A geostationary lightning pseudo-observation generator utilizing low frequency ground-based lightning observations

Felix Erdmann, Olivier Caumont, Eric Defer

A geostationary lightning pseudo-observation generator utilizing low frequency ground-based lightning observations

Felix Erdmann*

CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France & Laboratoire d'Aérologie, Université de Toulouse, CNRS, UPS, Toulouse, France

Olivier Caumont

CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Eric Defer

Laboratoire d'Aérologie, Université de Toulouse, CNRS, UPS, Toulouse, France

* *Corresponding author*: Felix Erdmann, Erdmann.professional@gmx.de

ABSTRACT

Coincident Geostationary Lightning Mapper (GLM) and National Lightning Detection Network (NLDN) observations are used to build a generator of realistic lightning optical signal in the perspective to simulate Lightning Imager (LI) signal from European NLDN-like observations. Characteristics of GLM and NLDN flashes are used to train different machine learning (ML) models, that predict simulated pseudo-GLM flash extent, flash duration, and event number per flash (targets) from several NLDN flash characteristics. Comparing statistics of observed GLM targets and simulated pseudo-GLM targets, the most suitable ML-based target generators are identified. The simulated targets are then further processed to obtain pseudo-GLM events and flashes. In the perspective of lightning data assimilation, Flash Extent Density (FED) is derived from both observed and simulated GLM data. The best generators simulate accumulated hourly FED sums with a bias of 2 % to the observation, while cumulated absolute differences remain of about 22 %. A visual comparison reveals that hourly simulated FED features local maxima at the similar geolocations as the FED derived from GLM observations. However, the simulated FED often exceeds the observed FED in regions of convective cores and high flash rates. The accumulated hourly area with FED>0 flashes per 5 km × 5 km pixel simulated by some pseudo-GLM generators differs by only 7 % to 8 % from the observed values. The recommended generator uses a linear Support Vector Regressor (linSVR) to create pseudo-GLM FED. It provides the best balance between target simulation, hourly FED sum, and hourly electrified area.

1. Introduction

Lightning is defined as electrical discharges within the atmosphere, more particularly within and between clouds (intra- and intercloud, IC) or between clouds and the ground (CG). Transient lightning phenomena also occur between the cloud and the upper atmosphere, e.g., Sprites and Jets. While cloud electrification and lightning initiation are still subject of studies, it is widely accepted that cloud ice and graupel are necessary to separate charges within clouds (e.g., Luque et al. 2020; Emersic and Saunders 2020; Lyu et al. 2019; Kolmasova et al. 2019; Takahashi et al. 2017; MacGorman and Rust 1998; Brooks et al. 1997). In particular, convection creates favorable conditions for lightning, and the updraft strength can be well correlated to the total lightning rate (e.g., Deierling and Petersen 2008). Ávila et al. (2010) found a high correlation between the occurrence of deep convection and lightning over land at a global scale. Hence, lightning is an effective tracer of deep convection.

The new generation of geostationary (GEO) satellites carry optical lightning sensors, among other instruments. The Geostationary Lightning Mapper (GLM) of the Geostationary Operational Environmental Satellite (GOES) R-series, the Lightning Mapping Imager (LMI) on board the Chinese Fengyun-4 satellites (Yang et al. 2017), and the upcoming Meteosat Third Generation Lightning Imager (MTG-LI, Dobber and Grandell 2014) will provide GEO lightning observations at a global scale. This satellite-based, large-scale, continuous observation of lightning offers new information for climate monitoring and studies. In addition, the assimilation of GEO lightning data in Numerical Weather Prediction (NWP) can help to improve the initial state of the model. Most of recent lightning data assimilation

studies use gridded Flash Extent Density (FED), for example Allen et al. (2016); Fierro et al. (2019).

To assimilate new observation types in NWP models it is desired to develop an assimilation scheme prior to the instrument launch and data availability. The simulation of appropriate realistic pseudo-observations precedes the development of any assimilation scheme, especially when the sensor is not yet in operation. Such synthetic observations can be derived from existing GEO sensors over other regions, i.e., GLM, and ground-based Lightning Locating Systems (LLSs). In addition, Low Earth Orbit (LEO) missions such as the Lightning Imaging Sensor (LIS) on the Tropical Rainfall Measurement Mission (TRMM) satellites (e.g., Christian et al. 1999; Cecil et al. 2005) and on board the International Space Station (ISS) (Blakeslee and Koshak 2016; Blakeslee et al. 2020) provide space-based lightning observations. One can also use ground-based networks, e.g., the National Lightning Detection Network (NLDN) (e.g., Cummins and Murphy 2009), Meteorage (e.g., Schulz et al. 2016; Erdmann et al. 2020), and Lightning Mapping Arrays (LMAs) (e.g., Rison et al. 1999; Thomas et al. 2004; Coquillat et al. 2019). While the satellite sensors detect visible light of lightning at 777.4 nm, the ground-based networks are operated at different frequencies that match electromagnetic radiation emitted by different lightning processes. NLDN and Meteorage use low frequency (LF) sensors that are most sensitive to discharge processes such as return strokes for CG flashes. Most LF networks can distinguish CG and IC signals. The CG flash detection (with return strokes) is usually reliable, whereas the IC flash detection efficiency (DE) increases within the network and for shorter baselines given one LF sensor type (personal communication, Stéphane Pedeboy, 2020/21). Global LF networks have lower DE and accuracy than national and regional LF networks (e.g., Nag et al. 2015). LMA stations sense very high frequency (VHF) signals of lightning leader propagation and

allow for 3-dimensional (3D) channel mapping (e.g., Rison et al. 1999). Their drawback is the limited range. A LMA network provides coverage within a radius of typically a few hundred kilometers (e.g., Thomas et al. 2004; Koshak et al. 2004; Chmielewski and Bruning 2016; Coquillat et al. 2019).

Biron et al. (2008) resampled TRMM-LIS data on a MTG-LI-like grid to assess the potential performance of the MTG-LI with emphasis on the influence of varying minimal detectable radiant energy. However, this method relying on LEO lightning data is not suitable for producing continuous pseudo-observations in the same area for operational applications because of the poor revisiting time. Stano (2013) demonstrated a simple method to create pseudo-GLM gridded products using LMA data. The pseudo-GLM data served to train forecasters on the use of GLM data products. GLM's Algorithm Working Group (AWG) investigated a transformation function that transforms LMA sources to optical lightning observations. The technique combines TRMM-LIS flash statistics and observed LMA flashes (Bateman 2013). The same method was applied by Schultz et al. (2016) to study automated storm tracking and lightning jump algorithms using GLM pseudo-observations. Höller and Betz (2010) present a simple statistical model for transforming stroke-type data of the LF network LINET (Betz et al. 2009) to pseudo-MTG-LI optical events. The statistical relations were studied comparing LINET strokes to concurrent TRMM-LIS groups. Then, they created a pixel matrix of the future MTG-LI and used TRMM-LIS statistics of radiance and event number per group to obtain pseudo-MTG-LI events. Their work aimed to propose a statistics-based method to create optical pseudo-observations of lightning from a given set of LF strokes. The available satellite lightning data solely emanated from the LEO TRMM-LIS mission, and in addition the number of cases was fairly limited (705 coincident flashes).

Recent studies assessing the GLM performance have shown that the DE varies within the field of view. GLM detects almost 90 % of the flashes in the south-eastern USA (e.g., Marchand et al. 2019; Murphy and Said 2020). The flash DE is statistically lower in other regions like Colorado. Rutledge et al. (2020) showed that the GLM performance depends on the charge structure and the hydrometeor distribution. In particular, electrically "anomalous" storms led to degrading GLM flash DE. The GLM flash DE also depends on the size and duration of flashes. Zhang and Cummins (2020) found that small, short duration flashes are more likely not observed by GLM than larger flashes.

This paper introduces in-depth techniques and results of creating GEO lightning pseudo-observations. The GEO lightning pseudo-observation generator is developed using NLDN records in the US and can be applied to all NLDN-like ground-based LLSs, e.g., Meteorage in France. One key part of the generator uses machine learning (ML) to relate NLDN-like observations to the extent and duration of the generated optical flashes. The generator simulates the GEO lightning pseudo-observations on the flash level including events and thus flash extent. FED grids can be derived from the generated pseudo-observations to serve as assimilation input data. This work prepares in particular the assimilation of pseudo-MTG-LI data in the Meteo-France operational mesoscale numerical weather prediction system AROME (Applications de la Recherche à l'Opérationnel à Méso-Echelle) in France. As MTG-LI will produce GLM-like data, and the French Meteorage network observes lightning similarly as NLDN in the US (Erdmann 2020, Chapter II.2.4), the developed GEO lightning pseudo-observation generator can be used to simulate realistic pseudo-MTG-LI data.

The main objective of this study is the generation of a realistic GEO lightning FED field. It does not aim at reproducing correctly individual flashes, but the FED product. Therefore, the most important characteristics are the overall flash number and the flash extent. There

is no direct dependency of FED on the flash duration and event number per flash, neither on flash energetics. The developed generator should provide synthetic MTG-LI FED over France for data assimilation studies (not in the scope of the present paper). The application in our study is not intended for an operational use even though the developed algorithm could be used for operational application or for training forecasters and users.

Section 2 introduces both NLDN and GLM instruments. It also describes the dataset with coincident GLM and NLDN flashes. Section 3 explains in-depth the strategy to mimic GLM data from NLDN observations. This includes a 2-part GEO lightning pseudo-observation generator and different ML models to relate GLM and NLDN flash characteristics. Section 4 presents pseudo-GLM observations, their comparison to real GLM observations, and the evaluation of the 2-part generator. FED from real and pseudo-GLM observations is compared for the different ML-based generators. Finally, recommendations for suitable GEO lightning pseudo-observation generators are given.

2. Instruments and Data

GLM and NLDN make use of different lightning detection and locating techniques. This section introduces important specifications of both instruments and the studied dataset. It briefly describes the developed methods to match and compare GLM and NLDN observations, and to infer flash characteristics needed for training ML models.

a. Geostationary Lightning Mapper (GLM)

The GLM is an optical sensor on board the GOES R-Series (currently GOES-16 at 75°W and GOES-17 at 137°W). This study uses the GOES-16 GLM data only. The GLM detects total lightning including IC and CG during day and night. Although it cannot directly

distinguish IC from CG signals, Koshak and Solakiewicz (2015) show that some ICs and CGs can be statistically differentiated. Especially due to the difficulty of the detection of daytime lightning against bright, sunlit clouds, thresholds and filters are applied to separate the lightning optical signal from background and other light sources. Lightning is detected in a narrow (1 nm) band centered at the 777.4 nm oxygen line in the near infrared. The wide field-of-view (FOV) image is focused on a high speed Charge Coupled Device (CCD) focal plane with a nearly hemispheric FOV coverage (1372×1300 pixels). The variable pitch pixel CCD allows for resulting pixels of about 8 km at nadir and only 14 km at the edge of the FOV (Goodman et al. 2013). Images are produced continuously and in time frames of 2 ms.

NASA's GLM lightning data algorithm produces Level 2 data with lightning information as events, groups and flashes. The x,y-coordinates of the focal plane are transformed to latitude and longitude coordinates of an estimated cloud top ellipsoid (with a height of 14 km at the equator and 6 km at the poles). Bruning et al. (2019) describes the effects of using this ellipsoid on GLM parallax with respect to any known ground-relative reference. GLM events are single illuminated pixels that pass the optical filters and are thus identified as lightning signals. Their location is defined as the center of the illuminated pixel. Adjacent events observed in the same 2 ms time frame are merged to form a group. Next, groups are combined into flashes. NASA's clustering algorithm uses a Weighted Euclidean Distance (WED) with limits of 16.5 km in latitude and longitude direction and 330 ms in time. Two groups with a WED of less than one are assigned to the same flash. The WED criterion is tested for pairs of events with one event in each group (Mach 2020).

The reader is referred to Goodman et al. (2013), the GLM Product Performance Guide for Data Users (Koshak et al. 2010), and Goodman et al. (2012) for further information on GLM details. Mach (2020) analyzed the GLM algorithms recently.

b. The National Lightning Detection Network (NLDN)

The NLDN (Cummins and Murphy 2009) consists of more than 100 Vaisala LS7002 ground sensors in the contiguous US (CONUS). It detects LF electromagnetic signals generated by fast lightning discharges such as return strokes or by intracloud components. Due to a combination of magnetic direction finding and time-of-arrival techniques, only two sensors are needed to construct the horizontal location (latitude and longitude, no altitude) and time of a signal. NLDN locates total lightning, including CG and IC discharges. According to Vaisala (2013), up to 95 % and better than 50 % of all CG and IC lightning, respectively, is detected. Zhu et al. (2016) found that one third of 153 IC pulses were detected by NLDN, and 86 % were classified correctly. NLDN detected 92 % of 367 return strokes, and also 92 % were correctly classified as CG. The median location accuracy approaches 250 m for CG strokes in the interior of the network. Lightning can be located at long range (1500 km), but the location accuracy in the interior of the network is significantly higher than outside. NLDN measures also the peak amplitude of the LF source. NLDN data used in this study include time (resolved at 1 ms), the location as latitude and longitude, the peak amplitude [kA], the polarity, the type (CG or IC) of the LF source, and quality parameters, e.g., the location error ellipse axes. Although Vaisala merges strokes to flashes (within 10 km and 1 s), this study retrieves NLDN flash level data using the algorithm developed by Erdmann et al. (2020) for Meteorage records in France. Hence, pulses/strokes are merged into a flash

if they occur within both 20 km and 0.4 s. The dataset is not further separated in this work, and the term *pulse/stroke* is used to represent all NLDN detections on the stroke-pulse level.

c. Database for the current study

The general dataset consists of six months of GLM and NLDN records, from March 15, 2018 to September 15, 2018. NLDN data were provided in a region between 30°N and 35°N and 95°W and 82°W. GLM data before 26 September 2018 need a time-of-flight (TOF) correction that takes into account the time lightning photons need to travel from the cloud tops (approximated at 10 km of altitude) to the GLM orbit. Our study applies a dynamical TOF correction with values ranging from 122.8 ms to 124.9 ms in the region of interest.

