication) using the flash clustering and noise reduction algorithms developed by van der Velde and Montanyà (2013). These algorithms identify noise sources based on their density in 3D space-time boxes with sides corresponding to the horizontal distance (XY), vertical distance (Z), and time difference (T). Source densities that do not meet their empirically-derived thresholds are not clustered into flashes and we only consider those LMA sources that meet the threshold values.

2.1.2 Earth Networks Global Lightning Network (ENGLN) Data

The COLLMA source data is augmented with ENGLN detections of CG strokes during the thunderstorm of interest. ENGLN combines observations from the Earth Networks Total Lightning Network (ENTLN: Zhu et al., 2017) and the World-Wide Lightning Location Network (WWLLN: Lay et al., 2004; Rodger et al., 2006; Jacobson et al., 2006; Hutchins et al., 2012) to detect and geolocate both CG and IC lightning. However, since we have the LMA for IC sources, we do not consider ENGLN ICs.

2.1.3 Geostationary Lightning Mapper (GLM) Data

GLM is the first lightning imager to be operated from geostationary orbit. It builds on the legacy of NASA's Optical Transient Detector (OTD: Christian et al., 2003) and Lightning Imaging Sensor (LIS: Christian et al., 2000; Blakeslee et al., 2020) imagers that have been flown in LEO over the past 25 years. These instruments consist of a Charge Coupled Device (CCD) imaging array behind the instrument optics, which includes a narrowband filter centered on the 777.4 nm Oxygen emission line triplet. The dissociation, excitation, and recombination

experienced by the atmospheric constituent gasses in response to the intense heating of the lightning channels cause strong emissions at these atomic lines, which permits lightning to be detected at all times of day, albeit with decreased sensitivity under sunlit conditions.

The basic unit of OTD / LIS / GLM detection is the "event," which is defined as a single pixel on the imaging array that exceeds the instrument threshold during a single integration frame. Events are clustered by the GLM Lightning Cluster Filter Algorithm (LCFA: Goodman et al., 2010) into "group" features that describe simultaneous emission over a contiguous area on the imaging array, and "flash" features that use close spatial and temporal group proximity to approximate complete and distinct single lightning flashes. We further define a feature level between groups and flashes to document persistent illumination over multiple quasi-sequential integration frames called "series" features (Peterson and Rudlosky, 2019). Our reprocessed data that includes these features and other improvements are available at Peterson (2021a).

2.1.3 Matching RF data to GLM Groups and Flashes

The matching scheme that we employ in this study is based on the GLM / ENGLN matching algorithm used in Peterson and Lay (2020). It works under the assumption that all RF emissions within the footprint of a GLM group contribute optical energy to that group. Thus, these RF sources can be considered "events" in the GLM sense and clustered into the GLM data hierarchy as children of groups. Groups are nominally assigned the contemporary LMA sources or ENGLN CG strokes that occur within their footprint. However, this approach is subject to the three important caveats discussed below.

The first caveat is due to what groups actually represent. While groups are intended describe individual optical pulses, this association is far from perfect. Optical pulses are

generally quick and localized – with durations shorter than a millisecond and extents smaller than an 8-km GLM pixel. In Peterson et al. (2021a), we saw that the active portions of the lightning channel as mapped by the LMA were typically around 2 km in lateral extent. Yet, multi-event groups are common, with the largest groups even illuminating cloud areas exceeding 10,000 km^2 (Peterson et al., 2017). Sources located near pixel boundaries (Appendix B in Zhang et al., 2020) explains how GLM groups are larger than LMA source extents in certain scenarios, but it does not explain how GLM flash footprints can exceed the LMA flash extent or encapsulate cloud regions that do not appear to be electrified. These oddities in the GLM data result from scattering in the cloud medium. Multiple scattering causes the optical emissions – even from a point source - to be spread laterally throughout the surrounding thunderclouds (Peterson, 2020a), causing the resulting GLM group footprints to overestimate the physical extent of the source. At the same time, radiative transfer effects can also cause groups to underestimate the scale of the lightning source if the cloud is able to block radiant energy from reaching orbit. In extreme cases, particularly opaque clouds generate "holes" in the group footprint where the cloud regions surrounding the poorly-transmissive cloud are illuminated while its center remains dark and free of events (Peterson, 2020b).

Of these two possibilities, groups underestimating the extent of the optical sources involved is the primary concern for this work. In these cases, we might not have a full picture of the altitudes of the charge layers that contributed optical energy to the group. We saw in Peterson et al. (2021a) that even in the larger groups, the extent of LMA sources within their footprints were either of comparable size to a GLM pixel or smaller. To include RF sources in the vicinity of GLM groups that do not occur within their footprints, we add a 10-km buffer to the group

assignment criteria. RF events are assigned to a GLM group if they occur within 10 km of any event that comprised that group.

The second caveat is that the RF sources might not be precisely aligned in time with the parent GLM groups. This can happen if the source occurs at the end of a 2-ms GLM integration frame, causing the optical energy to be split between two adjacent frames, or in long-lasting processes such as return stroke Continuing Current (CC) or in-cloud K-changes (Bitzer, 2017). The LMA might not even register impulsive sources if the channel remains ionized during one of these long-duration processes since RF emissions describe changes in current rather than current. Thus, the reported time of the RF event might be separated from the time of peak optical emission by a few milliseconds. Moreover, in these cases, there could be multiple GLM groups that the RF events could be assigned to. In these scenarios, we attempt to assign RF events to the peak of the light curve recorded by GLM. All GLM groups that meet the spatial matching criteria for the RF event and occur within 10 ms of the event are identified, and the brightest GLM group is selected for assignment.

The third and final caveat is related to the limited domain of the available LMA data. Because the LMA data were provided over a latitude / longitude box, there are cases of GLM flashes along the edges of the LMA box where some groups contain LMA matches while others do not. As in the previous parts of this study, we limit our analyses to flashes whose groups were entirely within the LMA box to mitigate biases from partial matches at the edges of the LMA domain. The end result is a combined GLM / RF dataset consisting of 2154 GLM flashes and 56,399 groups. Of these flashes, 471 (21.9%) contained ENGLN strokes and 90.1% matched with LMA sources. Of these groups, 631 (1.1%) matched with ENGLN strokes and 22,681

(40.2%) matched with LMA sources. See Table 1 in Peterson et al., 2021a for additional GLM/RF matching statistics.

2.2 Generating Machine Learning Feature (Input) and Label (Prediction) Data

We propose that the first GLM caveat listed above - of groups primarily describing thundercloud illumination rather than the geometry of the optical source - is key to retrieving altitude information optically. As optical signals traverse the cloud medium to the satellite, they become modified through absorption and scattering in the cloud. Even the same optical sources located at different altitudes would take on a different appearance to GLM based on the optical characteristics of the cloud medium along the paths their emissions traveled to the instrument. By interpreting the spatial energy distributions of GLM groups (termed “radiance patterns”), we are attempting to decode the cloud attributes contained within the optical lightning signals.

2.2.1 Radiance Patterns from High-Altitude and Low-Altitude Sources

The key mechanism behind the differences in appearance between low-altitude sources and high-altitude sources is the number of scattering interactions that the optical emissions encounter before reaching the satellite. The emissions from low-altitude sources experience more scattering events than high-altitude sources, which permit the optical energy to be spread over a larger area. As a result, the radiance patterns from modeled sources (Peterson, 2020a) are broader with a lower amplitude for low-altitude cases, and brighter and more concentrated when the source is placed near the cloud top.

We can see these trends in groups observed by GLM. Figures 1 and 2 show two examples of GLM groups from the Colombia thunderstorm that the COLLMA determined to be comprised

of primarily low-altitude sources between 5 and 10 km (Figure 1), and high-altitude sources around 15 km (Figure 2). Both figures are formatted following the convention of Figures 10-12 in Peterson et al., 2021a with a central panel (d) showing the normalized group radiance pattern (dark indicating low energy, light indicating high energy) with LMA sources (green boxes) and ENGLN strokes (asterisks where blue denotes -CGs and red denotes +CGs) overlaid. Plus symbols (+) also indicate the locations of events to clarify which pixels are illuminated. The upper panels show the longitude-altitude LMA / ENGLN source profiles in (c) and GLM energy distribution by longitude in (a). The bars in (a) denote totals, while plus symbols describe individual events. The panels to the right of the plan view in (d) repeat these two plots for latitude. The bottom two plots show timeseries of LMA / ENGLN altitude (g) and GLM group energy (i) along with a LMA altitude distribution for the full 15-minute period that contained the flash (h). Finally, the upper right panel (b) shows the GLM group area / group maximum event energy distribution for the flash with a polynomial fit overlaid and its reduced chi^2 value listed. Groups are color coded in (i) and (b) according to their order in the flash (dark: early, light: late) and the current group is indicated with a dashed line in the timeseries and as a red symbol in the energy / area distribution.

The group shown in Figure 1 corresponded to the second ENGN -CG from the flash. The GLM radiance pattern was broad – with events exceeding 10% of the maximum event energy occurring in 7 of the 8 columns and 6 of the 7 rows on the GLM CCD array spanned by the group footprint. The group area / max. energy curve in Figure 1b also shows that subsequent groups illuminated the surrounding cloud in the same way, such that group area could be predicted from maximum event energy following the polynomial fit. By comparison, the energy from the group in Figure 2 is highly-concentrated in the single brightest event. Despite being half

294 the size of the group in Figure 1, the peak energy of the high-altitude group in Figure 2 reached
295 200 fJ (compared to 30 fJ in Figure 1) and only two other events in the group (immediately to the
296 north and west of the brightest group) exceeded 10% of the maximum event energy. This is the
297 same behavior that we saw previously during GLM flashes that produced Gigantic Jets (GJ),
298 (Boggs et al., 2019), which extend upward from the cloud top. The GLM energy was not only
299 concentrated in a single pixel co-located with the GJ, but this pixel remained illuminated over
300 many frames during the GJ.

301 2.2.2 Selecting the Prediction Altitude

302 The flash case in Figure 1 demonstrates a key challenge for predicting the source altitude:
303 even through the flash acts like a confined feature in how it illuminates the cloud (Figure 1b), the
304 LMA source altitudes associated with individual groups range from 5 km to 10 km (or from the
305 ground in the case of the -CGs). Assigning a single altitude to optical sources that have a finite
306 vertical dimension is a difficult proposition. Any altitude that we select for this type of optical
307 source will be subject to biases from our assumptions of where the peak currents are located and
308 how we quantify GLM's detection advantage for higher-altitude sources. For example, we might
309 assume that peak emission occurs where the branches come together near the ground in this -CG
310 case – and thus the minimum LMA altitude would be the best choice. Or we might assume that
311 low-altitude sources are severely attenuated based on the previous modeling work in Peterson
312 (2020a), so the in-cloud emissions described by either the mean or maximum LMA source
313 altitude better represent the optical source altitude. We know from Peterson et al., 2021a that
314 GLM favors detecting sources near the cloud-top in the Colombia thunderstorm, and this can be
315 verified by comparing the vertical distributions of all LMA sources in Figure 3a to the
316 distribution of mean LMA altitude for all sources matched to a GLM group in Figure 3b over the

thunderstorm duration. These two panels show that GLM has difficulty detecting optical emissions from low-altitude sources (< 7 km) – particularly around 09:00 UTC and in the 10:00 UTC hour. If GLM does not detect these low-altitude sources, then we will not be able to include them in the retrieved GLM altitude distributions. Even if the algorithm performs very well, there will still be biases in the GLM-derived vertical altitude distributions from these missed events. As this is a particularly-complex issue that requires further investigation, we will choose to predict the LMA mean altitude for the groups that were detected here and accept biases from poor characterization of low-altitude sources as a potential source of error. A different method to derived the prediction altitude or normalization strategies to account for missed events can always be considered in future studies to mitigate this issue.

The other key challenge for predicting source altitude with GLM is that these altitudes are determined by top-down measurements of cloud illumination rather from the ground-up view provided by the LMA. Thus, the appearance of the group will depend on the cloud layers between the optical source and the local cloud-top height. This is not a new issue for GLM, whose observations are commonly interpreted under the assumption that the optical illumination is contained within the boundaries of the thunderstorm core where the local cloud-tops approximately reach the height of the tropopause (Virts et al., 2020). The true "detection altitude" where the light leaves the cloud might be taller or shallower than the prescribed ellipsoid altitude, and this results in parallax errors in GLM geolocations (Virts et al., 2020). Thundercloud illumination as viewed from space depends on the depth of cloud between the source altitude and the detection altitude. If we attempted to directly predict the altitude of the LMA measurements or predict an altitude normalized to the GLM ellipsoid, the resulting predictions would be subject to similar biases. These predictions might be reasonable for the

most active period of the storm in question, but performance is expected to suffer outside of this period or outside of the convective core.

