NOAA ROSES Semi-Annual Report

Reporting Period: September 2021 – February 2022 (3rd report)

PI: Daile Zhang
Co-PI(s):

Project Title: Downscaling of GLM Lightning Observations Using ISS-LIS Data

Executive Summary (1 paragraph max)

This report summarizes progress on four main project components: 1) intercomparison of the Geostationary Lightning Mappers (GLMs), and the International Space Station Lightning Imaging Sensor (ISS-LIS) using the coincidence datasets; 2) optimizing GLM parallax corrections and building the GLM blended dataset; 3) identifying individual GLM flash type using a decision tree machine learning algorithm; and 4) website development and transfer for data storage, data visualization, data query and other resources. In the following, we provide a summary of the work performed to date, including analysis results, data development, website design and development, as well as future work.

Progress toward FY21 Milestones and Relevant Findings (with any Figs)

1. GLM and ISS-LIS Intercomparison

An intercomparison study of coincidence of GLM and ISS-LIS data is conducted. Our main findings are: 1) GLM tends to detect and match with LIS when the LIS group radiance is higher, the cloud-top optical energy is higher, the group area is larger, and/or the number of children events in the groups are larger. 2) More than half of the time differences between the coincidence GLM and LIS are within ±2 ms. About half of the time, LIS started observing lightning before the first GLM events. The mean location differences are mostly less than 10 km (about the size of a GLM pixel) and centered at around 4-5 km (about the size of a LIS pixel). 3) GLM matches well with LIS in its center of the field of view, whereas has a relatively poorer performance on the edges. This location offset vector map is consistent with the GLM geolocation errors for original ellipsoid (Virts and Koshak, 2020), and can be further applied to GLM location offset correction and intercomparison, as well as our following work of machine learning optical pattern recognition and downscaling.

2. GLM Parallax Corrections and Blended Dataset

Two ways are used to correct the GLM parallax: GLM L2 LCFA data with 14 km at equator and 6 km at polar regions over the earth ellipsoid, and the optimal estimated detection heights (Virts and Koshak, 2020). We upscaled the original optimal heights 3º by 3º to 2 km by 2 km by the following steps: 1) For each 3º by 3º grid, the lat-lon geolocations of its four vertices to the fixed grid coordinates (y,x) (x: Fixed Grid E/W scan angle in radians; y: Fixed Grid N/S scan angle in radians) are converted to obtain the corresponding quadrilateral coordinates in fixed grid coordinates. 2) All GLM full-disk 2 km by 2 km grid centroids (in fixed grid coordinates) that located in the same 3º by 3º grids quadrilateral coordinates have been assigned to the same optimal height. The monthly mean optimal estimated detection heights for both GLM-16 and -17, and their differences with the LCFA lightning ellipsoid heights are compared. Overall, 70% differences are less than 3 km. Large differences mostly lie over the edges of the field of the view with a range of 17-24 km. There are small location differences (less than 2 km) over the CONUS and most of South America regions. GLM-17 shows larger location differences over most of the domain than those for GLM-16.

We created the first version of a 1-min blended lightning dataset (1-month period) that includes the GLM L2 data (with optimal heights corrected), and other ground-based data such as the Global Lightning Detection 360 (GLD360) and Earth Network Total Lightning Network (ENTLN). Overall, the geolocations of the blended data are more consistent with the ground-based measurements. A common GLM detection behavior of miss detecting signals in between other lightning areas (also called a GLM "hole") can be corrected by the blended data. Another common GLM detection behavior of miss detecting one side of the outer boundary of the lightning area can also be corrected by the blended data.

Ongoing work includes comparing the gridded blended data with regional Lightning Mapping Arrays for performance evaluation. Compared to the single GLM Flash Extent Density (FED), the blended FED are better geolocated with the D.C.Lightning Mapping Array (DCLMA) FED, which has a higher flash detection efficiency. The blended FED displays more lightning signals and covers a larger lightning area than that of single GLM FED. More testing and evaluation are still under investigation. The updated blended dataset will be used for future applications, such as lightning data assimilation.