In order to handle the large amount of GLM data and hence to limit the data processing time, a reduction of the 6-month dataset was necessary. The complete lightning dataset is studied to identify lightning-active days (start and end at 00 UTC), defined by the number of GLM flashes and the number of GLM events. Ten days with significant lightning activity and different storm types during both day and night are selected. Table 1 summarizes the number of GLM events and flashes as well as NLDN pulses and strokes and flashes recorded in the region during each of the ten selected days. Table 1 also states the dominant weather situation during each of the ten days. At least one day per month is selected to represent possible climatological differences of the lightning within the region. All further analyses use these ten days in order to reduce the immense amount of GLM event scale data. The resulting dataset comprises 1,133,671 GLM flashes and 1,115,675 NLDN flashes. Missing data is identified through an analysis of instrument activity during 20 s time windows equal to those of the GLM L2 data files. The amount of flashes is reduced to 1,132,051 GLM flashes

and 1,115,585 NLDN flashes due to possible[1] short periods of instrument inactivity. Hence, the difference in the number of observed flashes is less than 2 % of the flash counts, and both instruments operated continuously during the selected days. As the effect of downtimes of an instrument can be disregarded, the following analysis uses all available data. Three among the ten days are chosen to test the generators with uncorrelated data and to assess the variability in the results (test days). The test days (07 April 2018, 26 May 2018, and 31 July 2018) feature both thermally driven convection and dynamic forcing at air mass boundaries. In the following, the different weather regimes with different lightning activity are briefly described for the test days as the final FED product is in fact only analyzed for these three days.

For instance on 07 April 2018, the weather was dominated by a major cold front that traversed the region from northwest to southeast. Temperatures dropped by about 10 K behind the front. The strong dynamic forcing caused a mesoscale convective system (MCS) with linear structure. This system produced the vast majority of flashes observed during the test period of 07 April 2018 until it left the studied region at about 12 UTC.

26 May 2018 was characterized by relatively warm surface temperatures with slightly decreasing temperatures from west to east within the region. Moisture was induced into the region by a weak tropical depression over Cuba and later southern Florida. Convection occurred mainly in the local afternoon as a result of surface heating. Well defined cells formed and propagated slowly southward in the cyclonic flow.

[1]We do not know instrument downtimes from the data. Data may come with flags, but they do not give reliable information about the instrument status. We used a two-step approach to identify downtimes; (i) the flash DE is less than 50 % and (ii) the number of flashes observed is less than 10 % of the reference LLS.

Daytime temperatures widely exceeded 30 °C and remained at about 25 °C at night within the region on 31 July 2018. Moisture was advected into the region from the Gulf of Mexico while a dry line approached from the northwest. A multicell storm cluster formed in the convergence zone at local nighttime and propagated eastward driven by a short baroclinic wave aloft. The second peak of lightning activity results from thermal convection in the eastern portion of the region before the dry air moved in and inhibited further convection.

d. Data processing algorithms – Flash scale data and identification of matches

NLDN and GLM observe lightning independently of each other. The comparison of the two LLSs needs, however, coincident observations. This work uses the matching algorithm introduced by Erdmann et al. (2020). Coincident observations are defined at the flash scale for flashes within 20 km and 1.0 s. The criteria are tested for any pair of events and pulses/strokes. Two parent flashes are matched if one event (pulse/stroke) meets both the spatial and the temporal criteria to any pulse/stroke (event) of the given flash. The algorithm does not analyze the flash mean position but the event and pulse/stroke locations.

GLM flash level data are included in the GLM L2 science data and emanate from NASA's GLM L2 clustering algorithm. Mach (2020) found recently that NASA's GLM clustering algorithm was quite stable for different spatial and temporal merging criteria (mainly for storms with flash rates below about 40 flashes per minute). In the present study, the performance of NASA's GLM L2 clustering algorithm for one hour on May 26 was investigated. NASA's L2 GLM clustering algorithm succeeded to merge many events and to detect large flashes. The GLM operational algorithm still limits the maximum size of flashes due to computational restrictions. However, such cases are rare and hardly influence the data generators as statistical approaches are used for both training and testing here.

The matching of GLM and NLDN flashes (for the 10-day dataset) leads to 948,872 GLM and 971,102 NLDN flashes with match. Some flashes from one system are matched to more than one flash in the other system, and it happens more often that one GLM flash matches multiple NLDN flashes than vice versa. Considering the total number of GLM (NLDN) flashes, the relative flash DE is defined as ratio of flashes observed by both given and reference LLSs to the total number of flashes observed by the reference LLS. It yields 87.0 % (of 1,115,585 NLDN flashes) and 83.8 % (of 1,132,051 GLM flashes) for GLM and NLDN, respectively. Figure 1 illustrates the flash DE of both GLM and NLDN within the studied region, along with 2D density of observed flash centroids (gray iso-contour). The flash DE remains consistent within the entire domain. The local minimum in the northeast is caused by a low number of observed flashes for the two $1^\circ \times 1^\circ$ pixels in Figure 1. The high flash DE of GLM agrees with the results of Marchand et al. (2019), who found the GLM DE relative to ground-based Earth Networks Total Lightning Network (ENTLN) flashes exceeding 80 % for most of the southeastern CONUS. They used 35 km and 330 ms as spatial and temporal matching criteria, respectively. Murphy and Said (2020) compared among others GLM and NLDN relative DE, matching flashes within 20 km between GLM flash centroids and the first NLDN pulse/stroke per flash and 200 ms between the flash time windows between the start and end times, and report similar flash DE values on the large scale in the southeastern CONUS. A new approach to the GLM flash DE and false alarm ratio (FAR) is introduced by Bateman and Mach (2020); Bateman et al. (2021): combining several ground-based networks to provide reference data, and using coarse matching criteria of 50 km and 10 min, they found flash DE of over 90 % and FAR just above 5 % for the GLM on GOES-16.

3. Methods

This section defines the concepts and the strategy to generate GEO pseudo lightning observations. The methods are designed to use NLDN data and evaluated using real GLM observations. MTG-LI will provide total lightning observations with similar data structure as GLM observations. It will also consist of events, groups, and flashes. Although MTG-LI's spatial resolution (4.5 km at nadir versus 8.0 km at nadir) and the temporal resolution (1 ms versus 2 ms) will be higher than those of GLM, the methods presented here can still be applied to simulate MTG-LI observations. A comparison of ISS-LIS records over the domain of this study (USA) and the target region (France) revealed statistically similar flash characteristics (Erdmann 2020, thesis, Chapter II.1-2). In addition, for both regions of interest, statistics on NLDN and Meteorage LF lightning observations relative to ISS-LIS records were consistent. The FED as explained in the following section is simulated on a 5 km $\times$ 5 km resolution grid approximating the MTG-LI grid.

a. Definition of the Flash Extent Density (FED)

Flash extent density is a gridded product, summing over a given time integration period, the projections of the location of flash components, e.g., events and pulses/strokes on a given regular grid mesh. FED pixels with any lightning observation are identified, while pixels with multiple observations (e.g., multiple NLDN pulses/strokes) are counted once per flash. This gives a grid of pixels with either lightning (value 1) or no lightning (value 0) for each flash. The FED product considers all flashes within a given time integration period and sums up the occurrence of flash components per pixel. Hence, the FED product can have values greater than or equal to one flash per pixel. It shows the spatial distribution of

lightning activity within the given time period. For example, the propagation of convective cores can be tracked over several successive time integration periods.

The FED in this study is calculated on a regular latitude longitude grid with an average pixel size of 5 km × 5 km. To obtain the regular latitude longitude grid, the distance of 5 km is transformed to latitudinal and longitudinal distance as of the pixel at the center of the study region. Appendix A describes the details to transform GEO pixel grid to the regular FED grid. In the present study, FEDs are analyzed per 60 min time integration periods. The 1-hour period maintains information to locate tracks of convective cores and most electrified regions while it is also long enough to capture several storms distributed within the full domain. There might be, however, multiple storms at one location during 60 min. The FED integration period can be changed as needed since our GEO lightning pseudo-observation generator simulates data at the flash level. The sum of multiple short FED periods is equal to the FED of a corresponding long period, but the computation of one long period is more efficient. Hence, this work simulates FED per hour for computational reasons. It should be mentioned, however, that other FED time integration periods are currently under investigation, and the assimilation of MTG-LI will use a shorter FED time integration period.

b. Work flow – The simulation of GEO pseudo-observations of FED

The simulation of pseudo-GLM flashes from NLDN observations is performed in 2 parts. First, our GEO lightning pseudo-observation generator uses the flash database with the coincident GLM and NLDN flashes and their characteristics. This part called target generator employs ML techniques. It is based on statistical relationships between the NLDN characteristics (features) and the characteristics of the concurrently observed GLM flashes

(targets). The target generator is detailed in the following section. This part is conducted using different approaches, which will be explained thereafter. They include simple linear regressions as well as more sophisticated ML models. The second part of the GEO lightning pseudo-observation generator, described in the last section here, simulates pseudo-GLM events using the simulated GLM flash characteristics.

1) Simulate pseudo-GLM flash characteristics

Coincident NLDN and GLM flashes are analyzed regarding their characteristics including the flash extent and flash duration (both GLM and NLDN) as well as the event number per flash (GLM) or pulse/stroke number (NLDN) per flash. The flash extent is a characteristic distance for the illuminated area for GLM or simply the distance between point sources for NLDN. It sums up the distance between the lowest and highest latitude (the North-South [NS] extent) and the distance between lowest and highest longitude (the West-East [WE] extent) of events or pulses/strokes of the flash. GLM flash extent relies on the pixel center position but does not include the pixel extensions. Single pixel GLM flashes and single pulse/stroke NLDN flashes have an extent of 0.0 km. Flash duration is defined as the time between the frames; therefore, a single frame features a flash duration of 0.0 s, i.e., GLM flashes with all events at the same time and NLDN flashes with all pulses/strokes at the same time. The maximum and mean signal strengths, defined from the LF peak currents and radiant energies as measured by NLDN and GLM, respectively, are evaluated per flash to represent flash energetics. In addition, a CG stroke ratio is calculated for NLDN flashes dividing the number of CG strokes of the flash by the total pulse/stroke number. Previous studies (e.g., Thomas et al. 2000; Marchand et al. 2019; Erdmann et al. 2020; Murphy and Said 2020; Rutledge et al. 2020) found that characteristics of flashes observed by optical

satellite LLSs depend among others on the flash altitude. Flash components identified as CG strokes propagate on average at lower altitudes than the IC components. In total, there are 5 GLM flash characteristics (flash duration, event number per flash, flash extent, mean and maximum event radiant energy per flash) and 6 NLDN flash characteristics (flash duration, pulse/stroke number per flash, flash extent, mean and maximum LF amplitude per flash, CG stroke ratio). Details on the distributions of the flash characteristics are provided by Erdmann (2020, Chapter II.3.4).

Linear regressions between any two GLM and NLDN flash characteristics showed that GLM flash duration has Pearson correlation coefficients R above 0.64 to NLDN flash duration and the number of pulses/strokes per flash. GLM event number per flash and GLM flash extent feature R of 0.08 to 0.43 to the complete set of features. Mean and maximum event radiant energies per GLM flash are not correlated to any NLDN flash characteristic on the flash scale and then not relevant for synthetic FED generation. Hence, they are excluded from the ML targets. The remaining targets are GLM flash duration, event number per flash, and flash extent.

Building the GEO lightning pseudo-observation generator requires independent generator building (GB) and generator testing (GT) data for the generator design and for the verification of the generated product, i.e., the FED, respectively. The split of our dataset is illustrated in Figure 2. The GB data consist of seven days and the GT data of the remaining 3 days (test days) of the full dataset (see Section 2.c and Table 1). The GB includes an ML part. Here, only matched flashes are considered in order to compare feature and target values (see Figure 2). Features (input data) of the ML are the six NLDN characteristics, and targets (output data) are GLM flash duration, event number per flash, and flash extent. Feature and target sample sizes are given as the number of matched flashes detected by

GLM and NLDN, respectively, and are not equal in general (Section 2.d). Since training the ML models requires the same sample size for the features and targets, two (or more) flashes matched to the same flash of the other LLS are merged, and characteristics of the merged flashes are combined. The resulting ML data (dark orange in Figure 2) consist of 672,794 flashes, each sample with six NLDN features and three GLM targets. The ML part further splits this set of ML data randomly into independent ML training and ML validation data at a ratio of 90 % to 10 %. The ML models are thus trained with 605,515 flashes. The ML validation data serves to calculate goodness-of-fit scores for each applied ML technique. Then, the different ML models are compared and the model parameters (e.g., the number of trees or the number of neural network layers, see Appendix B.a) are tuned based on the scores. The 3-day GT dataset is used to evaluate each generator as a whole including the ML and event generation parts. The test exercise exploits both observed GLM and generated, NLDN-based pseudo-GLM datasets as two independent populations.