This issue might be addressed by normalizing the LMA source altitudes to the local cloud-top height. The Advanced Baseline Imager (ABI) Cloud Top Height (CTH) product is an attractive choice because ABI is on the same satellite as GLM and has a similar FOV. However, relying on ABI CTH data introduces a number of additional caveats. The ABI Cloud Height Algorithm (ACHA) is an operational algorithm based on joint measurements from the ABI infrared bands (CH14: 11.2 μm, CH15: 12.3 μm, and CH16: 13.3 μm), and its CTH estimates are subject to the uncertainties described in its Algorithm Theoretical Basis Document (ATBD) (Heidinger, 2012) and the less frequent sampling interval of ABI (10 minutes) relative to GLM (20 seconds). Perhaps the largest uncertainty for our application is its reliance on linear interpolations of temperature profiles supplied by Numerical Weather Prediction (NWP) models. These errors are then compounded by any parallax or location uncertainty in the LMA data being normalized (i.e., from lingering noise sources) where large CTH gradients exist.

The effect of these uncertainties on the LMA CTH normalization is shown in the timeseries of GLM-matched mean LMA source altitude in Figure 3b-e that span the duration of the Colombia thunderstorm. Figure 3b and d show the LMA measured altitudes, while Figure 3c and e show the CTH normalizations. Figure 3b and c contain all matched GLM groups while Figure 3d and e examine only the larger groups that consist of >5 GLM events. Both normalized timeseries contain activity above the ABI CTH (100%), and this activity is particularly common early in the storm (02:15 UTC - 07:30 UTC). As we showed in Peterson et al., 2021a (i.e., Figure 1), this time period corresponded to the thunderstorm moving into the area. As a result, much of

362 the activity contained within the LMA data domain occurred at the edge of the encroaching ABI
363 cold cloud feature (CH14 < 234 K) where strong gradients in ABI CTH exist.

364 If the optical emissions are able to more easily illuminate the storm edge than the dense
365 convective core, the group centroids in these edge cases can be located within the CTH gradient
366 region. While the LMA sources within the thunderstorm core might still be below their local ABI
367 CTH, the group centroid displaced towards the edge of the storm could be above its local ABI
368 CTH. This effect is particularly important with the densest thunderstorms where only edge
369 illumination is resolved by GLM (as in some cases noted in Peterson et al., 2021a from the
370 Colorado thunderstorm). Thus, while these apparent "above-cloud" sources might not make
371 intuitive sense, they are still a valuable inclusion in the dataset for describing this scenario that is
372 frequently encountered with GLM measurements.

373 2.2.3 Describing Radiance Patterns with Group-Level Metrics

374 A key strength of machine learning is that it can help to determine which combinations of
375 input parameters (features) best predict the parameters of interest (labels). In total, we have
376 devised 16 parameters in Table 1 that could be important for predicting altitude – 14 metrics that
377 describe the groups, and 2 series / flash level metrics that describe the context in which they
378 occur. The example groups in Figure 1 and Figure 2 provide guidance on some of the ways that
379 recorded radiance patterns from low-altitude sources and high-altitude sources differ, but these
380 differences could be quantified in many ways. We could focus on the spatial concentration of
381 energy or on the relationship between group area / energy (as discussed in Peterson et al.,
382 2021a). Alternatively, radiance anomalies including "holes" in GLM groups might provide better
383 predictors of source altitude.

Intuition based on data is an important place to start determining which parameters should be used in the analysis. For example, Figure 4 compares the percent of the group energy in the brightest event (GROUP_MAX_EVENT_PCT) with the overall group energy (GROUP_ENERGY). A two-dimensional histogram of GLM/LMA matches is shown in (a), the mean LMA altitude is shown in (b), the number of matches that describe ENGLN strokes is shown in (c), and the percent of all matches that originate at high altitudes (> 10 km) is plotted in (d). These plots show a clear distinction in source altitude with low-altitude sources at GROUP_MAX_EVENT_PCT < 25% and source altitudes increasing with GROUP_ENERGY and GROUP_MAX_EVENT_PCT. Most of the ENGLN strokes that occur in the matched GLM/LMA groups are also located along the bottom of the 2-D histogram (i.e., the lowest GROUP_MAX_EVENT_PCT for each GROUP_ENERGY) due to their low altitudes.

Machine learning provides an efficient framework for assessing how well different subsets of the parameters in Table 1 can predict the mean LMA altitudes associated with the diverse collection of GLM groups from the Colombia thunderstorm. We collect all of these GLM group metrics into a feature dataset and train random forest models from unique subsets of the parameters from Table 1 following the methods described in the next section. The top model from these tests will be used to analyze the Colombia thunderstorm in Section 3.

2.3 Scikit-Learn Random Forest Regression

Constructing machine learning models requires dividing the feature and label data into training and testing datasets. While we have 22,681 GLM groups matched to LMA sources, this sample of matches is not representative of generic GLM data for three reasons:

(1) The matching scheme prioritizes assigning LMA sources to the brightest groups in a series rather than the nearest group in time.

(2) The LMA sources are not distributed uniformly through the cloud depth, but rather are concentrated in the primary charge layers of the Colombia thunderstorm.

(3) The GLM groups were measured under a low instrument threshold that is not representative of thunderstorms elsewhere, particularly during the day.

To account for these biases, we take a judicious approach towards constructing the testing and training datasets. We limit the effect of the group matching preference in (1) by only including the brightest group in each unique series in the testing / training data. We reduce charge layer bias in (2) by adjusting the number of matches taken from each vertical level (LMA measured altitudes in 1-km bins) to ensure nearly-equal contributions from each CTH-normalized vertical layer (through, smaller numbers of sources near the top and bottom of the cloud are still allowed). Finally, we address the threshold concerns in (3) by recalculating the group parameters after imposing artificial thresholds between 1 and 10 fJ (as in Peterson et al., 2021b), and then adding the surviving groups at each threshold to the testing / training data. Thus, the random forest model is sensitive to how group characteristics change under varying instrument thresholds.

Once the feature and label data are compiled, we divide the matched groups into training (75%) and testing (25%) samples and begin the scikit-learn random forest regressor for various combinations of features. Note that in addition to the designated testing sample consisting of the brightest groups per series, we can also test the model with groups that had LMA matches but were not the brightest groups in their parent series, as this much larger dataset is not used for

training. We find that many of the 16 parameters that we devised in Table 1 were not useful for predicting altitude because they provided redundant information. For example, both the group energy Half Width of Half Max (GROUP_HWHM) and the percent of the group energy in the brightest event (GROUP_MAX_EVENT_PCT) describe the breadth of the spatial energy distribution of the group. While these parameters might provide some unique information in certain situations, the model assigns an importance score of 0 on a scale from 0 (not important) to 1 (the only important metric) to one of these parameters if the other is included as a feature. Moreover, these parameters have vastly different computational costs. While GROUP_MAX_EVENT_PCT is based on a simple sum of event energies, GROUP_HWHM requires modeling the radiance fall-off with distance from the brightest event in the group and then finding where this model falls below 50% of the maximum energy. As having both metrics does not improve the model, there is simply no benefit to using GROUP_HWHM. Other examples include group area / group event count, group area / convex hull area, and even group area / group energy.

This exercise revealed a set of five features that had considerable skill in predicting the LMA mean source altitude for the matched GLM groups: the maximum separation in the parent series (SERIES_GROUP_MAX_SEPARATION: importance: 0.39), which describes the horizontal extent of the lightning process that generated the group of interest; the percent of the group energy in the brightest event (GROUP_MAX_EVENT_PCT: importance: 0.23), which was shown in Figure 4; the distance between the group centroid and brightest event location (GROUP_MAX_LOC_DIS: importance: 0.16), which is sensitive to radiance anomalies in the group footprint; group footprint area (GROUP_AREA: importance 0.15); and the approximate GLM threshold for the parent flash (FLASH_THRESHOLD_APPROX: importance: 0.06). We

ran the random forest regressor with only these parameters included as features and then used the resulting machine learning model to predict the source altitudes for the GLM groups that were detected in the Colombia thunderstorm.

3 Results

This section will evaluate the GLM source altitudes retrieved by the random forest model. We will first evaluate model performance using the testing sample of matched GLM groups / LMA sources in Section 3.1. Then, Section 3.2 will compare GLM and LMA altitude trends within individual flashes and at the storm level over the duration of the Colombia thunderstorm.

3.1 GLM Source Altitude Model Performance with Testing Group Data

Histograms of LMA mean altitude, GLM predicted altitude, and the altitude difference between the LMA measurements and GLM predictions for the matched groups in the testing dataset are shown in Figure 5. Note that we do not include single-event groups in these analyses because they lack sufficient unique information for sources at different altitudes to be distinguished. The model mostly assigns these single-event detections to a single layer, which is not useful.

The rows in Figure 5 correspond to two-or-more event groups with various artificial thresholds applied. No threshold is applied in Figure 5a-c, a 2 fJ threshold is imposed in Figure 5d-f, a 4 fJ threshold is applied in Figure 5g-i, and a 6 fJ threshold is applied in Figure 5j-l. While the initial sample of LMA mean source altitudes in Figure 5a has a nearly equal number of sources between 40% and 100% of the ABI CTH, this near parity is not maintained at higher

thresholds (Figure 5d,g,j). The same sample group data from Figure 5a is used to generate these higher-threshold samples, but groups associated with LMA sources outside of the primary charge layer (~70% ABI CTH) preferentially fall below the higher imposed thresholds.

Similar biases can be found in the GLM predictions in Figure 5e,h, and i. Despite matched groups being chosen to ensure the LMA mean source altitudes were evenly-distributed between vertical layers, the illumination of the surrounding clouds leads to group radiance patterns that the model suggests come from the primary charge layer at 70% ABI CTH rather than elsewhere in the vertical profile. This could be an indication that the input data is not sufficiently robust to account for some group radiance patterns, as the filters described in Section 2 leave only on the order of 100 groups in each vertical level. If this is the case, then adding matched LMA-GLM data from additional thunderstorms might improve the model – particularly if the matched data is supplied from multiple LMAs across the GLM FOV and represent a diverse collection of thunderstorm charge structures. Another likely cause of this bias in the predictions is that our choice of estimating the optical source altitude from the mean LMA source altitude is not properly representing sources with a finite vertical extent (as we saw with the example flash in Figure 1). Rather than taking the mean or maximum LMA source altitude, a normalization scheme to account for GLM's detection advantage for high-altitude sources developed from Monte Carlo radiative transfer modeling could improve the agreement with observations.

Despite this apparent bias, the model errors in Figure 5c,f, and i remain low. With no artificial threshold imposed, the median absolute error is 9.7% of the ABI CTH, or 1.33 km. Generating similar plots from LMA-matched groups that were not the brightest in their series yields similarly-low errors. Histograms for the groups not included in the training or testing data

495 are shown in Figure S1. The median absolute errors for these predictions range from 6.62% (0.95
496 km) for >1 event groups to 4.18% (0.60 km) for >7 event groups.

497 In most cases, therefore, we can at least correctly predict which charge layer within the
498 Colombia thunderstorm the optical emissions originated. Interestingly, imposing a higher
499 threshold actually improves these error statistics. This could be an effect of the increasing
500 concentration of sources in the layer centered at 70% CTH, or it could signify that removing the
501 fainter events along the periphery of the GLM groups by imposing a higher threshold improves
502 the altitude estimate by limiting the cloud-edge illumination that results in CTH uncertainty.