3. Classification of GLM Flashes into CG and IC by Decision Tree Model

Since the GLM does not distinguish individual flash type of cloud-to-ground (CG) and intra-cloud (IC) lightning, we have explored the flash type classification of the GLM total lightning using a Decision Tree model. A Decision Tree is a supervised machine learning classification algorithm. Given a labeled dataset, i.e., GLM lightning dataset labeled as CG and IC, a classification tree learns a sequence of if-else questions about individual features or criteria to infer the labels (CG and IC). Each question involves a feature and a split-point. Fig. 1 shows the workflow chart of the binary tree graph that is used to assign for each data to a target value based on branches and nodes. The target values are presented in the tree leaves. Starting from the root on top, the sample goes through the internal nodes to reach the leaves at the bottom. At each node, a decision is made based on the selected sample's feature. The colors of the node represent the training status with deep colors meaning better-trained (or deep layers) and light colors not-well-trained (shallow layers). Three main steps are used as follows:

1. Matching: GLM and GLD360 datasets are matched at the group level, using the spatial and temporal thresholds of 50 km and 15 ms. Any groups in a GLM flash match with a GLD360-classified CG stroke was categorized as a CG flash. The rest of the flashes were categorized as an IC flash.
2. Machine learning: A dataset of full field-of-view GLM level-2 flashes in January 2021 are used for the decision tree model, among which 80% are used for training and 20% are used for testing. The scikit-learn model is used for the decision tree classifier with the maximum depth of six. There are seven different flash features selected for training as seen in Table 1. The detailed workflow diagram including the branches, nodes, leaves, values, scores, depth of layers, and classification is shown in Fig. 2.

3. Evaluation: To estimate the performance of the decision tree model, the accuracy scores are calculated for each feature and all seven features (Table 1).

The decision tree algorithm was chosen for this research because it requires low computation and it is easy to visualize. The limitations of the decision tree model are relatively low accuracy and possible overfitting issue. Ongoing work of fine tuning the model and adding ISS-LIS corrected GLM data are still under development. The results will be implemented into a deep learning model for cloud-top optical products pattern recognitions.

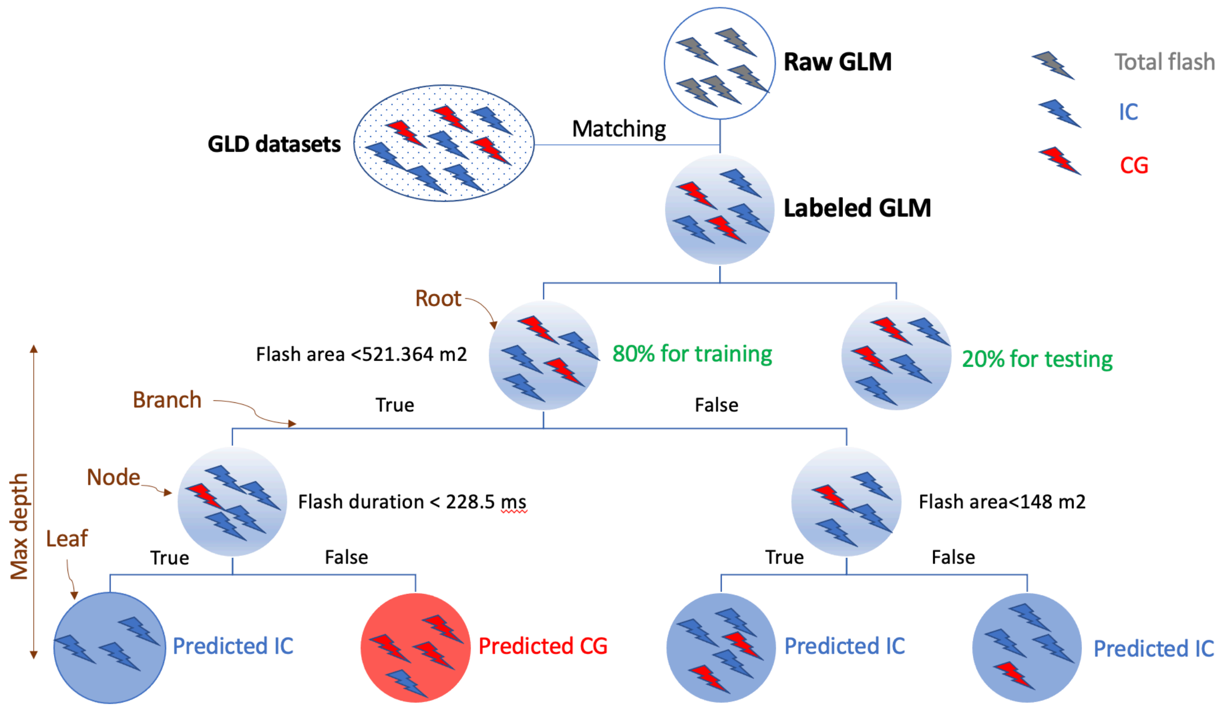

Fig. 1 A simplified workflow illustration of the decision tree model.