The generator simulates one pseudo-GLM flash for each observed NLDN flash. Thereby, it is assumed that flashes detected by GLM only and detected by NLDN only compensate each other. The assumption was justified as (i) GLM and NLDN feature flash DEs on the same order, (ii) both GLM-only and NLDN-only flashes were smaller in extent and shorter in duration than the flashes with coincident observations (see also e.g., Zhang and Cummins 2020; Erdmann et al. 2020), and (iii) most GLM-only and NLDN-only flashes were found in the same regions in proximity to convective cores where high flash rates were observed (as in Zhang and Cummins 2020). The overall GLM and NLDN flash numbers (see Table 1) vary by only a few percent. However, it can be seen that there are days and cases where NLDN detects more flashes than GLM, i.e., 07 Apr. and 14 Apr. and other days where GLM detected more flashes than NLDN, i.e., 26 May, 03 Jun., and 07 Aug.

Here, only GLM flashes will be simulated and only if there are NLDN records. The algorithms do not distinguish potential NLDN flashes that would not be detected by GLM. In addition, there is no algorithm developed to create flashes only detected by GLM. For those two configurations, developing dedicated algorithms would require taking into account the microphysical properties of the cloud profiles, but also a model that would generate the lightning activity as realistically as possible to mimic GLM-only and NLDN-only flashes. The goal of the lightning generator is to provide synthetic LI records with a better representativeness than what has been used so far, knowing that there are some limitations in our models, to develop a new proof of concept to assimilate space-based lightning observations. Another aspect concerns the detection of optical flashes at day and night. One can consider to develop a GEO pseudo-observation generator for both day and nighttime with potentially different relations between LF flash characteristics and GEO flash characteristics. However, as this paper includes a variety of methods and the first approach to use ML techniques to simulate GEO flashes, day and nighttime flashes are not separated. This also is the case for flashes over land and sea.

The aforementioned assumption means that flashes detected by NLDN only are treated similarly to those coincidently detected by both NLDN and GLM. As the number of NLDN-only flashes is significantly lower than the number of NLDN flashes with GLM match (given a GLM flash DE relative to NLDN of 87 % for the full 10-day dataset), the assumption only affects about 13 % of the simulated flashes. Statistics of GLM targets and FED fields inferred from the generated pseudo-GLM flashes are compared to those from all observed GLM flashes during the 3 days.

The comparisons of statistics of the observed and simulated targets include the distribution mean, median, minimum, and maximum. The root mean squared error (RMSE) between

characteristics of individual (simulated and real) GLM flashes is also computed, but only for the 295,313 NLDN flashes with GLM match (representing a GLM flash DE of 86.7 % for the test days). The evaluation makes an addition use of two statistical scores that are defined for the cumulative (in fact empirical) distribution functions (CDFs): the Kolmogorov-Smirnov statistic (KS, Massey 1951) and the Cramér-von Mises criterion (CvM, Anderson 1962) measure the distance between the observed and simulated CDFs of the targets. Both the KS and the CvM tests can verify the null-hypothesis that two samples belong to the same population.

2) ML-based target generators relating NLDN flash characteristics to GLM flash characteristics

The previous sections explained that our GEO lightning pseudo-observation generator consists of 2 parts, the ML-based target generator and the simulation of GEO pseudo-events. Appendix B.a briefly describes the different ML models used in the ML-based part of this generator. The ML-based algorithms relate NLDN flash characteristics to GLM flash characteristics in this work. Hence, all ML models are supervised models with the same training data. The models emanate from Python's scikit-learn library (sklearn, Pedregosa et al. 2011).

This study uses seven different ML model types (details in Appendix B.a): Multivariate linear regressions (LinReg), third-degree polynomial regressions (Poly), Extra Trees Regressors (ETR) as a form of Random Forests, Bagging with k-nearest neighbor regressors (BAGR KNN), Multilayer Perceptron neural networks (MLP), linear Support Vector Regressors (linSVR), and Histogram Gradient Boosting Regressors (HGBR).

3) Multi-step Approach

Targets of a multi-target ML training can be correlated, e.g., GLM event number per flash is strongly correlated to GLM flash extent with R of 0.74. To the best knowledge of the authors, models of Python's sklearn library do not take advantage of correlations between targets. Indeed, the so-called single target (ST) approaches do not consider correlations between targets, however, such correlations can help to improve the skill of ML models and thus the prediction of the generators. Borchani et al. (2015) summarize methods to deal with multi-target regressions and take advantage of correlations between targets. Their paper compares the ST approach to multiple multi-target approaches, e.g., multi-target regressor stacking (MTRS), regression chains (RC), multi-output support vector regression, multi-target regression trees, and rule methods. Spyromitros-Xioufis et al. (2016) introduced the stacked ST (SST) and ensemble RC (ERC). These methods can be computationally complex with high memory costs (Mastelini et al. 2019). As Aguiar et al. (2019) state, choosing the most suitable approach needs previous testing and depends on the task. The methods cited here are computationally expensive.

The flow chart in Figure 3 shows a computationally efficient multi-target approach that simplifies the SST. As a starting point, there are NLDN features and GLM targets as input for the ML training. The approach combines ST models (Figure 3a) of three classes (colored) for the training. The application case only uses the NLDN features as first input. Therefore, a multi-step approach is required. An application example is shown in Figure 3b. More details about our approach can be found in Appendix C.

In summary, the multi-step approach modifies the ML input feature set selection and thus the configuration of the corresponding generator. It is a form of multi-target regression that

can take advantage of correlations between the ML targets. Appendix B.b summarizes the available feature set selections for the ML as different configurations of generators. Figure 3b shows just one example of the application that is also detailed in Appendix C. Section 4 will demonstrate whether the additional GLM pseudo-features can help to tune the pseudo-GLM simulation towards observed GLM data.

4) Applied Scaling Methods

This study normalizes features to the [0,1] range with a Min-Max-scaler:

$$X_N = \frac{X - \min X}{\max X - \min X} \tag{1}$$

where X is a data vector, $\min X$ and $\max X$ define the minimum and maximum of X, respectively, and the resulting normalization X_N ranges from 0 to 1.

The targets are scaled with a common standard-scaler (also called z-value scaling) defined as

$$Z = \frac{X - \text{mean}\, X}{\text{std}\, X} \tag{2}$$

where X is a data vector, $\text{mean}\, X$ and $\text{std}\, X$ are the mean and the standard deviation of X, respectively. The resulting standardization Z is centered around 0.

The Min-Max-Scaler is an alternative standardization method that is more robust to small standard deviations and for different feature ranges than the common standard-scaler (sklearn documentation).

Some generators perform well with unscaled data (i.e., direct input of data with physical units) used as a reference input method during the ML part. All results presented in this paper are re-scaled to physical units.

5) Generate pseudo-GLM events

The studied domain is separated into regular adjustable size latitude-longitude pixels that represent the pseudo-GLM pixel matrix. Any given latitude-longitude position is projected on that pixel matrix to determine the corresponding pixel and thus the shape of one pseudo-GLM event. Using a regular grid simplifies and speeds up the simulation of pseudo-GLM events significantly. Each regularly shaped pseudo-GLM event covers an area equal to the average size of the observed, irregularly shaped GLM events in the region of interest. Analyzing simulations built on this regular pseudo-GLM grid should lead to statistically similar results as for the irregular grid of the GLM observations.

The target generator of the GEO lightning pseudo-observation generator simulates the targets based on the given NLDN flash characteristics. These pseudo-GLM targets provide the information to derive individual pseudo-GLM events. As the target generator may produce targets with values smaller than the observed (and physical) limits, the targets are adjusted to account for the known thresholds. For instance, negative flash extent and flash duration are set to zero, and there are at least 2 pseudo-GLM events per flash in accordance with NASA GLM data processing (Mach 2020). Pseudo-GLM flash NS and WE extents are calculated based on the simulated pseudo-GLM flash extent applying the same ratio as the NS and WE extents of the corresponding NLDN flash. If the NLDN flash contains a single pulse or stroke, the NS to WE ratio is set to one.

First, the locations of pseudo-GLM events are generated. Using the simulated pseudo-GLM flash extent and its NS and WE components, a rectangular sub-domain on the pseudo-GLM pixel matrix is defined. The center of this sub-domain houses the NLDN flash position centroid and the corresponding pixel constitutes the first event of the pseudo-GLM flash.

Any pixel within the sub-domain may also become a pseudo-GLM event of this pseudo-GLM flash. Three constraints have been designed to generate subsequent pseudo-GLM events: (i) each event of the flash has at least one adjacent or diagonal neighbor within one flash, thus, avoiding spatial gaps; (ii) pixels are primarily selected starting at the first event and propagating (meaning increasing distance to the first event) towards the sub-domain border to approximate the simulated flash extent; and (iii) additional pixels can be selected randomly within the rectangular area until the simulated event number is reached. In consequence one single pixel of the sub-domain can contain more than one pseudo-GLM event. Since pixels of the sub-domain are not guaranteed to contain a pseudo-GLM event, this random selection also affects the final FED product.

Then, the pseudo-GLM events get time stamped. In the present study, the matching of GLM and NLDN flashes revealed that the median time offset between the mean time of a given NLDN flash and the mean time of the matched GLM flash was about 8 ms. The NLDN and GLM average flash duration were 0.24 s and 0.39 s, respectively. Hence, the mean time of matched NLDN and GLM flashes are relatively close while GLM flashes last on average longer than NLDN flashes. As a consequence, the mean time of the NLDN flash defines the mean time of the pseudo-GLM flash that is also the time stamp of the first pseudo-GLM event. Our generator is built to generate realistic FED fields. Only the spatial distribution of the events is needed to infer FED. Hence, the temporal occurrences of pseudo-events are uniformly and arbitrary distributed during the duration of one flash. Pseudo-event times are then rounded to the time frames of the mimicked GEO LLS, i.e., to 2-ms-frames for pseudo-GLM data. The only constraint is that any adjacent pixel occurs within 330 ms (i.e., the time criterion to separated flashes in NASA's GLM L2 algorithm). One 2-ms-frame contains often several pseudo-GLM events.

4. Results

Figure 4 shows the example of one simulated pseudo-GLM flash created with the final GEO lightning pseudo-observation generator based on a *linSVR* model, the corresponding GLM and NLDN observations, and the observed and simulated GLM flash characteristics. One can see the difference between the real GLM grid and the regular pseudo-GLM grid of the simulation (Figure 4(c)). The difference between observed and simulated flash extent is within the size of one GLM pixel for this example. The simulated flash duration exceeds the observed flash duration significantly. There is also an overestimation of the number of GLM events by the generator.

Results are obtained from the 3-day test dataset. It contains 340,712 NLDN flashes that are used to simulate the same number of pseudo-GLM flashes. Statistics of the pseudo-GLM flashes are compared to the statistics of all 338,579 observed GLM flashes. First, the distributions of the simulated and observed GLM flash extent, flash duration, and event number per flash are compared. The best target generators are used to simulate pseudo-GLM events and eventually compute the pseudo-FED product. The FED is analyzed statistically for both observed and simulated GLM data of the three test days. The minimum discrepancy between observation and simulation will indicate the most suitable target generator configuration for the final GEO lightning pseudo-observation generator.

a. Evaluating the target generators – Distributions of GLM flash extent, flash duration and event number per flash

In a general sense, a wide range of values is observed for all target distributions. The GLM flash duration ranged from 0.0 s to 16.4 s. Observed GLM flashes comprised between

2 and 1395 events. The test data features GLM flash extent between 0 km and 277 km. The target generators should handle these ranges of values and predict target statistics similar to the statistics of observed GLM flashes.

Table 2 summarizes the findings , with statistics, the KS, and the CvM of the distributions of observed and simulated GLM flash duration, event number per flash, and flash extent for the full 3-day test data. The table contains distribution statistics for the respective target generator with smallest difference between observed and simulated characteristics over the test period. Results for the *linSVR num ext(a) plus* generator are shown as reference. Statistics of the simulated pseudo-GLM and the observed distributions are referred to as simulated statistics and observed statistics, respectively.

This analysis was also conducted for each test day. The results are presented in Appendix D.a.

The majority of the target generators features mean values similar to the observed means for all three target characteristics. The simulated medians, however, exceed the observed medians in most cases, especially for the number of events per flash, suggesting a tendency to overestimate the target values. The previously described behavior is true for all but the *linSVR*-based generators. *linSVR* filters the dataset in advance to build the prediction on the support vectors (Appendix B.a.(vi)). That results (in this study) in lower differences between the simulated and observed median values as compared to using the other ML model types. The mean values of *linSVR*-based predictions, are, however, often smaller than the observed mean, especially for the event number per flash. Table 2 demonstrates this behavior of *linSVR*-based generators. To detail one example, the recommended *linSVR*-based generator (see Section 4.b, bold in Table 2) underestimates median and mean flash extent by about 4.5 % and 11.7 %, respectively. The mean event number per flash is also underestimated

by about 29.6 %, however, the median event number per flash is overestimated by 20 %. The *linSVR*-based generator creates, compared to the observations, not enough flashes with an event number in the tails of the distribution, i.e., close to the observed minimum and maximum event numbers. Hence, it cannot mimic the full range of the observed event numbers per flash. This *linSVR*-based generator still outperforms all other generators with respect to the median considering the full 3-day test data.