503 To test if these reduced errors under higher threshold are physical, we construct new
504 altitude histograms based on event count under a 6 fJ threshold in Figure 6. As we saw in
505 Peterson et al., 2021a, the altitude profiles depend on group size with single-pixel groups
506 primarily originating from near the top of the cloud and large multi-pixel groups originating from
507 low altitudes. These trends are expected to be amplified under a high threshold. Indeed, while the
508 peak in the altitude distribution for all >1 event groups (Figure 6a-c) is at 70% ABI CTH,
509 increasing the event count to >3 events in Figure 6d-f, >5 events in Figure 6g-i, and >7 events in
510 Figure 6j-l causes the peak to descend in altitude. Meanwhile, the median absolute errors in
511 Figure 6c,g,i, and l decrease from 4.56% (0.64 km) to 3.83% (0.54 km), 3.45% (0.51 km), and
512 1.89% (0.3 km) as the groups increase in size and the peak becomes displaced vertically from the
513 primary charge layer in the thunderstorm. Thus, higher thresholds probably do improve the
514 altitude estimates by reducing the influence of ABI CHT gradients on the predictions. However,
515 these improvements come at the cost of limiting the number of predictions that can be made – as
516 the abundant dim groups most quickly fall below threshold.
517

3.2 GLM Source Altitude Model Predictions of Flash / Thunderstorm Trends

The GLM source altitude prediction model is next applied to all GLM groups from the Colombia thunderstorm – regardless of whether they match any LMA sources or occur as part of a larger series. Applying the model generally will allow us to examine how well it captures major LMA altitude trends at the flash and thunderstorm level.

We begin by using the LMA-matched data to reproduce the altitude timeseries from Figure 3b-e with GLM predictions in Figure 7. Figure 7a and c are identical to Figure 3, while Figure 7b and d replace the ABI CTH timeseries with GLM-retrieved altitude timeseries. Note that these GLM altitudes have been converted back to units of kilometers using the local ABI CTH at each group centroid for direct comparison with Figure 7a and c. As before, the first two panels consider all matched groups (including single-event groups) while the last two panels consider only groups with >5 constituent events.

Despite the expected uncertainty from ABI CTH gradients and the use of LMA mean source altitudes as a measure of optical source altitude, the GLM predictions are able to reproduce the primary features in the LMA altitude distribution over the thunderstorm duration – including periods of intensification leading to increases in source altitude at 07:00 UTC, 09:00 UTC, and 10:00 UTC and maturation causing source altitude to decrease after 11:00 UTC. Still, the GLM altitude timeseries for all groups (Figure 7b) and >5 event groups (Figure 7d) over-estimate the peak source altitudes during periods of intensifications. This appears to be due to the ABI CTH normalization. The group radiance profile suggests that the source is above the local ABI CTH value, but the ABI CTH is high enough that the altitude retrieved from the GLM data is predicted to be between 17 km and 20 km. If we re-run the model without the normalization (not shown for brevity), these 17-20 km predicted altitudes disappear, but the model then over-

estimates the altitudes of low-altitude sources that are embedded in low clouds. A 90th percentile altitude product or something similar applied to the ABI CTH normalized data might balance preserving these low sources while still permitting changes in source altitudes to be tracked.

GLM-retrieved altitudes could also be used to generate new GLM gridded products (Bruning et al., 2019). Figure 8 examines the spatial distributions of these LMA measured and GLM predicted altitudes by computing a Mean Source Altitude (MSA) grid over a 1.5 hour interval between 07:30 UTC and 12:00 UTC. LMA measurements of MSA are shown in the left column (Figure 8a,e,i,m) and the LMA vertical profile is shown in the second column (Figure 8b,f,j,n). These plots are then repeated for the GLM predicted altitudes in the right two columns. The MSA grid at 07:30 UTC contains a single concentrated feature with high source altitudes surrounded by a small number of matched groups around its edge. This MSA feature describes an isolated thunderstorm that was active during this period before the larger and more mature storm system moved into the LMA data domain. As we saw in Figure 7b, the GLM predictions overestimate the tallest LMA source altitudes at this point in time, though the peak in the altitude profile (Figure 8d) is nearly identical to the LMA (Figure 8b). The isolated matched groups around the storm edges are also at low altitudes (3-6 km) in both the LMA and GLM plots. Normalizing by ABI CTH allows the GLM predictions to pick up on these lower edge sources.

The MSA grids are more complex by 09:00 UTC (Figure 8e-h) with multiple lightning centers containing flashes at different altitudes. By this point of the storm, the larger and more mature thunderstorm feature had moved into the LMA domain and was generating the low-altitude propagating flashes. These horizontal flashes occur between 5 km and 9 km in the LMA data (Figure 8e) and the GLM predicted altitudes largely agree (Figure 8g). The key difference between the LMA measurements and GLM predictions here are in the quantity of low-altitude

564 sources (Figure 8f and h), not the average source altitudes.

565 The previous trends for 07:30 UTC and 09:00 UTC persist to the 10:30 UTC time step
566 (Figure 8i-l). The GLM predictions are occasionally higher than the LMA measurements, but the
567 peak of the distribution is identical and both MSA grids show the same trends of higher sources
568 in the eastern convective feature while low-altitude sources dominate the propagating flashes on
569 the western flank of the storm. Finally, by 12:00 UTC (Figure 8m-p), the low-altitude
570 propagating flashes overtake the higher-altitude convective flashes, causing both the LMA and
571 GLM altitude profiles to peak at just 7 km altitude.

572 To evaluate the performance of the GLM altitude prediction model at the flash level, we
573 repeat the analyses in Figures 1 and 2 while adding a new overlay to represent the GLM
574 predicted altitude for every multi-event group during the flash of interest. GLM altitude
575 predictions for the low-altitude flash in Figure 1 are shown in Figure 9 while the predicted
576 altitudes from the high-altitude flash in Figure 2 are shown in Figure 10. These new GLM
577 altitude overlays are added to the longitude / altitude cross sections (Figure 9c, 10c), latitude /
578 altitude cross sections (Figure 9e, 10e) and altitude timeseries (Figure 9g, 10g) in the same style
579 as GLM groups in the plan view (Figure 9d, 10d) and area / energy distribution (Figure 9b, 10b).
580 The GLM groups are depicted with larger box symbols whose color corresponds to the time-
581 ordered group index. GLM predicted altitude histograms are also added to Figure 9h and 10h.

582 As with the previous thunderstorm trends, the GLM predicted altitudes from the low-
583 altitude flash in Figure 9 are largely consistent with the vertical range of LMA source altitudes
584 (Figure 9h). While differences arise between GLM and the LMA for individual groups, much of
585 this can be attributed to the vertical extent of LMA sources involved in each match. GLM
586 likewise correctly predicts that the LMA sources in the high-altitude flash in Figure 10 occur

around 15 km altitude. However, GLM adds more detail to this flash case, as the LMA only recorded one source before 550 ms into the GLM flash (which could be noise due to its low altitude and horizontal separation from the other sources). All of the GLM predicted source altitudes are above 10 km in this case, which is consistent with the LMA flash in question.

Figure 11 performs the same analysis as Figures 9 and 10 for the ascending flash discussed in Peterson et al., 2021a. This flash produced LMA sources primarily in the 5 km layer early on and generated two ENGLN -CGs before developing upward into the 10 km layer between 300 ms and 400 ms into the GLM flash. We see the same behavior in the GLM predictions in Figure 11g. There were 5 groups in the early portion of the flash (before 300 ms), and the model predicted that 4 were located in the 5 km layer. The later development into the upper layer was accompanied by sustained optical illumination, and the GLM-predicted source altitudes during this period likewise ascend into the upper layer. As discussed in Peterson et al., 2021a, the upward development of the flash causes the group area / energy distribution to have a "forked" appearance due to the low-altitude source producing a different area / energy relationship than high-altitude sources. This can be seen in Figure 11b here. These differences in how clouds are illuminated by sources at different altitudes are key to being able to predict source altitude with GLM.

The final flash that we examine in Figure 12 is the case of a long horizontal lightning flash that descended in altitude as it developed from the rear of the convective line into the stratiform region. This flash spawned a single ENGLN +CG and was unique from a GLM perspective for generating large, elongated groups that traced significant fractions of the existing lightning channel. Despite the limited quantities of stratiform flashes in the testing / training datasets, the GLM predictions are able to map the descent of the LMA flash from 14 km altitude

at its origin in the northwest down to 5 km as it traversed the electrified stratiform region. The longitude / altitude (Figure 12c), latitude / altitude (Figure 12 e) and timeseries (Figure 12g) all show reasonable matches between the LMA measurements and GLM predictions until the end of the flash (beyond 1500 ms). After this point, GLM predicts vertical development to high altitudes (10-15 KM). While LMA sources are not present at this point to confirm or refute these GLM altitudes, we do see this behavior with the LMA sources earlier in the flash around the time of the +CG.

The storm-level analyses in Figure 7 and 8, and the flash-level analyses in Figures 9 to 12 demonstrate that the GLM altitude prediction model is able to resolve the temporal and spatial variations in LMA altitude that respond to changes in the kinematics of the Colombia thunderstorm and are consistent with the physical structure of the flashes mapped by the LMA. The ability of the model to predict storm-scale and flash-scale trends in underlying LMA data that are not supplied as training data to the random forest regressor confirms that its skill does not come from overfitting the data, but instead that altitude information can be extracted from GLM measurements of thundercloud illumination.

4 Conclusion

In this third part of our thundercloud illumination study, we use machine learning methods to determine whether source altitude information can be retrieved from the spatial energy distributions of GLM groups. To do this, we find the LMA sources that match the GLM groups recorded from a thunderstorm in Colombia, construct group-level metrics to describe attributes of their radiance patterns that are relevant to thundercloud illumination, and then use the Python scikit-learn random forest regressor to construct a model for predicting mean LMA source altitude (normalized by ABI Cloud Top Height) from these group-level metrics.

We find that the machine learning model can retrieve source altitudes in the testing dataset (and data not used for testing or training) well enough to determine which charge layer the optical emissions originated from (median absolute error: 1.33 km). The model also has skill in capturing changes to the thunderstorm LMA source distributions in response to convective invigoration or maturation and resolving the vertical extent of individual lightning flashes – including cases where the flash ascends or descends in the cloud.

Additional work is needed to expand these methods into a general source altitude retrieval algorithm that can work with arbitrary thunderstorms. Future work will expand our collection of matched GLM-LMA data to enable the construction of such a retrieval. The eventual goal is to be able to derive flash-level, storm-level, and climatological lightning altitude trends over the full 25-year global lightning dataset provided by OTD, LIS, GLM, and other similar instruments. Currently, these analyses are only possible with a reasonable accuracy over limited regional domains (for example, within ~300 km of an LMA). Adding this capability to all of the lightning imagers will provide an unparalleled view of the three-dimensional extent of global lightning and its response to a changing climate.

Acknowledgments

This work was supported by the US Department of Energy through the Los Alamos National Laboratory (LANL) Laboratory Directed Research and Development (LDRD) program under project number 20200529ECR. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001). The work by co-author Douglas Mach was supported by ASA 80MFSC17M0022 “Cooperative Agreement with Universities Space Research

Association" and NASA Research Opportunities in Space and Earth Science grant NNX17AJ10G "U.S. and European Geostationary Lightning Sensor Cross-Validation Study." We would like to thank the operators of the Colombia LMA at the Technical University of Catalonia and Dr. Jesús López for sharing their processed LMA data for the presented case. The data used in this study is available at Peterson (2021b and c).

References

Albrecht, R. I., Goodman, S. J., Buechler, D. E., Blakeslee, R. J., & Christian, H. J. (2016). Where are the lightning hotspots on Earth?. *Bulletin of the American Meteorological Society*, *97*(11), 2051-2068.

Aranguren, D., Lopez, J., Montanya, J., & Torres, H. (2018, September). Natural observatories for lightning research in Colombia. In *2018 International Conference on Electromagnetics in Advanced Applications (ICEAA)* (pp. 279-283). IEEE.

Bitzer, P. M. (2017). Global distribution and properties of continuing current in lightning. *Journal of Geophysical Research: Atmospheres*, *122*(2), 1033-1041.

Blakeslee, R.J., Lang, T.J., Koshak, W.J., Buechler, D., Gatlin, P., Mach, D.M., Stano, G.T., Virts, K.S., Walker, T.D., Cecil, D.J., Ellett, W., Goodman, S.J., Harrison, S., Hawkins, D.L., Heumesser, M., Lin, H., Maskey, M., Schultz, C.J., Stewart, M., Bateman, M., Chanrion, O. and Christian, H. (2020), Three Years of the Lightning Imaging Sensor Onboard the International Space Station: Expanded Global Coverage and Enhanced Applications. *J. Geophys. Res. Atmos*., **125**: e2020JD032918. https://doi.org/10.1029/2020JD032918

Blyth, A. M., Christian Jr, H. J., Driscoll, K., Gadian, A. M., & Latham, J. (2001). Determination of ice precipitation rates and thunderstorm anvil ice contents from satellite observations of lightning. *Atmospheric Research*, *59*, 217-229.