TABLE 1. ACCURACY SCORES OF GLM FALSHES CHARACTERISTICS INPUT AS FEATURES INTO THE DECISION TREE (DT) MODEL

		Definition	Accuracy Score
FEATURE 1	Flash time offset of first event	Time of occurrence of the first constituent event in flash	0.88548951048951
FEATURE 2	Flash time offset of last event	Time of occurrence of the last constituent event in flash	0.88548951048951
FEATURE 3	Flash duration	The time duration of a flash, i.e., the difference of the two variables above	0.881993006993007
FEATURE 4	Flash latitude	Flash centroid latitude (mean	0.880244755244755

		constituent event latitude weighted by their energies) latitude coordinate	
FEATURE 5	Flash longitude	Flash centroid latitude (mean constituent event latitude weighted by their energies) longitude coordinate	0.880244755244755
FEATURE 6	Flash energy	Flash radiant energy	0.88548951048951
FEATURE 7	Flash area	Flash area coverage (pixels containing at least one constituent event only)	0.886363636363636
ALL		Considering all features above	0.876748251748251

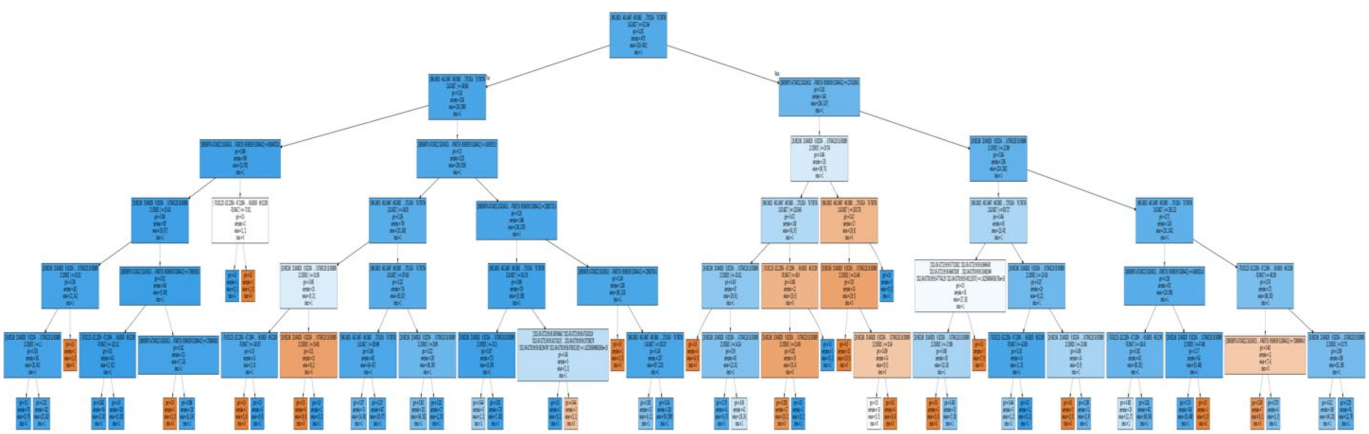

Fig. 2 Decision tree visualization with graphviz method

4. GLM Website Development

The Atmospheric Electricity Research and Training (AERT) website has finished its first-round (6-month) testing period, and it is currently under final security check by our IT engineer, and ready to transfer to the new server, which will be permanent. The testing website is temporarily available at: http://essic5.umd.edu/~wwwdeploy/index.html. It features GLM L2 data status, data inquiry, trainings, visualizations, news, and safety. The overall performance is stable with multiple users accessing and one admin user updating. Once the transfer is done, we will work closely with our IT engineer to ensure the new server is stable and all the functionalities are running correctly.

The sponsored high school student, who has been working on the website development and maintenance is recently admitted to the University of Maryland. The student will continue working on the website updates once it is online. In the meantime, the student is doing research on GLM blended dataset evaluation.

Plans for Next Reporting Period

- Develop the light source illumination deconvolution model for the coincidence GLM and ISS-LIS, based on our former model for intercomparing GLM and TRMM-LIS.
- Develop an updated version of the blended lightning dataset and the blended FED products with ISS-LIS location corrected GLM data, and evaluate the performance with the regional LMA data.
- Fine-tune the Decision Tree machine learning model for GLM flash type classification by adding ISS-LIS location corrected GLM data.
- Develop a deep learning model for optical pattern recognitions using the GLM and ISS-LIS data.