Some general conclusions can be drawn regarding the generator performances for the observed range and variability of the target values. The target generator minimum often approximates or slightly exceeds the observed minimum, whereas the maximum is underestimated in most cases. This particular behavior can even be seen for the best target generators (Table 2) because the number of small flashes with characteristics close to the minimum observed target values is relatively high. The rare, highest observed values are often underrepresented in the statistical approach. It is further found that observed GLM flash statistics can vary for a given set of the six observed NLDN features. This is the case as our six NLDN features cannot completely explain the range of target values even if the statistics derived here are significant in terms of the large sample size. The large values of the RMSE per flash in Table 2 and also Appendix D.a result from the deterministic nature of the ML models in combination with this lack of information in the features, e.g., cloud properties. The RMSE values of the GLM flash extent are similar to the mean values, whereas they reach twice the mean for both GLM flash duration and event number per flash. Here, the optimization of our GEO lightning pseudo-observation generator for FED that depends mostly on the flash extent is evident. A relatively wide range of target values is in particular found for small NLDN flashes with NLDN pulse/stroke number, extent, and duration near the lower end of the distributions (not shown). Large (meaning long extent, long duration,

and many pulses/strokes or events) NLDN flashes usually coincide with large GLM flashes. As the NLDN features are somewhat correlated to the GLM targets, the high RMSE due to a small NLDN flash as input also leads to a high RMSE when predicting small GLM flashes.

KS and CvM assign a quantitative value to measure the distance between two samples. While KS is normalized (values of 0 to 1), the CvM value depends in general on the distance between simulated and observed CDFs and the sample size. As the sample size is kept constant for all generators, CvM in fact provides a common measure of the agreement between observed and simulated targets. Both KS and CvM feature lower values for the GLM flash duration than for both the GLM flash extent and the GLM event number per flash considering the full test dataset (Table 2). This result is in accordance with the strong correlation coefficients between observed GLM flash duration and NLDN features (see also Section 1). KS and CvM for the flash duration rely mainly on the underestimation of long duration flashes. As an exception, the recommended *linSVR num ext(a) plus* generator not only underestimates the maximum flash duration but also cannot produce single-frame flashes. Therefore, KS and CvM are higher for the flash duration than for the flash extent here[2]. KS and CvM reach their highest values, i.e., when comparing the 3 target distributions, for the GLM event number per flash, for which the weakest correlations to features were observed.

The presented selection of generator performance for the 3-day test data and each test day (Appendix D.a) indicate that the choice of a suitable target generator can be situational. The objective now is to find a configuration that best approximates the observed GLM flashes and target distributions. Therefore, the differences between the simulated and observed

[2]It can be noticed from Table 2 that this *linSVR*-based generator also overestimates the minima of event number per flash and flash extent. For those two targets, the simulated maxima are closer to the observed maxima than for the best performing generator in Table 2 causing overall similar and even lower KS and CvM for the *linSVR num ext(a) plus*.

statistics (i.e., mean, median, minimum, maximum, RMSE, KS, CvM) are calculated and normalized for each statistic. The normalization divides each absolute difference by the maximum absolute difference of all target generators for a given statistic. A value of 1 represents the worst target generator for the given statistic, while a value of 0 indicates no difference to the observation. In addition, and to summarize all the information, the so-called Normalized Difference Average (NDA) is introduced to average the normalized absolute differences and scores for a given generator. The perfect generator would yield an NDA of zero. NDAs of the target generators can be directly compared in order to identify the highest performer. NDA is calculated per target and for all three targets overall.

Overall NDAs for all three targets range from 0.35 for the *linSVR num ext raw* generator to 0.87 for the *MLP num ext(a) raw* generator. The best (i.e., lowest NDA) 24 target generators all use a *linSVR*, and the performance of the best target generators varies only within the range of uncertainty given in Section 1. For example, the difference between the 1st and 10th ranked target generator is only 0.04 NDA. The NDA ranking of target generators reveals a clustering explained by the ML model type, with *linSVR*-based generators performing the best, followed by *BAGR KNN dist*-based, *ETR*-based, and polynomial regression-based generators. MLP- and HGBR-based generators exhibit the highest NDAs.

The generators yielding the lowest NDA values are mostly those using the multi-step approach. In addition, the use of all six (*plus*, Appendix B.b) instead of only four (*default*) NLDN features improved the performance of the majority of tested generators. The feature and target scaling had little effect on the generator performance, although scaling is usually recommended for ML applications. The ML model type has in fact the highest impact on the simulation of pseudo-GLM flashes and thus on the target generator performance.

Figure 5 visualizes the statistics of all tested target generators for the flash extent as the most impactful characteristic on FED. It groups the results for each statistic by ML model type. Seven ML model types were used to build the generators (Section 2 and Table B1 except *RF*). Each distribution contains the results of 28 generators using this ML model type including 7 feature set selections and 2 optional attributes (Appendix B.b, Table B2). Figure 5 shows these results as normalized differences and scores for the 3 test days combined. It reveals that the boxplot minima for the *linSVR* type generators are the closest to zero for most statistics. *BAGR KNN dist*-based generators feature the second lowest values of KS and CvM. The finding is supported by results for each test day (Appendix D.a) showing best performances for the targets by *BAGRN KNN dist*-based generators on 07 April 2018 and by *linSVR*-based generators on 26 May and 31 July 2018. Some boxplots exhibit a wide range of outcomes. The range shows that all ML model types are sensitive to the configuration. The NDA of the best generator, i.e., *linSVR num ext(a2)*, equals 0.28. The associated outcomes for flash duration and event number per flash statistics (AppendixD.b) confirm *linSVR*-based generators as most suitable to simulate GLM targets for the entire test period. Hence, results for the individul targets agree with the overall NDA analysis.

1) Confidence in the results

The confidence in the outcomes is evaluated for the two parts of the GEO lightning pseudo-observation generator. The uncertainty of the outcomes is expressed as the range of outcomes given the same configuration. First, (only) three selected generators with constant configuration are trained 10 times using the same full training dataset (for computational efficiency). Herewith, the training variability of the ML model is assessed. The selected generators are *BAGR KNN dist num plus*, *MLP alpha8 num raw*, and *linSVR num ext(a)*

plus. They are labeled by their ML model type as *BAGR KNN dist*, *MLP*, and *linSVR*, respectively. Figure 6 shows the distributions (boxplots) of targets for the full test data for the 3 generators (x-axis) each trained 10 times for pseudo-GLM flash duration (a), pseudo-GLM number of event per flash (b), and pseudo-GLM flash extent (c), respectively. The predicted target range of the 10 trained generators is smaller than the variability due to different ML model types and due to different configurations of one ML type. The 10 *BAGR KNN dist* based simulations feature a very narrow range of outcomes for all statistics. The 10 trainings of both this *linSVR* and this *MLP* yield a range of values of 0.2 to 0.4 normalized absolute difference for most statistics. The range of the minimum event number per flash (Figure 6b) and the minimum flash extent (Figure 6c) reaches about 0.5 and up to 0.7 for the *linSVR* and *MLP*-based generators, respectively. This range is in the order of magnitude as the variability enforced by using different ML model types. The range of normalized absolute difference for the maximum event number predicted based on 10 equally configured *linSVR* models is also about 0.6. In addition, the range of normalized absolute differences is always wider for the mean than for the median. Despite a relatively high uncertainty in some statistics, the overall trends as described in the previous section remain valid. Statistics sensitive to distribution outliers, i.e., the mean and minimum, exhibit higher uncertainties due to the GB than more robust statistics, i.e., the median, KS, and CvM. Some target generators, i.e., the *BAGR KNN dist*-based one, appear to provide very robust predictions. The uncertainty range is usually smaller than the overall range of values for each statistic.

The test of the variability in the results enforced by the second part of the GEO lightning pseudo-observation generator, i.e., generating pseudo-GLM events (not shown) is much

smaller than for the ML part. Hence, the overall range of targets for a given generator configuration is similar to those shown in Figure 6.

b. Evaluating observed and simulated Flash extent density (FED)

Hourly FED maps are calculated for both GLM observed and simulated flashes. They will be referred to as observed and simulated FED, respectively, in the following. The evaluation includes the hourly FED summed-up over the domain (termed FED sum), the electrified areas defined as pixels with positive FED (i.e., greater than 0 flashes per 5 km × 5 km pixel per hour), and a visual inspection of convective cores. As the choice of the ML model type has the highest impact on the overall performance of the GEO lightning pseudo-observation generator, the results are mainly discussed with respect to the ML model types.

Figure 7 presents the observed FED (a) to the simulated FED of 3 selected generator configurations (b-d) for the example of 26 May 2018, 20:00 to 21:00 UTC. The 3 generator configurations represent a selection of the variety of tested generators with different ML model types, feature set selections, and scaling, as detailed below (see also Appendix B). Simulated FED fields capture the coarse geographical distribution of the observed FED. One can identify the most active regions (highest FED values), that are situated at similar locations for the observed and simulated FEDs. The numbers in the top corners of the panels (a)-(d) indicate the number of lightning pixels with FED>0 flashes per 5 km × 5 km pixel per hour on the left and the FED sum on the right. The product of the number of lightning pixels and the area per pixel yields the electrified area. The *linSVR* (*num ext(a) plus*, Table B2) in Figure 7b uses GLM duration as additional feature when simulating GLM number, and then GLM duration and GLM event number to simulate GLM extent. This *linSVR*-based generator performs among the best for the simulation of GLM targets overall,

and it appears to be among the best also for the FED sum. It underestimated the electrified area in most cases (as in the example in Figure 7a,b). The *MLP*-based simulation (*num raw*, Table B2) of the FED of Figure 7c uses unscaled features and targets. GLM flash extent and flash duration relate only to 4 NLDN features (without mean amplitude and CG stroke ratio). The attribute *num* means that pseudo-GLM flash duration and flash extent are obtained directly from ST approaches. Those simulated targets serve as pseudo-features to derive the pseudo-GLM event number. This *MLP*-based generator performs among the best for the electrified area, but overestimates GLM flash extent, GLM event number per flash, and eventually the FED sum. Figure 7d maps the FED as simulated by the *BAGR KNN dist* based generator (*num plus*, Table B2) that uses all six NLDN features and the *num* method (see above). It is the best performing generator using the *BAGR KNN dist* ML model type. Although this generator overestimates the target medians and the FED sum, it belongs to the best 25 % of generators for both FED sum and and electrified area. It performed best for 07 April 2018 test case with the dominant squall line that produced most of the large-extent lightning flashes. In general, all 3 generators overestimate the 1-hour FED sum in Figure 7. The *linSVR*-based generator simulates an FED sum significantly closer to the observed FED sum than using both the *MLP* and the *BAGR KNN dist*. The *linSVR*, however, underestimates the number of lightning pixels, that is best simulated by the *MLP*-based generator here.

The results are further investigated for the 3-day test period by comparing pixel-to-pixel simulated and observed hourly FED. Figure 8 shows the 2D-histograms, computed for the entire 3-day test dataset, for the same *linSVR* (a), *MLP* (b), and *BAGR KNN dist* (c) based generators as used in Figure 7. In general, the Pearson correlation coefficients R of 0.91 to 0.92 indicate well correlated distributions of observed and simulated FED. Figure 8 also

shows the range of simulated FED is wider than the range of observed FED (grey box). The corresponding trend to overestimate the FED in the simulation is proofed by the regression lines (light green) that feature steeper slopes than the equal-value line (black). In particular, the *MLP*-based (Figure 8b) and the *BAGR KNN dist* based generator (Figure 8c) overestimate the FED usually more than the *linSVR*-based generator (Figure 8a). Y-intercepts near 0 indicate good agreement for regions without lightning activity. These findings agree with the example in Figure 7.

To summarize and quantify the evaluation of both FED sum and electrified area, the metrics *normalized difference* D_{real} and *absolute normalized difference* D_{abs} are defined:

$$(a)\ D_{real} = \frac{\sum_{n=1}^{H} S_n - O_n}{\sum_{n=1}^{H} O_n} \quad \text{and} \quad (b)\ D_{abs} = \frac{\sum_{n=1}^{H} |S_n - O_n|}{\sum_{n=1}^{H} O_n} \tag{3}$$

where S_n is the simulated hourly FED sum or electrified area, O_n is the observed hourly FED sum or electrified area, and H is the total number of time steps (here 72 hours). D_{real} and D_{abs} can be used to compare the different GEO lightning pseudo-observation generators and identify the generator with the lowest difference to the observation.

All 196 generators are evaluated for D_{real} and D_{abs} of both FED sum and electrified area. As the ML part of the generator enforces significantly higher differences than the derivation of pseudo-GLM events (the second part), again results are mainly discussed regarding the different ML configurations.

D_{real} and D_{abs} are calculated for the 3-day test period. For the FED sum, the 28 *linSVR*-based generators tested are ranked as best 28 configurations in the comparison, i.e., lowest D_{abs}. Table 3 presents the results for the best 20 and worst 5 generators as ranked by D_{abs} of FED sum. The best GEO lightning pseudo-observation generators exhibit an D_{abs} of 22 % to 25 %, while D_{real} is close to zero, i.e., balance between situations with over- and under-

estimated FED sum. The worst generators (some of *MLP* and *ETR* based configurations) lead to almost twice as high FED sum as the observed values. Similar, positive values of both D_{real} and D_{abs} for the FED sum mean that most generators overestimate the FED sum. This agrees well with Figure 8. The exception is found for the *linSVR* type generators that often underestimate the FED sum with D_{real} ranging from $-22\,\%$ to $+39\,\%$. Figure 8a shows one example of a *linSVR* with positive D_{real}.