Boggs, L. D., Liu, N., Peterson, M., Lazarus, S., Splitt, M., Lucena, F., ... & Rassoul, H. K. (2019). First observations of gigantic jets from geostationary orbit. *Geophysical Research Letters*, *46*(7), 3999-4006.

Bruning, E. C., Rust, W. D., MacGorman, D. R., Biggerstaff, M. I., & Schuur, T. J. (2010). Formation of charge structures in a supercell. *Monthly Weather Review*, *138*(10), 3740-3761.

Bruning, E. C., Weiss, S. A., & Calhoun, K. M. (2014). Continuous variability in thunderstorm primary electrification and an evaluation of inverted-polarity terminology. *Atmospheric Research*, *135*, 274-284.

Bruning, E. C., Tillier, C. E., Edgington, S. F., Rudlosky, S. D., Zajic, J., Gravelle, C., ... & Meyer, T. C. (2019). Meteorological imagery for the geostationary lightning mapper. *Journal of Geophysical Research: Atmospheres*, *124*(24), 14285-14309.

Carey, L. D., Murphy, M. J., McCormick, T. L., and Demetriades, N. W. S. (2005), Lightning location relative to storm structure in a leading-line, trailing-stratiform mesoscale convective system, *J. Geophys. Res.*, 110, D03105, doi:10.1029/2003JD004371.

695 Cecil, D. J., Goodman, S. J., Boccippio, D. J., Zipser, E. J., & Nesbitt, S. W. (2005). Three years
696 of TRMM precipitation features. Part I: Radar, radiometric, and lightning characteristics.
697 *Monthly Weather Review*, *133*(3), 543-566.
698 Cecil, D. J., Buechler, D. E., & Blakeslee, R. J. (2014). Gridded lightning climatology from
699 TRMM-LIS and OTD: Dataset description. *Atmospheric Research*, *135*, 404-414.
700 Christian, H. J., R. J. Blakeslee, S. J. Goodman, and D. M. Mach (Eds.) (200). Algorithm
701 Theoretical Basis Document (ATBD) for the Lightning Imaging Sensor (LIS),
702 NASA/Marshall Space Flight Center, Alabama. (Available as
703 http://eospso.gsfc.nasa.gov/atbd/listables.html, posted 1 Feb. 2000).
704 Christian, H. J., Blakeslee, R. J., Boccippio, D. J., Boeck, W. L., Buechler, D. E., Driscoll, K. T.,
705 ... & Stewart, M. F. (2003). Global frequency and distribution of lightning as observed
706 from space by the Optical Transient Detector. *Journal of Geophysical Research:*
707 *Atmospheres*, *108*(D1), ACL-4.
708 DeMaria, M., DeMaria, R. T., Knaff, J. A., & Molenar, D. (2012). Tropical Cyclone Lightning
709 and Rapid Intensity Change, *Monthly Weather Review*, *140*(6), 1828-1842. Retrieved
710 Mar 2, 2021, from https://journals.ametsoc.org/view/journals/mwre/140/6/mwr-d-11-
711 00236.1.xml
712 Ely, B. L., Orville, R. E., Carey, L. D., and Hodapp, C. L. (2008), Evolution of the total lightning
713 structure in a leading-line, trailing-stratiform mesoscale convective system over Houston,
714 Texas, *J. Geophys. Res.*, 113, D08114, doi:10.1029/2007JD008445.
715 Fierro, A. O., Stevenson, S. N., & Rabin, R. M. (2018). Evolution of GLM-Observed Total
716 Lightning in Hurricane Maria (2017) during the Period of Maximum Intensity, *Monthly*
717 *Weather Review*, *146*(6), 1641-1666. Retrieved Mar 2, 2021, from
718 https://journals.ametsoc.org/view/journals/mwre/146/6/mwr-d-18-0066.1.xml
719 Gatlin, P. N., & Goodman, S. J. (2010). A total lightning trending algorithm to identify severe
720 thunderstorms. *Journal of atmospheric and oceanic technology*, *27*(1), 3-22.
721 Goodman, S. J., D. Mach, W. J. Koshak, and R. J. Blakeslee. (2010). *GLM Lightning Cluster-*
722 *Filter Algorithm (LCFA) Algorithm Theoretical Basis Document (ATBD)*. Retrieved from
723 https://www.goes-r.gov/products/ATBDs/baseline/Lightning_v2.0_no_color.pdf, posted
724 24 Sept. 2010
725 Goodman, S. J., Blakeslee, R. J., Koshak, W. J., Mach, D., Bailey, J., Buechler, D., ... & Stano,
726 G. (2013). The GOES-R geostationary lightning mapper (GLM). *Atmospheric research*,
727 *125*, 34-49.
728 Hutchins, M. L., Holzworth, R. H., Brundell, J. B., and Rodger, C. J. (2012), Relative detection
729 efficiency of the World Wide Lightning Location Network, *Radio Sci.*, 47, RS6005,
730 doi:10.1029/2012RS005049.
731 Heidinger (2012): Algorithm Theoretical Basis Document (ATBD) for ABI Cloud Height,
732 NOAA/NESDIS/STAR. (Available as
733 https://www.star.nesdis.noaa.gov/goesr/documents/ATBDs/Baseline/ATBD_GOES-
734 R_Cloud%20Height_v3.0_July%202012.pdf).
735 Jacobson, A.R., R. Holzworth, J. Harlin, R. Dowden, and E. Lay, 2006: Performance Assessment
736 of the World Wide Lightning Location Network (WWLLN), Using the Los Alamos
737 Sferic Array (LASA) as Ground Truth. *J. Atmos. Oceanic Technol.,* **23**, 1082–1092,
738 https://doi.org/10.1175/JTECH1902.1

739 Jayaratne, E. R., Saunders, C. P. R., & Hallett, J. (1983). Laboratory studies of the charging of
740 soft-hail during ice crystal interactions. *Quarterly Journal of the Royal Meteorological*
741 *Society*, *109*(461), 609-630.
742 Jiang, H., & Ramirez, E. M. (2013). Necessary Conditions for Tropical Cyclone Rapid
743 Intensification as Derived from 11 Years of TRMM Data, *Journal of Climate*, *26*(17),
744 6459-6470. Retrieved Mar 2, 2021, from
745 https://journals.ametsoc.org/view/journals/clim/26/17/jcli-d-12-00432.1.xml
746 Lay, E. H., Holzworth, R. H., Rodger, C. J., Thomas, J. N., Pinto, O., and Dowden, R. L. (2004),
747 WWLL global lightning detection system: Regional validation study in Brazil, *Geophys.*
748 *Res. Lett.*, 31, L03102, doi:10.1029/2003GL018882.
749 Liu, C., Cecil, D., & Zipser, E. J. (2011). Relationships between lightning flash rates and passive
750 microwave brightness temperatures at 85 and 37 GHz over the tropics and subtropics.
751 *Journal of Geophysical Research: Atmospheres*, *116*(D23).
752 Liu, C., Cecil, D. J., Zipser, E. J., Kronfeld, K., & Robertson, R. (2012). Relationships between
753 lightning flash rates and radar reflectivity vertical structures in thunderstorms over the
754 tropics and subtropics. *Journal of Geophysical Research: Atmospheres*, *117*(D6).
755 López, J. A., Montanyà, J., van der Velde, O., Romero, D., Aranguren, D., Torres, H., ... &
756 Martinez, J. (2016, September). First data of the Colombia lightning mapping array—
757 COLMA. In *2016 33rd International Conference on Lightning Protection (ICLP)* (pp. 1-
758 5). IEEE.
759 Lyons, W. A., Bruning, E. C., Warner, T. A., MacGorman, D. R., Edgington, S., Tillier, C., &
760 Mlynarczyk, J. (2020). Megaflashes: Just how long can a lightning discharge get?.
761 *Bulletin of the American Meteorological Society*, *101*(1), E73-E86.
762 Mansell, E. R., MacGorman, D. R., Ziegler, C. L., & Straka, J. M. (2005). Charge structure and
763 lightning sensitivity in a simulated multicell thunderstorm. *Journal of Geophysical*
764 *Research: Atmospheres*, *110*(D12).
765 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay,
766 E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning*
767 *research*, *12*, 2825-2830.
768 Peterson, M. (2020a). Modeling the transmission of optical lightning signals through complex 3-
769 D cloud scenes. *Journal of Geophysical Research: Atmospheres*, 125, e2020JD033231.
770 https://doi.org/10.1029/2020JD033231
771 Peterson, M. (2020b). Holes in Optical Lightning Flashes: Identifying Poorly-Transmissive
772 Clouds in Lightning Imager Data. *Earth and Space Science*, 7,
773 e2020EA001294https://doi.org/10.1029/2020EA001294
774 Peterson, M. (2021a). GLM-CIERRA http://dx.doi.org/10.5067/GLM/CIERRA/DATA101
775 Peterson, M. (2021b). Coincident Optical and RF Lightning Detections from a Colombia
776 Thunderstorm. https://doi.org/10.7910/DVN/5FR6JB, Harvard Dataverse, V1
777 Peterson, M. (2021c). Machine Learning Models for Predicting Lightning Altitude.
778 https://doi.org/10.7910/DVN/VM1YEI, Harvard Dataverse, V1
779 Peterson, M., & Liu, C. (2011). Global statistics of lightning in anvil and stratiform regions over
780 the tropics and subtropics observed by the Tropical Rainfall Measuring Mission. *Journal*
781 *of Geophysical Research: Atmospheres*, *116*(D23).
782 Peterson, M., Rudlosky, S., & Deierling, W. (2017). The evolution and structure of extreme
783 optical lightning flashes. *Journal of Geophysical Research: Atmospheres*, 122, 13,370–
784 13,386. https://doi.org/10.1002/2017JD026855

Peterson, M., Light, T., & Mach, D. (2021a). The Illumination of Thunderclouds by Lightning: Part 1: The Extent and Altitude of Optical Lightning Sources. *Journal of Geophysical Research: Atmospheres*.

Peterson, M., Light, T., & Mach, D. (2021b). The Illumination of Thunderclouds by Lightning: Part 2: The Effect of GLM Instrument Threshold on Detection and Clustering. *Journal of Geophysical Research: Atmospheres*.

Rodger, C. J., Werner, S., Brundell, J. B., Lay, E. H., Thomson, N. R., Holzworth, R. H., & Dowden, R. L. (2006, December). Detection efficiency of the VLF World-Wide Lightning Location Network (WWLLN): initial case study. In *Annales Geophysicae* (Vol. 24, No. 12, pp. 3197-3214). Copernicus GmbH.

798 **Table 1.** GLM metrics that were considered as potential features for the machine learning model.
799 Entries with an asterisk symbol were used in the final model.
800

Parameter Name	Units	Description
GROUP_ENERGY	fJ	Group total energy
GROUP_MAX_EVENT_PCT*	%	Percent of group energy in brightest event
GROUP_AREA*	km^2	Group footprint area
GROUP_CONVEX_AREA	km^2	Area of convex hull around all events in the group
GROUP_MAX_LOC_DIS*	km	Distance between group centroid and brightest event location
GROUP_EVENT_MAX_SEPARATION	km	Maximum great circle distance between events
GROUP_HWHM	km	Half Width of Half Maximum of constituent event energy
GROUP_ELONGATION	ratio	Group elongation factor (major axis length / minor axis length)
GROUP_EVENT_COUNT	#	Number of events in the group
GROUP_N50	#	Min. number of events to capture 50% of the group energy
GROUP_N75	#	Min. number of events to capture 75% of the group energy
GROUP_N90	#	Min. number of events to capture 90% of the group energy
GROUP_LOCAL_MAX_COUNT	#	Number of local maxima in the group footprint
GROUP_HOLE_COUNT	#	Number of holes (pixels with no events) in the group footprint
SERIES_GROUP_MAX_SEPARATION*	km	Maximum separation of groups in the parent series feature
FLASH_THRESHOLD_APPROX*	fJ	Approximation of the GLM threshold for the parent flash

801

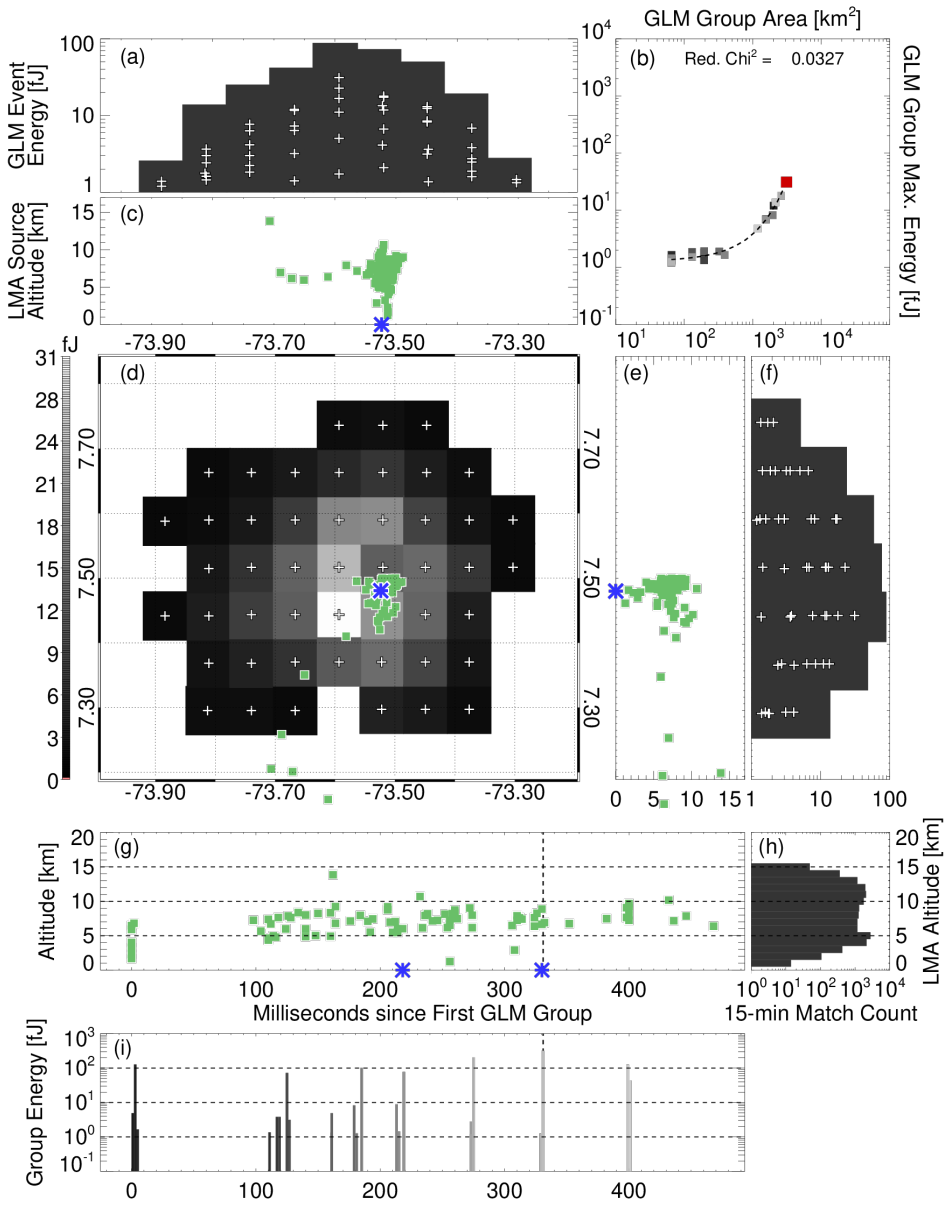

Figure 1. The largest group in an example low-altitude GLM flash. The plan view (d) shows an image of the group (dark: low energy, light: high energy) with events indicated with a + symbol, LMA sources overlaid with small green boxes, and ENGLN -CG (blue) or +CG (red) strokes indicated with asterisk symbols. Panels (c) and (e) show LMA cross sections by altitude and either longitude (c) or latitude (e). Panels (a) and (f) show GLM longitude energy cross sections by longitude (a) or latitude (f). Plus (+) symbols in (a) and (f) indicate individual events while bars show column totals. Timeseries for LMA source altitude (g) and GLM group energy (i) are shown below the map. An LMA source altitude distribution is provided in (h), while the group energy / area distribution for the GLM flash is shown in (b). The groups in (i) and (b) are color coded by their time-ordered index number. A polynomial fit is also applied to the data in (b) and shown as a dashed line with its reduced chi^2 value overlaid.

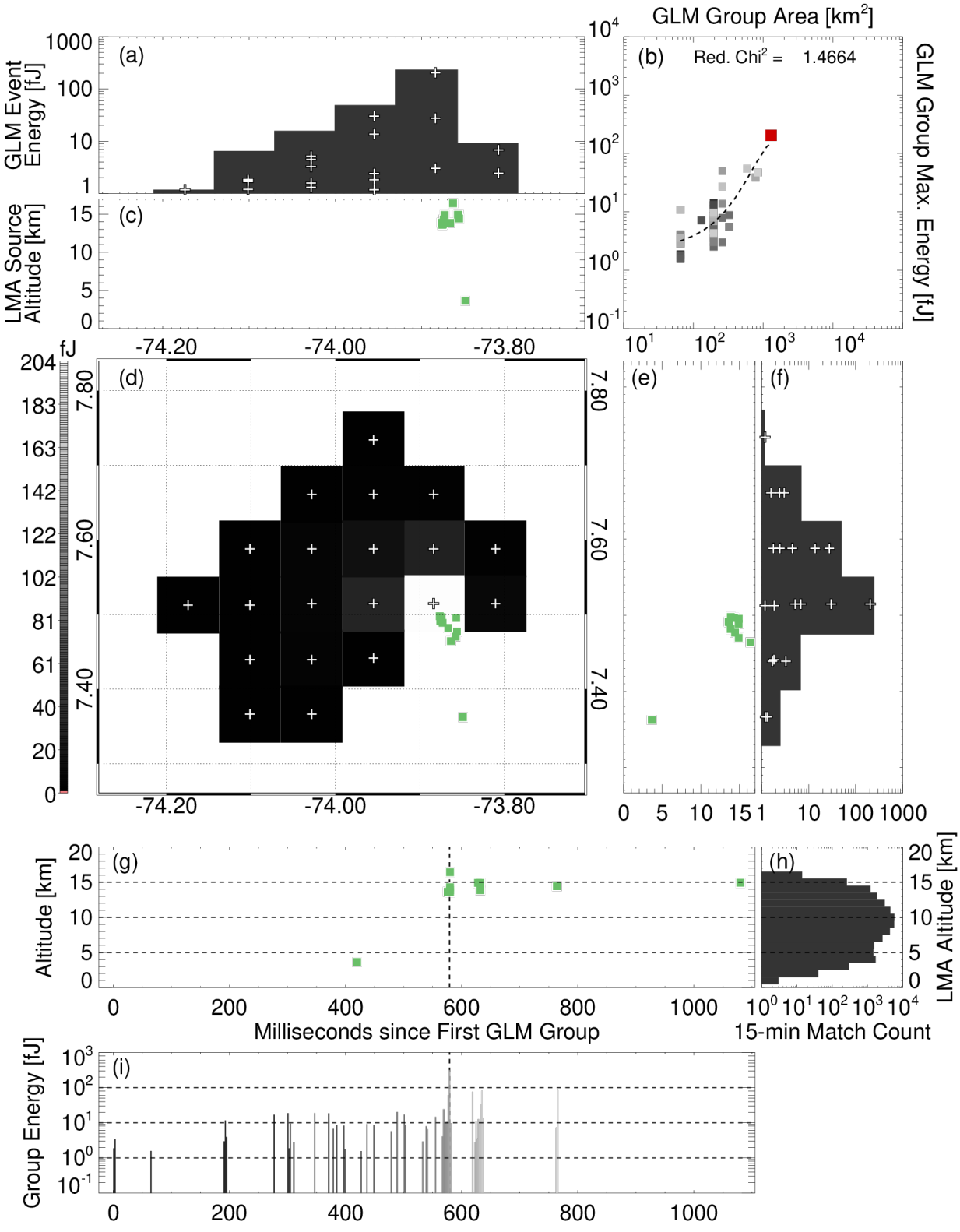

Figure 2. As in Figure 1, but for the largest group in an example high-altitude GLM flash.

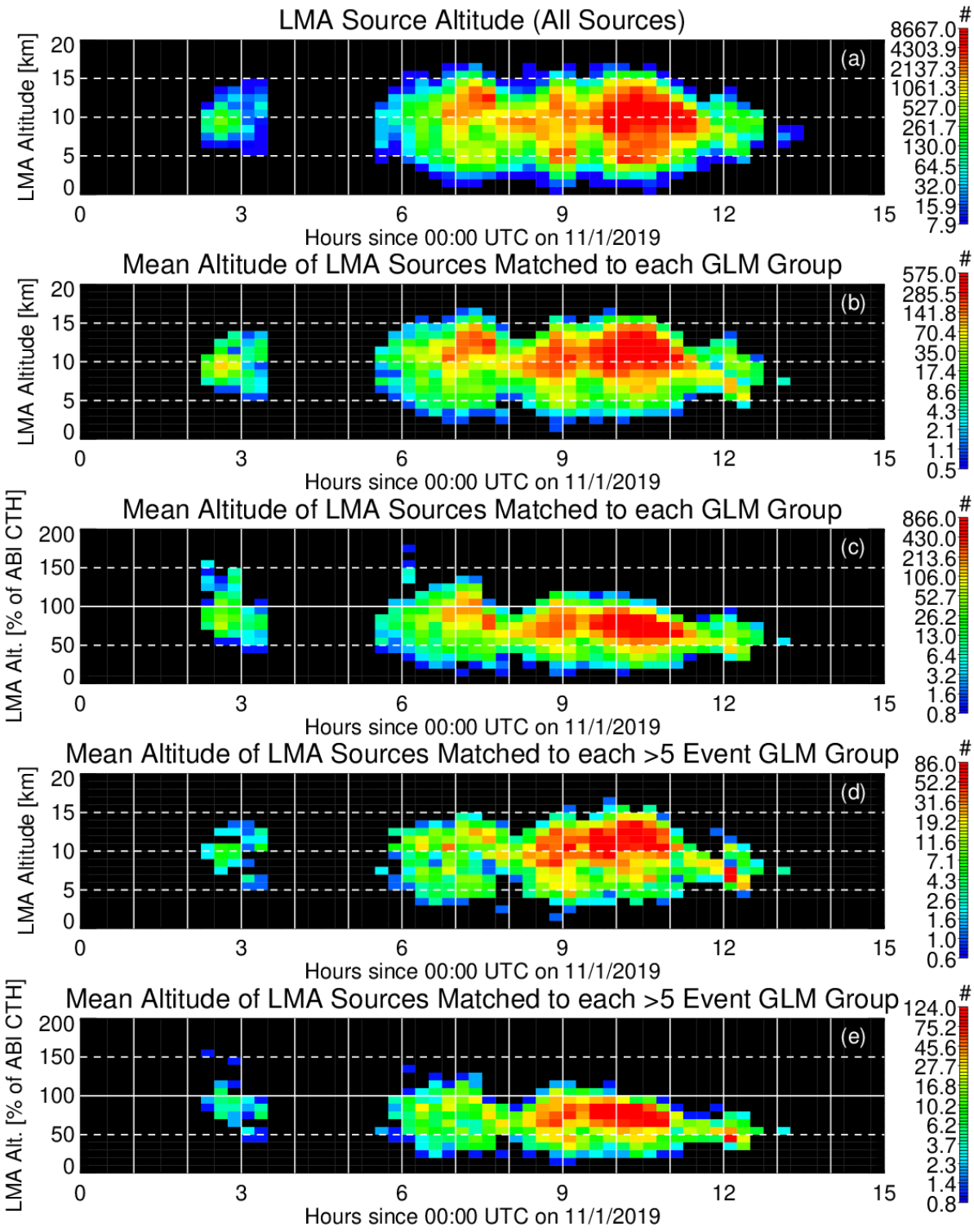

Figure 3. Timeseries of LMA source altitude (a) and the mean altitudes of LMA sources matched to GLM groups (b-e). Measured LMA altitudes are shown for all matched GLM groups in (b) and for groups with >5 events in (d), while LMA altitudes normalized to the local ABI Cloud Top Height (CTH) are shown in (c) for all groups and (e) for groups with >5 events.