As mentioned, the best 28 generators for the FED sum are all of type *linSVR*. The best 10 generators use the multi-step approach (*num* and *num ext*, Table 3). The use of mean LF amplitude and CG fraction (*plus*) as additional NLDN features has a minor effect on the simulation of FED sum.

Results for the electrified area are in general closer to the observation than the FED sum. They are shown in Table 4 for the best 20 and worst 5 generators as ranked by D_{abs} of the electrified area. The generators with the lowest D_{abs}, *HGBR* type, differ absolutely by about $7.5\,\%$ from the observed electrified area. The vast majority of all tested target generators underestimate the electrified area (negative D_{real}). Multiple generators of various types feature D_{abs} of less than $10\,\%$, e.g., using *HGBR*, *Poly*, *BAGR KNN dist*, *ETR*, or *MLP* models. The *linSVR*-based generators, that performed best for the FED sum, exhibit the highest differences to the observation here with D_{abs} from $15\,\%$ to $35\,\%$ (all with negative D_{real}). For example, the best performer for the FED sum is ranked as third worst for the electrified area with a high underestimation of the area.

The best 20 generators for the electrified area take advantage of the multi-step approach in 15 cases. Also 15 of those 20 ML-based generators use all NLDN features (Table 4). Comparing only the *linSVR*-based generators, all 10 leading generators use 6 rather than

only 4 NLDN features. This result strengthens the meaning of including all NLDN features and of the multi-step approach.

The computational cost of our multi-step approach is still higher than the ST approach, however, only needed for the training of the generator. The application of trained multi-step generators is relatively fast, i.e., similar duration as applying ST generators. The best generator without multi-step approach (*linSVR(a) raw*) exhibits D_{abs} more than 14 % higher than the best generator for the FED sum (Table 3). In addition, D_{real} exceeds 23 % indicating that the FED sum is mostly overestimated. FED sum simulation is most sensitive to the choice of the generator and, hence, particularly important to obtain realistic synthetic FED. The multi-step approach helps in particular to obtain more realistic FED sum than ST-based generators. For the electrified area, however, generators not using the multi-step approach can perform as well as the best generators (Table 4). If only electrified area is of interest, common ST models can be used. The multi-step generator *linSVR num ext(a) plus* is successfully applied to simulate GLM FED (Section 4) and also MTG-LI FED over France (not in the present paper).

The recommended GEO lightning pseudo-observation generator balances the simulation of all pseudo-GLM target distributions, FED sum, and electrified area. It is named *linSVR num ext(a) plus* generator. This configuration features an overall NDA of 0.39, and an D_{abs} to observed FED sum and electrified area of 24.9 % and 21.3 %, respectively. This generator used all available features and utilizes the multi-step approach. First, GLM flash duration is predicted from all six NLDN features, and then used as additional pseudo-feature to predict the event number per flash. Finally, the pseudo-GLM flash extent is simulated from NLDN features and the pseudo-features GLM flash duration and event number. Both features and targets are scaled (Section 4). The *linSVR* ML technique is more time-efficient than the

MLP and bagging-based, e.g., *BAGR KNN dist* and *ETR*, techniques for the training and also needs less disk space to be stored. These are two other advantages of the *linSVR num ext(a) plus* generator.

Figure 9 presents hourly FED sum (a) and electrified area (b) with the overall value (1) and the difference to the observation (2) for 31 July 2018 test case. The observed FED and results for the 10 generators with lowest D_{abs} are plotted. Figure 9(a) includes in addition results of the best generator for electrified area (lime), and Figure 9(b) the results of the best generator for FED sum (orange). The figure also shows the number of hourly simulated pseudo-GLM flashes (histogram). Similar figures for the other two test days are also evaluated but not shown here because identical conclusions are drawn. The absolute values (Figure 9 panels 1) show that the FED sum (a) reacts directly to the number of (simulated) flashes. The electrified area curves (b) appear to have a time offset relative to changes in the flash number, suggesting that within 1 h a lower number of relatively large flashes can electrify a similar area as a higher number of smaller flashes. An increasing (decreasing) flash rate during the development (decay) of convective storms does not automatically mean a larger (smaller) electrified area, since even less flashes can still illuminate a large portion of the cloud via scattering. The simulated FED adapts this behavior very well. In particular, the simulated FED features similar hours with highest FED and electrified area as the observed FED.

It is observed that the simulated FED sum usually exceeds the corresponding observation during the phases of highest flash amounts within the region (Figure 9a). This could mean that NLDN detects significantly more flashes than GLM during these times, and thus the number of simulated flashes is significantly higher than the number of observed GLM flashes. These findings agree with Zhang and Cummins (2020), who found that the GLM

DE decreases for high flash rates and with shorter extent and duration flashes, which are observed during the mature phase of a thunderstorm.

It should be noted that the absolute values (Figure 9a1 and b1) and difference to the observation (Figure 9a2 and b2) for the FED sum (Figure 9a) have the same order of magnitude. In contrast, the difference (9b2) is one order of magnitude smaller than the absolute values (Figure 9b1) for the electrified area. Hence, the difference to observed FED and also the spread between generators with different configurations are much greater for the FED sum than for the electrified area. Therefore, it is decided to put more weights on the ranking of the FED sum than on the ranking of generators by electrified area when choosing the recommended generator. Eventually, the *linSVR*-based generator returns as the recommendation in an overall evaluation context. If, however, for a certain objective the electrified area is most important, several *HGBR*, *MLP* or even *ETR* based generators perform better than the recommended *linSVR*-based generator.

In a Monte Carlo approach, FEDs for 10 of in total 100 realizations of the recommended *linSVR* generator are calculated for the three test days. Figure 10 illustrate the median (line) and range (shaded) of FED sum and electrified area on 31 July 2021. The variability of both the FED sum and the electrified area has the same order of magnitude as the difference between the leading generators (Figure 9). Figure 10 also confirms that the *linSVR*-based generator tends to underestimate the electrified area. The vast majority of the time, all 10 realizations simulate lower electrified area than the GLM observations indicate. However, all 10 realizations remain relatively close to the observed FED sum at most times (except for the cases with intense convection, as discussed earlier). It should be noted that this *linSVR*-based generator does not appear among the best 10 generators for the electrified area (Figure 9b).

5. Summary

This study analyzed in detail the simulation of GEO lightning pseudo-observations in two parts: First pseudo-GLM flash characteristics are simulated and then pseudo-GLM events are derived. The data generator uses only LF ground-based data. There is no additional cloud information used in the generator. The entire process is non-trivial because relations (correlations) between characteristics of coincident LF ground-based and optical satellite lightning observations are often weak at the flash scale.

A multivariate analysis using several features and targets is conducted to achieve more robust flash characteristics. Simulated GEO flash characteristics (targets) are obtained via machine learning (ML) models. Targets include GLM flash extent, GLM flash duration, GLM event number per flash. An independent test data set is then introduced to compare the statistics of simulated pseudo-GEO flashes to the observed GEO, i.e., GLM, flash characteristics. In a second part, the simulated targets are used to mimic individual GEO events on a regular latitude longitude grid.

After testing different ML models used in the first part of our generator, a linear SVR (*linSVR*) based GEO lightning pseudo-observation generator is recommended. The results of multiple *linSVR* configurations turned out to be similar. In more detail, our recommendation is to use a *linSVR* with feature and target scaling, that uses all NLDN and pseudo-GLM features in a multi-step approach.

The type of the ML model chosen in the first part of our GEO lightning pseudo-observation generated has a major impact on the simulated flashes. In fact, the performance ranking of tested target generators reveals clusters per ML model type. Whereas the vast majority of generators produces pseudo-GLM flashes with flash characteristic means close to the

observed ones, they simultaneously overestimate the medians of flash characteristics. Therefore, they produce insufficient small flashes as compared to the GLM observations. Only *linSVR*-based generators were able to simulate pseudo-GLM flash characteristics with distribution medians close to the observation for the 3-day test dataset. This gain is achieved at the expense of slightly underestimating the target means. It is then found that FED sums from *linSVR*-based generators are closer to the observed FED sum than for all other generators, however, the electrified area is at least 10 % smaller than the observed electrified area.

Besides the type of the ML model, the set of features and the feature scaling impact the results. In particular, including (pseudo) GLM flash characteristics in the set of features improved the predictions of most ML models as target generators and thus the overall performance of the GEO lightning pseudo-observation generator.

In general, generators that perform well for the FED sum exhibit high D_{abs} for the electrified area and vice versa. For example, the best generator for the electrified area with D_{abs} (D_{real}) of 7 % (−2 %) highly overestimates the FED in most cases with D_{abs} and D_{real} of 75 % and 72 %, respectively. On the other hand, the best generator for the FED sum with D_{abs} (D_{real}) of 22 % (2 %) always underestimates the electrified area with D_{abs} (D_{real}) of 27 % (−27 %). Figure 9 illustrates this finding on the example of test day 31 July 2020.

The developed GEO lightning pseudo-observation generator provides exactly one pseudo-GEO flash for each LF flash. It does not distinguish whether an LF flash, i.e., an NLDN flash, is detected by the GEO LLS, i.e., GLM. During the application of the generator, there is no information whether a given NLDN flash could be detected by the GEO LLS. Additional assumptions, e.g., using flash characteristics, would then be needed to distinguish the LF flashes with and without GEO match. In addition, our GEO lightning data generator

does not include a specific part to simulate GEO flashes that are not directly coincident to any LF flash. Here, the pragmatic approach of using all LF flashes as input is justified with similar flash DE of the LF (i.e., NLDN) and the GEO (i.e., GLM) LLS thus giving overall similar amounts of GLM and NLDN flashes. Then, NLDN and GLM flashes without any coincident observation are analyzed. They are referred to as NLDN-only and GLM-only flashes, respectively. It was observed that both the NLDN-only and GLM-only flashes occurred mostly in proximity to the convective cores and regions of overall high flash rates. The number of observed GLM-only and NLDN-only flashes was in general of the same order of magnitude. It is assumed that pseudo-GLM flashes simulated from the NLDN-only flashes substitute the observed GLM-only flashes. It should be mentioned that some simulated pseudo-GLM flashes might overlap as the pseudo-GLM flash extent is usually greater than the NLDN flash extent. Overlapping pseudo-GLM should actually be merged, however, this is not further studied here. As one possible consequence, the simulated pseudo-GLM FED can be somewhat higher than the observed GLM FED (as seen for most configurations of generators). In particular, the simulated hourly FED values are often higher than observed in situations when many NLDN flashes were observed. On the other hand, lower simulated than observed FED at the rim of cells indicate that NLDN flashes cannot represent the scattering of light as seen by GLM. Peterson et al. (2020) showed that optically detected flashes can appear large near storm edges due to light reflected off nearby clouds. Simulated FED (based on NLDN observations) could then be closer to the actual flash channel extent as derived from LMA-type observations than the observed FED, especially at the rim of cells. Nevertheless, the simulation might differ from what the satellite sensor sees.

Our methodology is configured and refined for NLDN Vaisala sensors. NLDN flash statistics were compared to coincident GLM flashes and their extent, duration, and event number.

For an application in other regions than the US and/or with different LF networks, NLDN operational specification and observations might be compared with the ones of the other LF network in order to identify the necessity for adapting the input data. This comparison can be of direct (e.g., NLDN and GLD360) or indirect (e.g., NLDN and Meteorage compared to ISS-LIS as common reference) nature.

The studied dataset is limited to a region in the SE USA and for the months of March to September. GLM features high flash DE (e.g., Marchand et al. 2019; Murphy and Said 2020) in this region satisfying our objective to build a high-fidelity generator to simulate GEO lightning data. However, the limited dataset lacks winter storms that may have different characteristics. For the application of our generator in Europe, this should be a minor limitation as winter storms rarely occur here. Taszarek et al. (2020) found that 3.6 % of flashes over Europe occurred during the European winter. Wintertime flashes might be important over SE Europe and the Mediterranean Sea. The performance of the data generator will depend on the LF network performance, e.g., flash DE. Realistic data can only be expected in regions where the LF network provides good coverage. The simulated data are, thus, restricted by the quality and range of the LF data input. The SE US region features mostly normal polarity storms while storms with different charge structure occur more often in other parts of the US. For example, Rutledge et al. (2020) show that flash characteristics and GLM flash DE are altered for storms with anomalous charge structure. In addition, the data used to train our GEO lightning data generator were recorded in this region well covered and far from the edges of the GLM's (on GOES-16) field of view. Simulating data of a GEO LLS near the edges of the field of view needs caution regarding parallax effects and an increase in the area one event covers.

The GLM data includes a parallax correction. Our GEO lightning pseudo-observation generator assumes that GLM observations are correctly located. The simulated flashes are placed according to the LF lightning data. If the GEO LLS that should be mimicked uses a different parallax correction than GLM, an adaption may become necessary to obtain realistic data of this LLS.

A comparison of GLM and NLDN during day and night, and for intra-cloud (IC) and cloud-to-ground (CG) flashes revealed similar relationships between NLDN and GLM flash characteristics. The dataset for the ML includes all observed flashes, without a separation of these flash types. In addition, all applied ML models aim to optimize average characteristics. This study uses deterministic approaches without a definition of a confidence interval of the outcomes. As one result, the tails of the characteristics' distributions, e.g., exceptionally small flashes, are underrepresented in the simulation compared to the observation.

Supplementary data might improve the present GEO lightning data generator. Cloud information and brightness temperature data could provide additional features for the ML, e.g., cloud top height, and also information about more likely scattering directions, e.g, in anvils of convective clouds or in stratiform cloud lightning. (Doppler-)Radar data would provide even more versatile possibilities to include cloud structures, dynamics and microphysics.

Data availability statement. NLDN data are available from Vaisala, and data as presented in this paper were provided by Ronald L. Holle.

GLM data are in general available from NOAA: GOES-R Algorithm Working Group and GOES-R Series Program, (2018): NOAA GOES-R Series Geostationary Lightning Mapper (GLM) Level 2 Lightning Detection: Events, Groups, and Flashes. Flash and Event subsets

used. NOAA National Centers for Environmental Information. doi:10.7289/V5KH0KK6. Access date: 01 November 2018. The GLM data as presented in this paper were downloaded from the French AERIS/ICARE Data and Services Center of the Université de Lille where the files were stored 19 April 2018.

Acknowledgments. This work is part of the Ph.D. thesis of Felix Erdmann funded by the Centre National d'Études Spatiales (CNES) and Météo-France. This article is funded by Météo-France, the SOLID project (Funding ID: n/a), and the EXAEDRE project (Funding ID: ANR-16-CE04-0005). This work was supported by the French National program LEFE (Les Enveloppes Fluides et l'Environnement), project ASMA. We thank the AERIS/ICARE Data and Services Center for providing access to GLM data.

The authors thank Ronald L. Holle (Vaisala) for providing the NLDN data and reviewing the NLDN specific information. We would like to acknowledge the guidance of Chien Wang (MOPGA Recipient Scientist at Laboratoire d'Aérologie, Toulouse, France) regarding machine learning applications.

The authors would like to thank the 3 anonymous reviewers and Eric Bruning for their detailed and constructive critics.

The authors declare that there is no conflict of interests.

APPENDIX A

GEO pixel slicing for FED

Deriving FED requires knowledge about flash locations or, in case of satellite observations, the positions of lightning data pixels. GLM products do not come with this necessary information. Therefore, the real GLM grid is reconstructed locating the centers of all events

of the full half-year dataset. This large dataset was used to ensure that the reconstruction of the GLM grid would be complete, i.e., there was at least one event at each GLM pixel. A time invariant real GLM grid is assumed. As individual pixels appear to wobble locally with time and do not appear on a regular grid; due to micro-vibrations of the satellite platform, spacecraft jitter, and variable pitch CCD; a k-means-clustering analysis is performed to identify the statistical mean location of each pixel center. Corner points of pixels are then defined as the mean locations between the centers of the 4 pixels adjacent to each point. It is assumed that corner points can be connected by straight lines in order to represent the pixel shapes. This assumption is not entirely true, as the regular CCD grid is projected on the Earth (more precisely, on the cloud top ellipsoid, Section 2.a), however, the FED should be less impacted by this assumption than by assuming a regular GLM grid. Shapes of GLM events do usually not match the FED grid pixel shapes. One GLM event with average side length of 8.7 km can overlap multiple FED pixels with side length of 5 km to some part. The fractions of the GLM event within each pixel of the FED grid are summed up while integrating over the time period. This slicing of GLM events reduces the effect of producing gaps or double counts of GLM pixel when transformed to the regular FED grid, as recently described by Bruning et al. (2019).

APPENDIX B

Definitions of the machine learning (ML) algorithms

a. ML model types

This section defines the seven ML model types that are trained in the study. The basic idea of each ML model type is introduced, and specifications and important parameters

for their tuning are briefly described. As mentioned in Section 2, Python's sklearn-package is used. Model names are given as they appear in the sklearn library and documentation (`https://scikit-learn.org/stable/`) that provides further details.

(i) Multivariate Linear Regression. The first approach is the most commonly used linear regression *sklearn.linear_model.LinearRegression*. It is applied simultaneously to all features and targets and is, thus, a multivariate linear regression (*LinReg*). The algorithm seeks for the minimum sum of squared errors between the features and the targets by using linear functions. It is an ordinary least square fit in a space with dimensions equal to the number of features times the number of targets.

(ii) Multivariate Polynomial Regression. The Polynomial Regression (*Poly*) is an adjustment to the multivariate linear regression. It fits a polynomial of degree 3 (rather than a linear function) to minimize the sum of squares between predicted targets and the corresponding observations in the validation dataset. The cubic polynomial model is chosen based on the initial correlation analysis with relations between any one feature and one target. The low polynomial degree allows fast computation.

(iii) Random Forest Regressor. A Random Forest (*RF*) is a ML algorithm using bootstrapping and applying single decision trees to each bootstrap sample. The overall result is the average of the outcomes of all the decision trees. The minimum leaf size defines the minimum size at the end of the decision tree. A specific form of the *RF* is called Extra Trees *sklearn.ensemble.ExtraTreesRegressor* (*ETR*, Geurts et al. 2006). *ETR* enforces randomness by not only selected random features in each subset but also splits depending on the best randomly produced thresholds instead of looking for the most distinctive threshold (as in *RF*). *ETR* usually reduces the variance and increases the bias of the model compared to

RF. In general, a higher number of trees improves the performance but also the computation time. Our *RF* implementation uses a *ETR* model with 50 decision trees. The number of decision trees results from a sensitivity test (*ETRs* with 5, 10, 20, 50, 100, and 500 trees were tested) between performance as R^2-score (see skilearn documentation) and computational effort. Here, the GB dataset with independent ML training and validation (i.e., calculating the R^2-scores) data (see Section 1) is used. The minimum leaf size is set to two, i.e., a remaining sample of two data points defines the end of the branch. Single point leaf size would increase the variability of the trees and would lead to a higher likelihood of overfitting.

(iv) Bagging Regressor with k-Nearest Neighbor Regressor. Bootstrap Aggregation, short Bagging (Breiman 1996), uses subsamples drawn by bootstrapping from the entire dataset. This step is similar to the *RF* regressor. The algorithm used to treat the subsamples can, however, be chosen (not always a decision tree). This paper applies the bagging regressor *sklearn.ensemble.BaggingRegressor* combined with a K-Nearest Neighbor (*KNN*) regressor (e.g., Altman 1992) *sklearn.neighbors.KNeighborsRegressor* on each of 50 subsamples. The number of neighbors to use by default is set to the 5 closest points and distance weighting is applied for Euclidean distances. The *KNN* finds closest neighbors with a K-dimensional tree (KD tree) method (Bentley 1975). It reduces the number of distance calculations compared to a brute-force approach calculating distances between all data points. The *KNN* regressor in combination with distance weighting should represent the actual range of the subsample training data better than a decision tree (as used in *RF* and *ETR*). The expense might be an increase in overfitting of the data.

(v) Multilayer Perceptron Neural Network. MultiLayer Perceptrons (*MLPs*) are a form of Neural Networks in supervised ML (Glorot and Bengio 2010). They consist of differ-

ent layers of neurons, where the input layer neurons represent the features and the output layer neurons represent the simulated targets. An adjustable number of hidden layers can connect the input and output layers. Each neuron initially transforms the values from the previous layer in a weighted linear summation. Then, a (non-)linear activation function is used. Parameters of our *MLP* model *sklearn.neural_network.MLPRegressor* were determined after testing different configurations to balance computation time and accuracy. It uses one hidden layer with 50 neurons. The activation function is the rectified linear unit function. Additionally, an early stopping criterion is applied if there is no improvement over 20 consecutive iterations. The early stopping requires splitting the training dataset randomly, whereby 10 % are used to verify the improvement of the model and 90 % remain as actual training dataset. The tolerance for the stopping criteria is reduced from default 10^{-4} to 10^{-8} to allow a higher number of iterations. The alpha parameter for the L2 penalty was also reduced from default 10^{-4} to 10^{-8} after testing different values. The lower alpha led to faster training while maintaining the model skill. This change is indicated by naming *alpha8* of the MLP-based generators. Furthermore, the default Adam solver (Kingma and Ba 2014) and a constant learning rate are used, along with adjusted parameters beta1 (0.7), beta2 (0.9), and epsilon (10^{-10}) for the decay rates and the numerical stability in the Adam solver.

(vi) Support Vector Regressor. The Support Vector Regressor (*SVR*) is based on Support Vector Machine (*SVM*) algorithms. A set of hyper-planes is constructed. Therefore, a defined kernel function is applied to achieve a separation of data clusters (by the hyper-planes) for the regression. The kernel function can be a linear or non-linear function (i.e., polynomial or Radial Basis Function). Linear *SVR* (*linSVR*) is faster and uses less memory than

SVR with non-linear kernel-functions. Non-linear *SVR* provides usually better separation of different clusters in the data and thus a higher score than linear SVR. The distances of the nearest data points to the hyper-planes (so-called functional margins) are maximized. Points with a larger functional margin lead to less uncertainty for the prediction than data close to the hyper-planes. *SVM* in general analyzes all data while the cost function (L1 loss) depends on a subset of the training data, referred to as support vectors. Support vectors are a set of data points with some distance from the target values that still allow the correct prediction. The systematic reduction of the training data makes this model type fundamentally different from the remaining model types of this study. Further information is also provided by Smola and Schölkopf (2004).

Due to our large sample size (672,794 flashes), only the *linSVR sklearn.svm.LinearSVR* is used in this study in its default configuration. As for the *MLP*, an early stopping criterion is used for a lack of improvement between consecutive iterations.

(vii) Histogram-based Gradient Boosting Regression Tree. Boosting is, besides bagging, another approach to reduce overfitting of ML models. It combines an ensemble of weak learners to one strong learner. The Histogram Gradient Boosting Regression *sklearn.ensemble.HistGradientBoostingRegressor* (*HGBR*) is much faster than regular Gradient Boosting Regressors. Data is first binned into 256 integer-valued bins. The algorithm can then leverage histograms instead of relying on sorted continuous values when building the decision trees. The number of splitting points is reduced and the algorithm becomes time efficient, inspired by LightGBM (Ke et al. 2017). The first step of the *HGBR* averages the target values and calculates residuals (average difference of observation to prediction) with a least-squares loss function. Based on these residuals, a small decision tree is built,

along with a learning rate. The learning rate limits the influence of a single small decision tree in the final ensemble to avoid overfitting. Then, new predictions are computed using the averages and the decision tree for residuals. Based on new predictions, new residuals are calculated and a new decision tree is created. The final model combines several of these decision trees to pull the target averages towards the observations. The used maximum number of iteration is 500 and the early stopping criteria kicks in after 50 iterations without significant improvement of the loss value.

b. Naming convention for the GEO lightning pseudo-observation generator configurations

This section defines the meaning of names given to different configurations of a target generator. The names and abbreviations of the ML model types can be found in Table B1. The given ML model types are used in the first part of the GEO lightning pseudo-observation generator referred to as target generator[3]. Table B2 summarizes the feature usage that is available for each ML model type available for the target generator. The feature set selections indicate whether a single-target or multi-step approach is used. The feature set selection called *NLDN* is the default configuration as described. Generators with extension of only *default*, *plus*, *raw*, or *raw plus* are single-step approaches, i.e., using 3 times the model of class (1) in Figure 3. Multi-step simulations always simulate the GLM flash duration in the first step here. The order of the targets number of events per flash and the GLM flash extent is not fixed. The extension *num* indicates one additional step only for the pseudo-GLM event number per flash using the pseudo-GLM flash duration as pseudo-feature. GEO lightning pseudo-observation generator configurations with extension *num ext* and *num ext(a)* have two additional steps using different pseudo-features as shown in Table B2. The *num ext(a2)*

[3]RF is included in the table for completeness. Only ETR as a special RF model is used in the study.

generators use only the GLM flash duration as pseudo-feature, thus two models of class (2) as of Figure 3.

The attributes define a modification of the feature set selections with binary character. The *plus* attribute indicates that NLDN LF amplitude and CG fraction are added to the list of features. Attribute *raw* means that no feature and target scaling were used. Combinations of the given feature set selections and attributes are possible, e.g., an unscaled model with NLDN mean LF amplitude and CG stroke ratio as additional features that uses the GLM flash duration as pseudo-feature for the event number per flash gets the extension *num(a) raw plus*. The total number of generator configurations equals 196: There are 7 ML model types (Table B1 except *RF*). For each ML model type there are 7 feature set selections resulting from the single and multi-step approaches, and for each combination of ML model and feature set selection again 4 different attribute usages (Table B2), i.e., none, *plus*, *raw*, or *raw plus*. The 196 generator configurations (28 for each ML model type) define the base for the statistical results presented in Section 4.

APPENDIX C

The multi-target multi-step approach

This section describes a multi-target regression that simplifies the idea of the stacked single target (SST) approach (Spyromitros-Xioufis et al. 2016). In this study, there are six NLDN features (as physical input) and three GLM targets (as physical simulated variables) per sample, i.e., per flash. The three GLM targets are denoted T_i, T_j, and T_k. T_i can represent any of the 3 targets. The indexes i, j, k indicate the order of obtaining the final targets. Targets that are used like features are referred to as pseudo-features, i.e., T_j and T_k in Figure 3a. With this dataset, there are in general 4 different ways to simulate the target T_i.