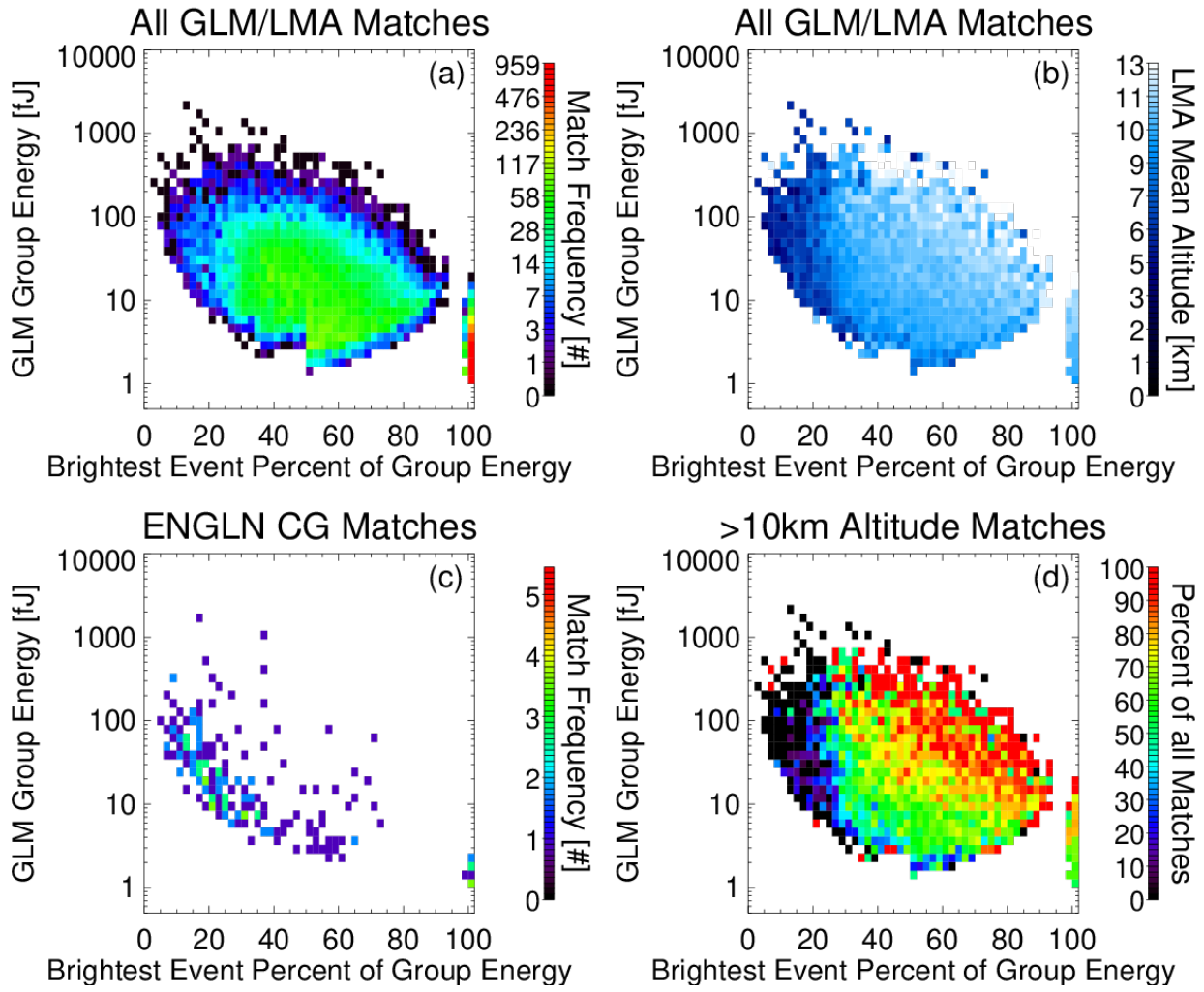

Figure 4. LMA / ENGLN attributes of matched GLM groups with varying group energy and brightest event percent of group energy values. (a) Two-dimensional histogram of LMA matches. (b) Average LMA source altitude contour plot. (c) Two-dimensional histogram of ENGLN CG matches. (d) Fraction of high-altitude (>10 km) matches in each bin.

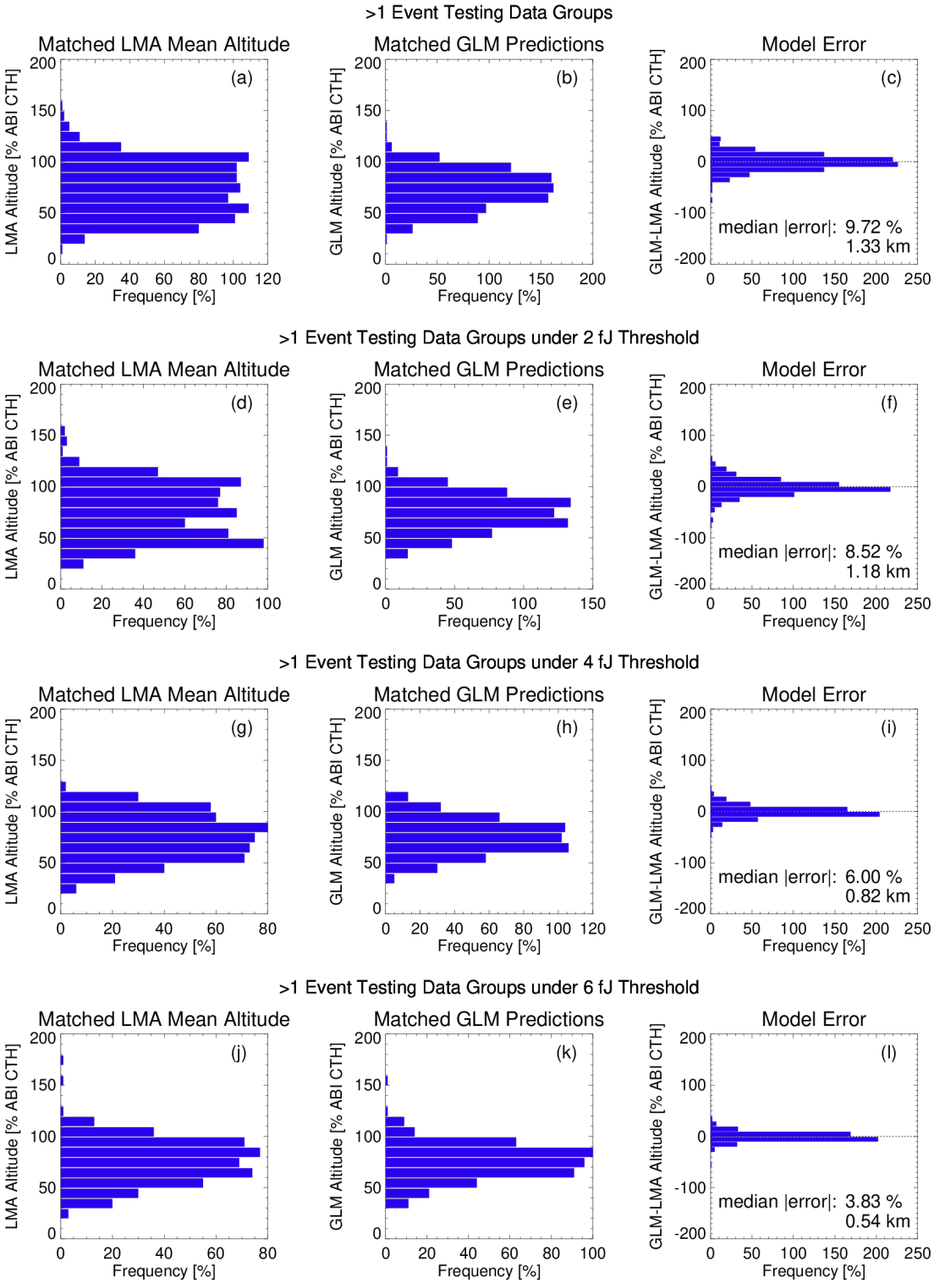

Figure 5. Comparisons between LMA measured altitudes (a,d,g,j) and GLM predicted altitudes b,e,h,k) in the model testing dataset. Model errors are shown in (c,f,i,l). Each row corresponds to a different imposed threshold on the GLM groups: 0 fJ (a-c), 2 fJ (d-f), 4 fJ (g-i), or 6 fJ (j-l).

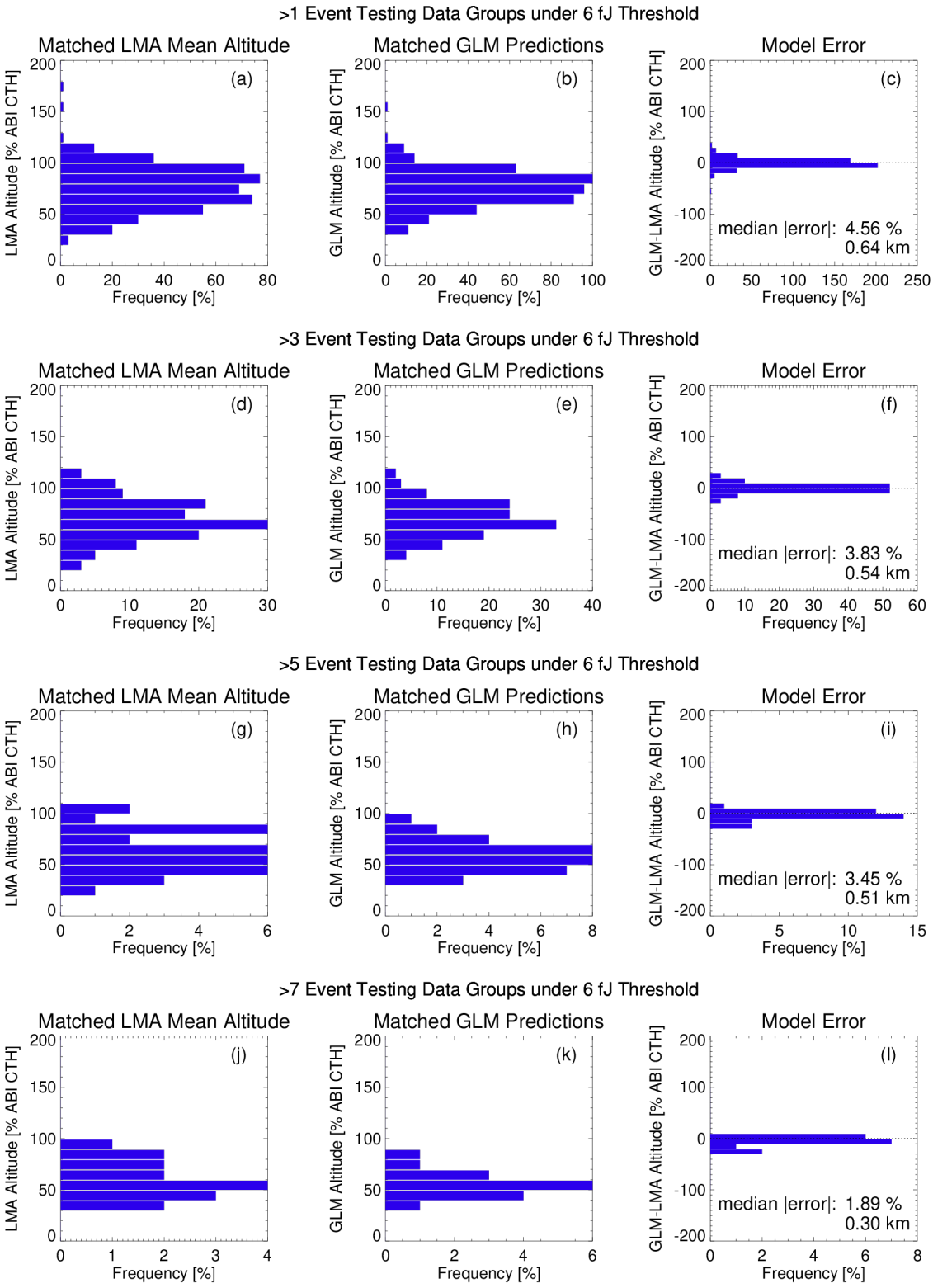

Figure 6. Comparisons between LMA measured altitudes (a,d,g,j) and GLM predicted altitudes b,e,h,k) for a 6 fJ threshold in the model testing dataset. Each row corresponds to a minimum number of events per group: >1 event (a-c), >3 events (d-f), >5 events (g-i), or >7 events (j-l).