The 4 ST models are shown as the training part in Figure 3a. There are 3 classes of models: Yellow is the model class (1) without pseudo-features, gray indicates model class (2) using 1 pseudo-feature, and the red for model class (3) uses 2 pseudo-features. The model $M_{\rightarrow i}$ constitutes the common ML model, i.e., class (1), with only the NLDN features as input. One (i.e., T_j or T_k) or two (i.e., T_j and T_k) of the three targets can be added to the input as pseudo-features in order to simulate the target T_i. The resulting models $M_{j\rightarrow i}$ (using T_j as pseudo-feature), $M_{k\rightarrow i}$ (using T_k with the features), and $M_{j,k\rightarrow i}$ (using T_j and T_k with the features) may indeed take advantage of correlations between the predicted target and the targets that are used as pseudo-features.

The application case only uses the NLDN features as first input. Therefore, a multi-step approach is required. Figure 3b presents the example application for a 3-step approach that first predicts the pseudo-GLM flash duration, then the pseudo-GLM event number per flash, and finally the pseudo-GLM flash extent. This configuration is denoted *num ext(a)* (see Table B2 for details on the configuration naming). The first step, $M_{\rightarrow i}$, uses the NLDN features and predicts the first pseudo-GLM characteristic $M_{\rightarrow i}\left(NLDN\right)$, i.e., pseudo-GLM flash duration. The second step, $M_{i\rightarrow j}$, uses the NLDN features and the result of the first step, $M_{\rightarrow i}\left(NLDN\right)$, i.e., the pseudo-feature GLM flash duration. This model of class (2) predicts the second pseudo-GLM characteristic $M_{i\rightarrow j}\left[NLDN, M_{\rightarrow i}\left(NLDN\right)\right]$, i.e., the pseudo-GLM event number per flash. Both predicted pseudo-GLM characteristics (i.e., GLM flash duration and event number per flash) can then be used as pseudo-features to predict the third target with the class (3) model $M_{i,j\rightarrow k}$. Hence, the final target prediction $M_{i,j\rightarrow k}\left\langle NLDN, M_{i\rightarrow j}\left[NLDN, M_{\rightarrow i}\left(NLDN\right)\right]\right\rangle$ depends on the NLDN features and both previous predictions for this configuration. In general, a model of class (3) can also use two pseudo-features produced by two models of class (1). Also, 2 models of class (2) could be

used to simulate the remaining 2 targets after the first step. Utilizing 3 times a model of class (1) is equal to the common ML ST approach. Hence, several combinations of models of different classes are possible and have been investigated here.

The ML training for the multi-step approach can be performed in parallel for the models $M_{\rightarrow i}$, $M_{j\rightarrow i}$, and $M_{j,k\rightarrow i}$. The approach can use all ML model types as the training creates independent learners. Our multi-step approach adapts the idea of the SST, but uses GLM observations instead of simulated pseudo-GLM targets during the ML training. A trained generator can be applied even if the observations are not available using the corresponding pseudo-observation in their place. This method assumes similarity between observations and pseudo-observations, however, the pseudo-observations only approximate the real observations. Our approach does not propagate errors in successive steps. However, the training is more efficient than for an SST approach as all generator parts can be trained simultaneously rather than waiting for the pseudo-observations to be created. Computational efficiency was necessary due to the large number of generators tested in this paper and in the perspective of an operational-like application. The results (Section 4) showed that our multi-step approach aids in simulating realistic pseudo-GLM observations and the performance is often better than with using common ST models without pseudo-features.

Although the correlations between the NLDN features and both GLM flash extent and event number per flash are relatively weak, the NLDN features improve the prediction during each step as seen through feature drop tests (not shown). Indeed, all features have a positive effect on the model score. Due to strong correlations between GLM flash duration and NLDN features flash duration and pulse/stroke number, and to reduce the number of ML-based target generators, only the multi-step approaches which predict the GLM flash duration in the first step ($M_{\rightarrow i}$) are considered. There remains only one model of class (2) in Figure 3a

and three ways to simulate a target T_i. The GLM flash duration is also weakly correlated to both GLM flash extent and event number (R of about 0.10 and 0.17), and GLM flash extent and event number per flash are well correlated (R of about 0.74). Thus, the first step always provides the pseudo-GLM flash duration from the NLDN flash characteristics as features. The second step uses the simulated flash duration in addition to the NLDN features to simulate one or both of pseudo-GLM flash extent and event number per flash. The pseudo-feature used in model class (2) (Figure 3) is fixed in this paper to be flash duration leaving only one realization of model class (2) to simulate a second target. To further reduce the number of multi-step configurations, the approaches that simulate the flash extent but not the event number per flash through a multi-step process are not further considered since (i) GLM flash duration shows weaker correlation with GLM flash extent than with the event number per flash and (ii) the ST approach for event number per flash from NLDN features exhibits the lowest skill of the three targets. A potential third step may simulate the last GLM target based on NLDN features and the two remaining simulated pseudo-GLM characteristics as additional pseudo-features. The paper describes generator configurations using only the GLM duration (strongest correlations) or using GLM duration and a second target as additional pseudo-features to simulate the remaining target (GLM flash extent or event number per flash).

Our multi-step approach aims at producing more realistic pseudo-GLM flash extent and event number per flash than using the NLDN features alone. The NLDN features also remain important as the correlations between some targets are weak.

APPENDIX D

Supplementary results for each test day and target distribution statistic

This section contains detailed results for each test day that are presented in the main paper for the combined 3-day period. The second part includes the Figures and analysis of the normalized difference between distribution statistics of observed and simulated GLM flash duration and event per flash. The results are presented in a similar way as for the flash extent statistics in the main paper.

a. Target generator results for each test day

Tables D1, D2, and D3 present the results for 07 April 2018, 26 May 2018, and 31 July 2018, respectively. As explained for Table 2 with results for the 3 days combined, the tables show the observed statistics for each target distribution, the outcomes using the best performing generator, and statistics of data simulated with the *linSVR num ext(a) plus* recommended generator. The most common behavior of the target generators exibihts simulated mean values close to the observation statisitcs of the three target distribution. Median values are usually underestimated by the target generators as seen in Tables D2 and D3. Results for the 07 April 2018 test case differ from the general behavior (Table D1). That day saw exceptionally large flashes with high event numbers per flash that likely occurred within the MCS and the squall line. As a consequence, the ML-based target generators underestimated the means of the observed flash characteristics for that test case, but the medians of simulated and observed targets are similar.

The results for each test day resample the results for the combined 3-day test data (see Table 2) overall. Minimum values are often only slightly overestimated for the three targets, while the simulated maxima cannot reach the observed maxima for none of the targets and on none of the 3 test days (Tables D1 to D3). *linSVR*-based target generator outperform all other generator types on 26 May an on 31 July 2020. On April 07 with extensively

large flashes, different *BAGR KNN dist*-based generators are found as best performers for all three targets (Table D1). The choice of the most suitable generator appears to be situational, i.e., there is no generator that performs better than all other generators in all cases. The recommended *linSVR num ext(a) plus* (bold in Tables D1 to D3) performs on one level with best generator for the event number per flash and flash extent on 26 May an on 31 July 2020. The event number per flash is significantly underestimated by this *linSVR*-based generator on 07 April 2020 for the mentioned reason. The flash extent, as most important target for the FED, is also underestimated on that day, however, the CvM is only about half of the CvM for event number per flash meaning a more realistic simulation of the flash extent distribution than the event number per flash distribution.

b. Normalized statistics for difference between observation and simualation for GLM flash duration and event number per flash

Figures D1 and D2 group the results for each statistic by the seven ML model types. As explained in Section 4.a, each distribution contains the results of 28 generators (see also Appendix B.b, Table B2).

Figure D1 shows the normalized differences and scores of different target generators for the GLM flash duration for the 3 test days combined. The GLM flash duration distribution is equally well simulated by a variety of ML-based target generators as the narrow spread of the medians (green line) indicates. In detail, a *linSVR*-based generator and a *MLP*-based generator can predict the mean well, a *MLP*-based generator and an *ETR*-based generator are best for the maximum, while a *linSVR*-based generator exhibits the lowest differences for the median as well as both KS and CvM scores. In total, a *linSVR*-based

target generator (*linSVR num ext raw*) best approximates the observed distribution of the GLM flash duration in this comparison with an NDA value of 0.30.

For the GLM event number per flash in Figure D2, *linSVR* and *BAGR KNN dist* models make the best target generators. The lowest NDA of 0.45 is obtained for several *linSVR* and *BAGR KNN dist*-based generators, e.g., *linSVR num ext(a) raw plus* and *BAGR KNN dist num ext raw*. For test day 07 April 2018 (a dominant mesoscale system with above-average mean and median GLM event numbers per flash), all generators underestimated the event number per flash. As generators using a *linSVR* usually predict lower values than the other generators, they underestimate the observed statistics even more on 07 April 2018. Nevertheless, for the full test data, there are *linSVR*-based generators that predict the mean event number equally well as the best target generator, i.e., *MLP*-based, as demonstrated by the lower whiskers in Figure D2. *LinSVR*-based generators are again most suitable to predict the event number distribution median.

References

Aguiar, G. J., E. J. Santana, S. M. Mastelini, R. G. Mantovani, and S. Barbon, Jr, 2019: Towards meta-learning for multi-target regression problems. URL https://arxiv.org/abs/1907.11277.

Allen, B. J., E. R. Mansell, D. C. Dowell, and W. Deierling, 2016: Assimilation of pseudo-GLM data using the ensemble Kalman filter. *Monthly Weather Review*, **144**, 3465–3486, doi:10.1175/MWR-D-16-0117.1.

Altman, N. S., 1992: An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, **46 (3)**, 175–185, doi:10.1080/00031305.1992.10475879,

URL https://amstat.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879.

Anderson, T. W., 1962: On the distribution of the two-sample Cramér-von Mises criterion. *The Annals of Mathematical Statistics*, **33 (3)**, 1148–1159, doi:10.1214/aoms/1177704477.

Ávila, E. E., R. E. Bürgesser, N. E. Castellano, A. B. Collier, R. H. Compagnucci, and A. R. Hughes, 2010: Correlations between deep convection and lightning activity on a global scale. *Journal of Atmospheric and Solar-Terrestrial Physics*, **72 (14)**, 1114 – 1121, doi:https://doi.org/10.1016/j.jastp.2010.07.019, URL http://www.sciencedirect.com/science/article/pii/S1364682610002154.

Bateman, M., 2013: A high-fidelity proxy dataset for the geostationary lightning mapper (GLM). *AMS Sixth Conference on the Meteorological Application of Lightning Data*, Austin, TX, USA, 725.

Bateman, M., and D. Mach, 2020: Preliminary detection efficiency and false alarm rate assessment of the Geostationary Lightning Mapper on the GOES-16 satellite. *Journal of Applied Remote Sensing*, **14 (3)**, 1 – 10, doi:10.1117/1.JRS.14.032406, URL https://doi.org/10.1117/1.JRS.14.032406.

Bateman, M., D. Mach, and M. Stock, 2021: Further investigation into detection efficiency and false alarm rate for the geostationary lightning mappers aboard GOES-16 and GOES-17. *Earth and Space Science*, **8 (2)**, e2020EA001 237, doi:https://doi.org/10.1029/2020EA001237, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020EA001237, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020EA001237.

Bentley, J. L., 1975: Multidimensional binary search trees used for associative searching. *Communications of the ACM*, **18 (9)**, 509–517, doi:10.1145/361002.361007, URL https://doi.org/10.1145/361002.361007.

Betz, H. D., K. Schmidt, P. Laroche, P. Blanchet, W. P. Oettinger, E. Defer, Z. Dziewit, and J. Konarski, 2009: LINET—an international lightning detection network in europe. *Atmospheric Research*, **91 (2)**, 564 – 573, doi:10.1016/j.atmosres.2008.06.012, URL http://www.sciencedirect.com/science/article/pii/S0169809508002305, 13th International Conference on Atmospheric Electricity.

Biron, D., L. D. Leonibus, P. Laquale, D. Labate, F. Zauli, and D. Melfi, 2008: Simulation of Meteosat Third Generation-Lightning Imager through tropical rainfall measuring mission: Lightning Imaging Sensor data. *Remote Sensing System Engineering*, P. E. Ardanuy, and J. J. Puschell, Eds., SPIE, International Society for Optics and Photonics, Vol. 7087, 77 – 88, doi:10.1117/12.794764.

Blakeslee, R., and W. Koshak, 2016: LIS on ISS: Expanded global coverage and enhanced applications. *The Earth Observer*, **28**, 4–14.

Blakeslee, R. J., and Coauthors, 2020: Three years of the lightning imaging sensor onboard the international space station: Expanded global coverage and enhanced applications. *Journal of Geophysical Research: Atmospheres*, **125 (16)**, e2020JD032 918, doi:10.1029/2020JD032918, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD032918.

Borchani, H., G. Varando, C. Bielza, and P. Larrañaga, 2015: A survey on multi-output regression. *WIREs Data Mining and Knowledge Discovery*, **5 (5)**, 216–233, doi:10.1002/widm.1157, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1157.

Breiman, L., 1996: Bagging predictors. *Machine Learning*, **24(2)**, 123–140, doi:10.1023/A:1018054314350.

Brooks, I. M., C. P. R. Saunders, R. P. Mitzeva, and S. L. Peck, 1997: The effect on thunderstorm charging of the rate of rime accretion by graupel. *Atmospheric Research*, **43 (3)**, 277 – 295, doi:10.1016/S0169-8095(96)00043-9, URL http://www.sciencedirect.com/science/article/pii/S0169809596000439.

Bruning, E. C., and Coauthors, 2019: Meteorological imagery for the geostationary lightning mapper. *Journal of Geophysical Research: Atmospheres*, **124 (24)**, 14 285–14 309, doi:10.1029/2019JD030874.