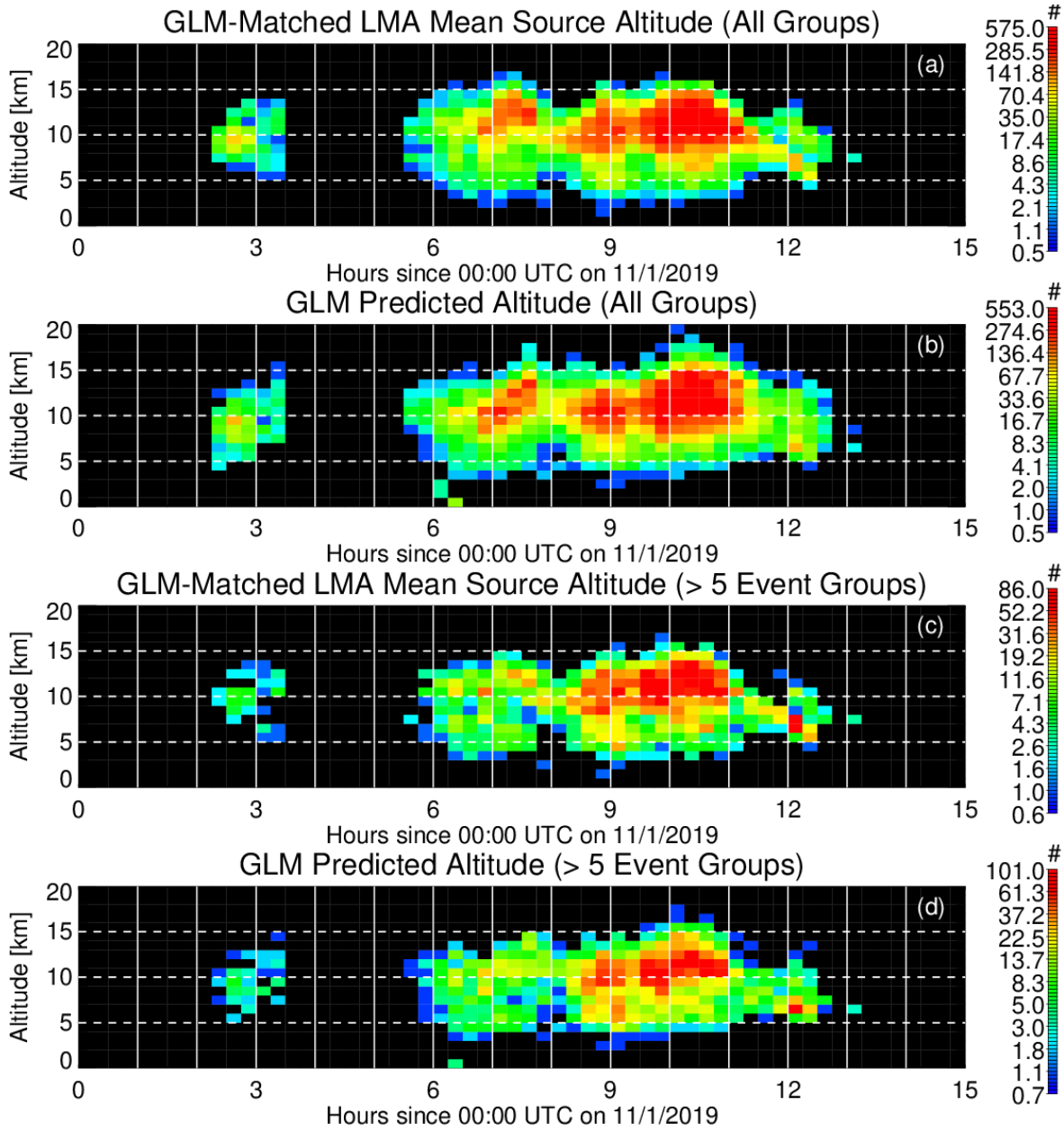

Figure 7. Timeseries of the mean altitudes of LMA sources matched to GLM groups (a,c) and GLM predicted altitudes from matched groups (b,d). As in Figure 3, (a) and (c) include all matched groups while (b) and (d) only consider groups with >5 events.

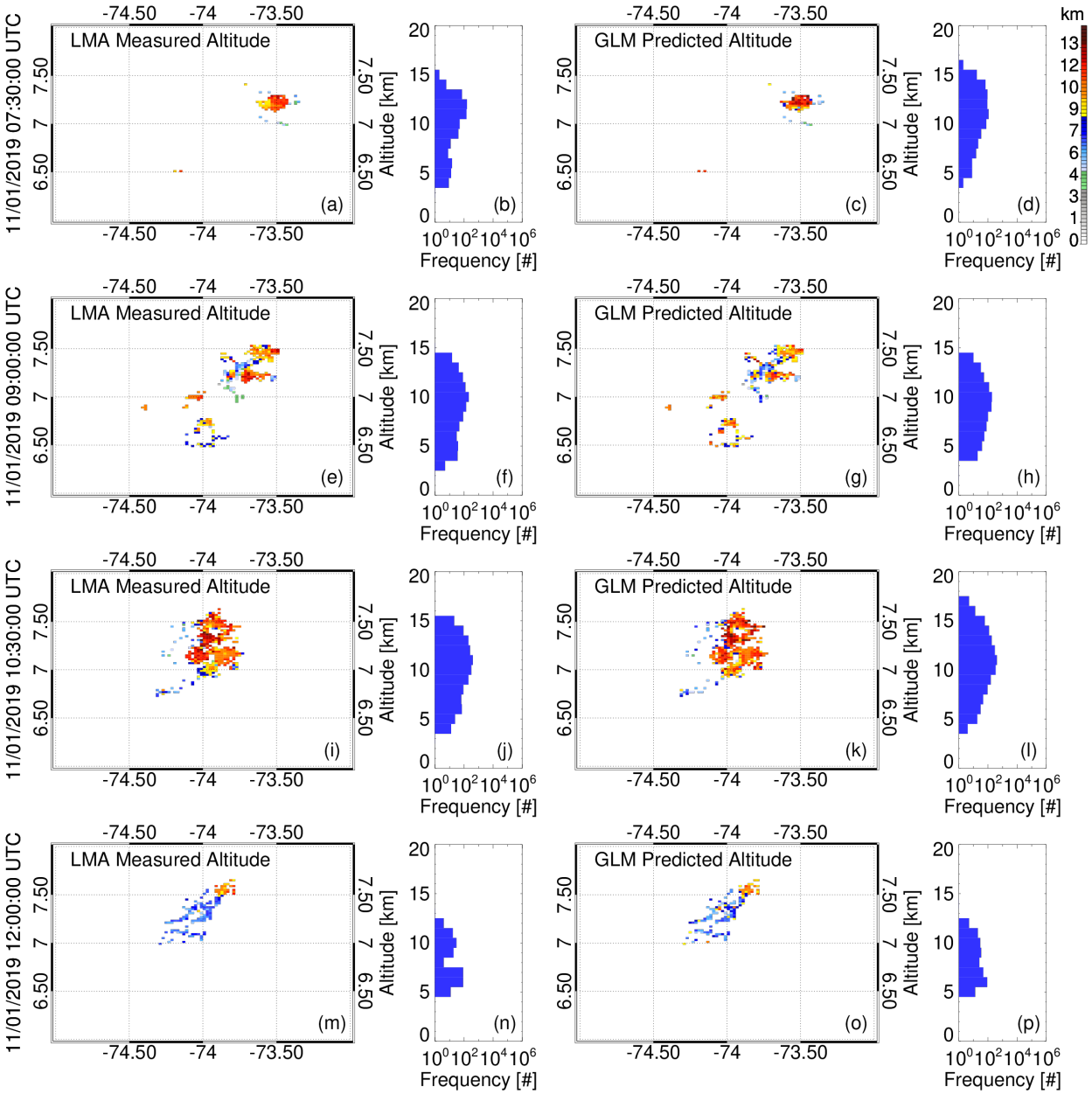

Figure 8. Mean Source Altitude (MSA) grids (left) and source altitude profiles (right) constructed from LMA measured altitudes (a-b,e-f,i-j,m-n) and GLM predicted altitudes (c-d,g-h,k-l,o-p). Each row corresponds to a unique time during the Colombia thunderstorm in 1.5 hour increments starting at 07:30 UTC (a-d).

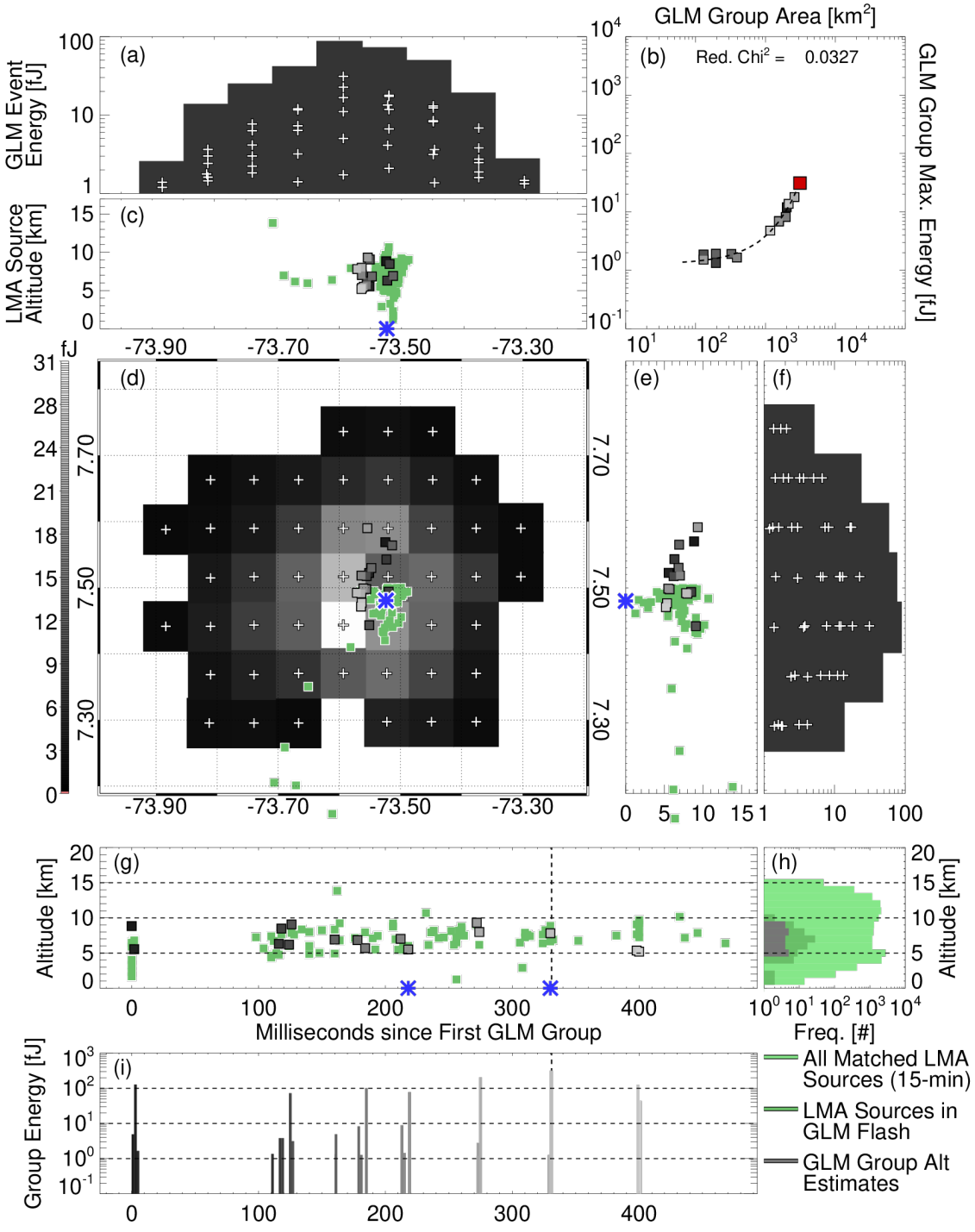

Figure 9. As in Figure 1, but with GLM predicted source altitudes (greyscale boxes) added to (c), (e), and (g). Box colors are identical to (b), (d), or (i), but single-event groups are not shown. LMA source (green) and GLM group (grey) altitude profiles for the specific flash in question are added to (h).

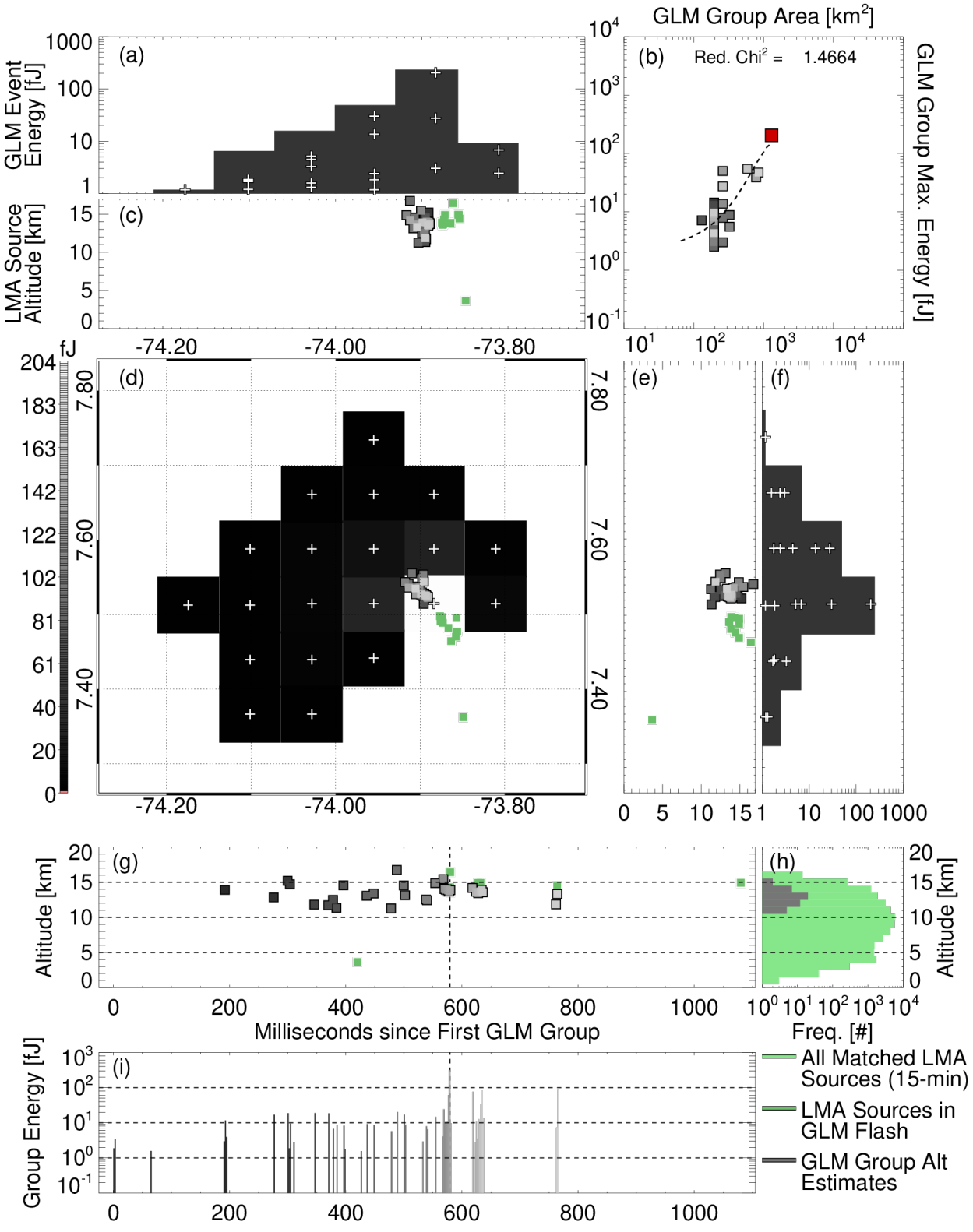

Figure 10. As in Figure 9, but for the high-altitude GLM flash in Figure 2.

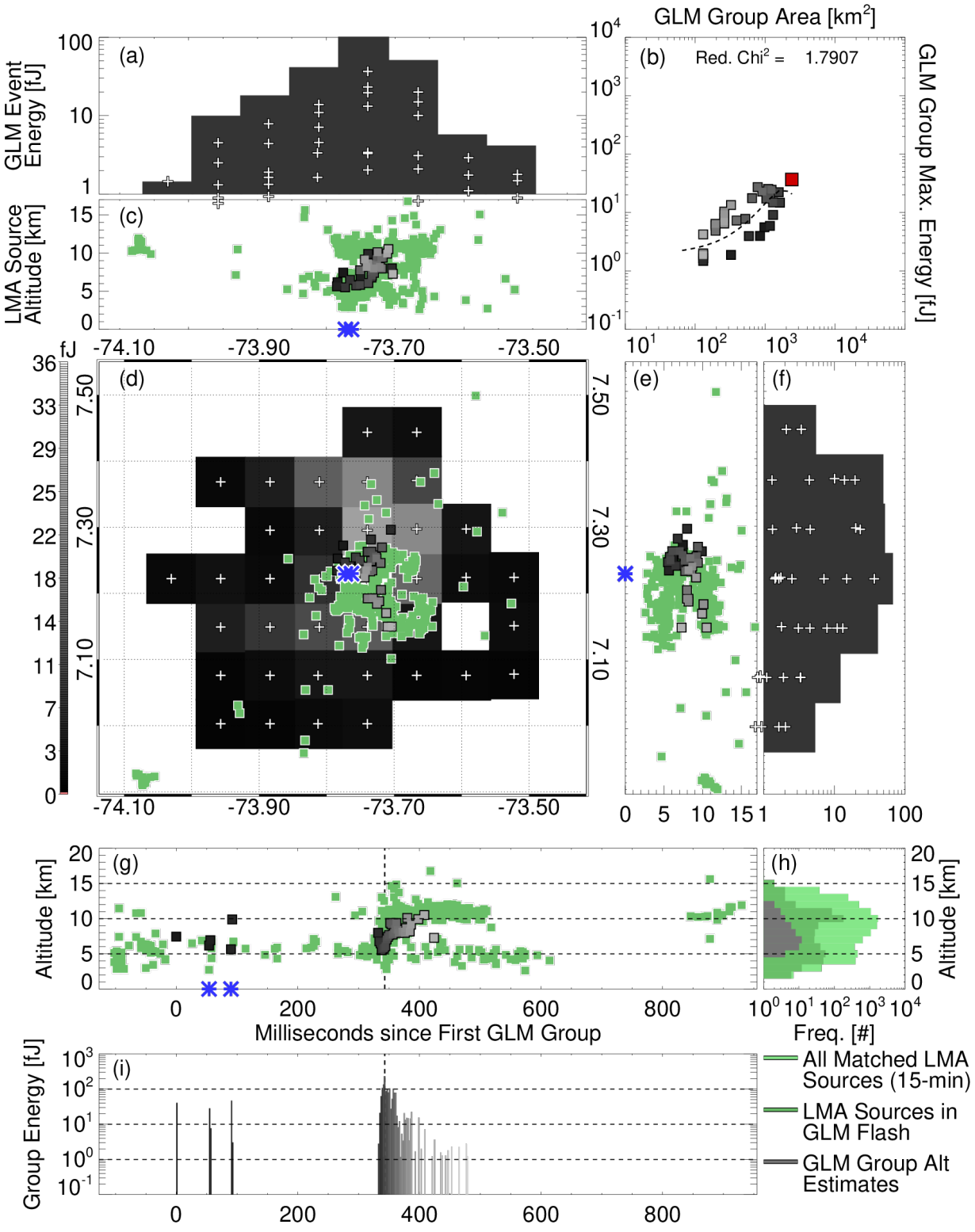

Figure 11. As in Figure 9, but showing the GLM predicted altitudes following the ascent of LMA sources in the upward-developing GLM flash that was discussed in Peterson et al., 2021a.

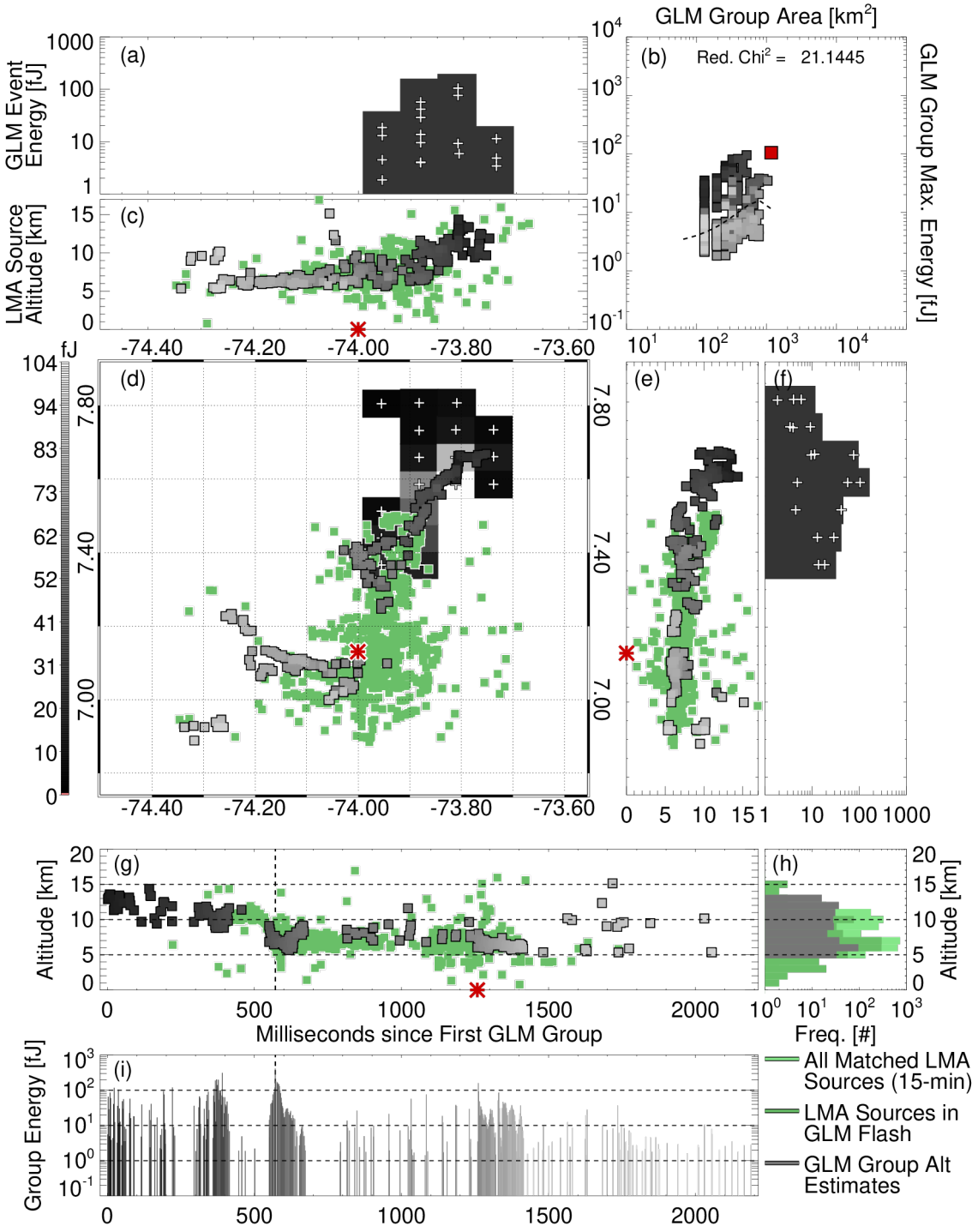

Figure 12. As in Figure 9, but showing the GLM predicted altitudes resolving the descent of LMA sources in a long horizontal flash.