Cecil, D. J., S. J. Goodman, D. J. Boccippio, E. J. Zipser, and S. W. Nesbitt, 2005: Three years of TRMM precipitation features. Part I: Radar, radiometric, and lightning characteristics. *Monthly Weather Review*, **133**, 543–566, doi:10.1175/MWR-2876.1.

Chmielewski, V. C., and E. C. Bruning, 2016: Lightning mapping array flash detection performance with variable receiver thresholds. *Journal of Geophysical Research: Atmospheres*, **121 (14)**, 8600–8614, doi:10.1002/2016JD025159, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JD025159.

Christian, H. J., and Coauthors, 1999: The lightning imaging sensor. *11th International Conference on Atmospheric Electricity*, 746–749, NASA Conf. Publ. NASA/CP-1999-209261a.

Coquillat, S., and Coauthors, 2019: SAETTA: high-resolution 3-D mapping of the total lightning activity in the Mediterranean Basin over Corsica, with a focus on a mesoscale

convective system event. *Atmospheric Measurement Techniques*, **12 (11)**, 5765–5790, doi: 10.5194/amt-12-5765-2019, URL https://amt.copernicus.org/articles/12/5765/2019/.

Cummins, K. L., and M. J. Murphy, 2009: An overview of lightning locating systems: History, techniques, and uses, with an in-depth look at the U.S. NLDN. *IEEE Trans. Electromag. Compat.*, **51(3)**, 499–518, doi:10.1109/TEMC.2009.2023450.

Deierling, W., and W. A. Petersen, 2008: Total lightning activity as an indicator of updraft characteristics. *Journal of Geophysical Research*, **113**, D16 210, doi:10.1029/2007JD009598.

Dobber, M., and J. Grandell, 2014: Meteosat Third Generation (MTG) Lightning Imager (LI) instrument performance and calibration from user perspective. *Proceedings of the 23rd Conference on Characterization and Radiometric Calibration for Remote Sensing (CALCON)*, 11-14 August 2014, Utah State University, Logan, Utah, USA. 13 pages.

Emersic, C., and C. Saunders, 2020: The influence of supersaturation at low rime accretion rates on thunderstorm electrification from field-independent graupel-ice crystal collisions. *Atmospheric Research*, **242**, 104 962, doi:10.1016/j.atmosres.2020.104962, URL http://www.sciencedirect.com/science/article/pii/S0169809519312803.

Erdmann, F., 2020: Préparation à l'utilisation des observations de l'imageur d'éclairs de météosat troisième génération pour la prévision numérique à courte échéance (Preparation for the use of Meteosat Third Generation Lightning Imager observations in short-term numerical weather prediction). Ph.D. thesis, Université Toulouse 3 – Paul Sabatier, Toulouse, France.

Erdmann, F., E. Defer, O. Caumont, R. J. Blakeslee, S. Pédeboy, and S. Coquillat, 2020: Concurrent satellite and ground-based lightning observations from the optical lightning imaging sensor (ISS-LIS), the low-frequency network meteorage and the SAETTA lightning mapping array (LMA) in the northwestern Mediterranean region. *Atmospheric Measurement Techniques*, **13 (2)**, 853–875, doi:10.5194/amt-13-853-2020, URL https://www.atmos-meas-tech.net/13/853/2020/.

Fierro, A. O., Y. Wang, J. Gao, and E. R. Mansell, 2019: Variational assimilation of radar data and GLM lightning-derived water vapor for the short-term forecasts of high-impact convective events. *Monthly Weather Review*, **147**, 4045–4069, doi:10.1175/MWR-D-18-0421.1.

Geurts, P., D. Ernst, and L. Wehenkel, 2006: Extremely randomized trees. *Machine Learning*, **63**, 3–42, doi:10.1007/s10994-006-6226-1.

Glorot, X., and Y. Bengio, 2010: Understanding the difficulty of training deep feedforward neural networks. *JMLR W&CP: 13th International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, Vol. 9, 249–256.

Goodman, S., D. Mach, W. Koshak, and R. Blakeslee, 2012: GLM lightning cluster-filter algorithm. Algorithm theoretical basis document, NOAA NESDIS Center for Satellite Application and Research.

Goodman, S. J., and Coauthors, 2013: The GOES-R Geostationary Lightning Mapper (GLM). *Atmospheric Research*, **125-126**, 34 – 49, doi:10.1016/j.atmosres.2013.01.006, URL http://www.sciencedirect.com/science/article/pii/S0169809513000434.

Höller, H., and H.-D. Betz, 2010: Study on inter-comparison of LIS and ground-based lightning location system observations. Report ITT No. 09/996, EUMETSAT, Deutsches Zentrum für Luft- und Raumfahrt.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, 2017: LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 3146–3154, URL http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. URL https://arxiv.org/abs/1412.6980.

Kolmasova, I., T. Marshall, S. Bandara, S. Karunarathne, M. Stolzenburg, N. Karunarathne, and R. Siedlecki, 2019: Initial breakdown pulses accompanied by VHF pulses during negative cloud-to-ground lightning flashes. *Geophysical Research Letters*, **46**, 5592–5600, doi:10.1029/2019GL082488.

Koshak, W., D. Mach, M. Bateman, P. Armstrong, and K. Virts, 2010: GOES-16 GLM level 2 data full validation data quality - product performance guide for data users. Guide, NASA Marshall Space Flight Center.

Koshak, W. J., and R. J. Solakiewicz, 2015: A method for retrieving the ground flash fraction and flash type from satellite lightning mapper observations. *Journal of Atmospheric and Oceanic Technology*, **32 (1)**, 79 – 96, doi:10.1175/JTECH-D-14-00085.1, URL https://journals.ametsoc.org/view/journals/atot/32/1/jtech-d-14-00085_1.xml.

Koshak, W. J., and Coauthors, 2004: North Alabama Lightning Mapping Array (LMA): VHF source retrieval algorithm and error analyses. *Journal of Atmospheric and Oceanic Technology*, **21 (4)**, 543 – 558, doi:10.1175/1520-0426(2004)021⟨0543:NALMAL⟩2.0.CO;2, URL https://journals.ametsoc.org/view/journals/atot/21/4/1520-0426_2004_021_0543_nalmal_2_0_co_2.xml.

Luque, M. Y., F. Nollas, R. G. Pereyra, R. E. Bürgesser, and E. E. Ávila, 2020: Charge separation in collisions between ice crystals and a spherical simulated graupel of centimeter size. *Journal of Geophysical Research: Atmospheres*, **125 (3)**, e2019JD030 941, doi:10.1029/2019JD030941, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD030941.

Lyu, F., S. A. Cummer, Z. Qin, and M. Chen, 2019: Lightning initiation processes imaged with very high frequency broadband interferometry. *Journal of Geophysical Research: Atmospheres*, **124**, 2994–3004, doi:10.1029/2018JD029817.

MacGorman, D. R., and W. D. Rust, 1998: *The electrical nature of storms*. 1st ed., Oxford University Press, 198 Madison Avenue, New York, New York 10016.

Mach, D. M., 2020: Geostationary lightning mapper clustering algorithm stability. *Journal of Geophysical Research: Atmospheres*, **125 (5)**, e2019JD031 900, doi:10.1029/2019JD031900, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD031900.

Marchand, M., K. Hilburn, and S. D. Miller, 2019: Geostationary lightning mapper and earth networks lightning detection over the contiguous United States and dependence on flash characteristics. *Journal of Geophysical Research: Atmospheres*, **124 (21)**, 11 552–

11 567, doi:10.1029/2019JD031039, URL https://agupubs-onlinelibrary-wiley-com-s.docadis.ups-tlse.fr/doi/abs/10.1029/2019JD031039.

Massey, F. J., 1951: The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, **46 (253)**, 68–78, URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500769.

Mastelini, S. M., V. G. T. da Costa, E. J. Santana, F. K. Nakano, R. C. Guido, R. Cerri, and S. Barbon, 2019: Multi-output tree chaining: An interpretative modelling and lightweight multi-target approach. *Journal of Signal Processing Systems*, **91 (2)**, 191–215, doi:10.1007/s11265-018-1376-5, URL https://doi.org/10.1007/s11265-018-1376-5.

Murphy, M. J., and R. K. Said, 2020: Comparisons of lightning rates and properties from the U.S. National Lightning Detection Network (NLDN) and GLD360 with GOES-16 Geostationary Lightning Mapper and Advanced Baseline Imager data. *Journal of Geophysical Research: Atmospheres*, **125 (5)**, e2019JD031 172, doi:10.1029/2019JD031172.

Nag, A., M. J. Murphy, W. Schulz, and K. L. Cummins, 2015: Lightning locating systems: Insights on characteristics and validation techniques. *Earth and Space Science*, **2**, 65–93, doi:10.1002/2014EA000051.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Peterson, M., S. Rudlosky, and D. Zhang, 2020: Changes to the appearance of optical lightning flashes observed from space according to thunderstorm organization and structure. *Journal of Geophysical Research: Atmospheres*, **125 (4)**, e2019JD031 087, doi:10.1029/2019JD031087.

Rison, W., R. J. Thomas, P. R. Krehbiel, T. Hamlin, and J. Harlin, 1999: A GPS-based three-dimensional lightning mapping system: Initial observations in central New Mexico. *Geophysical Research Letters*, **26 (23)**, 3573–3576, doi:10.1029/1999GL010856.

Rutledge, S. A., K. A. Hilburn, A. Clayton, B. Fuchs, and S. D. Miller, 2020: Evaluating geostationary lightning mapper flash rates within intense convective storms. *Journal of Geophysical Research: Atmospheres*, **125 (14)**, e2020JD032 827, doi:10.1029/2020JD032827, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD032827.

Schultz, E. V., C. J. Schultz, L. D. Carey, D. J. Cecil, and M. Bateman, 2016: Automated storm tracking and the lightning jump algorithm using GOES-R geostationary lightning mapper (GLM) proxy data. *Journal of Operational Meteorology*, **4(7)**, 92–107, doi:10.15191/nwajom.2016.0407.

Schulz, W., G. Diendorfer, S. Pedeboy, and D. R. Poelman, 2016: The European lightning location system EUCLID – Part 1: Performance analysis and validation. *Nat. Hazards Earth Syst. Sci.*, **16**, 595–605, doi:10.5194/nhess-16-595-2016.

Smola, A. J., and B. Schölkopf, 2004: A tutorial on support vector regression. *Statistics and Computing*, **14**, 199–222, doi:10.1023/B:STCO.0000035301.49549.88.

Spyromitros-Xioufis, E., G. Tsoumakas, W. Groves, and I. Vlahavas, 2016: Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, **104 (1)**, 55–98, doi:10.1007/s10994-016-5546-z, URL https://doi.org/10.1007/s10994-016-5546-z.

Stano, G. T., 2013: Fusing total lightning data with aviation weather center and storm prediction center operations during the GOES-R visiting scientist program. *AMS Sixth Conference on the Meteorological Applications of Lightning Data*, Austin, TX, USA, 724.

Takahashi, T., S. Sugimoto, T. Kawano, and K. Suzuki, 2017: Riming Electrification in Hokuriku Winter Clouds and Comparison with Laboratory Observations. *Journal of the Atmospheric Sciences*, **74 (2)**, 431–447, doi:10.1175/JAS-D-16-0154.1, URL https://doi.org/10.1175/JAS-D-16-0154.1.

Taszarek, M., J. T. Allen, P. Groenemeijer, R. Edwards, H. E. Brooks, V. Chmielewski, and S.-E. Enno, 2020: Severe convective storms across Europe and the United States. Part I: Climatology of lightning, large hail, severe wind, and tornadoes. *Journal of Climate*, **33 (23)**, 10 239 – 10 261, doi:10.1175/JCLI-D-20-0345.1, URL https://journals.ametsoc.org/view/journals/clim/33/23/jcliD200345.xml.

Thomas, R. J., P. R. Krehbiel, W. Rison, T. Hamlin, D. J. Boccippio, S. J. Goodman, and H. J. Christian, 2000: Comparison of ground-based 3-dimensional lightning mapping observations with satellite-based LIS observations in Oklahoma. *Geophysical Research Letters*, **27 (12)**, 1703–1706, doi:10.1029/1999GL010845.

Thomas, R. J., P. R. Krehbiel, W. Rison, S. J. Hunyady, W. P. Winn, T. Hamlin, and J. Harlin, 2004: Accuracy of the lightning mapping array. *Journal of Geophysical Research*, **109**, D14 207, doi:10.1029/2004JD004549.

Vaisala, 2013: Vaisala thunderstorm advanced total lightning sensor LS7002. URL https://www.vaisala.com/sites/default/files/documents/WEA-LS7002-Datasheet-B211284EN-A-LOW.pdf, accessed on 2018-06-28.

Yang, J., Z. Zhang, C. Wei, F. Lu, and Q. Guo, 2017: Introducing the new generation of chinese geostationary weather satellites, Fengyun-4. *Bulletin of the American Meteorological Society*, **98 (8)**, 1637–1658, doi:10.1175/BAMS-D-16-0065.1.

1503 Zhang, D., and K. L. Cummins, 2020: Time evolution of satellite-based optical proper-
1504 ties in lightning flashes, and its impact on GLM flash detection. *Journal of Geophysi-*
1505 *cal Research: Atmospheres*, **125 (6)**, e2019JD032 024, doi:10.1029/2019JD032024, URL
1506 https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD032024.

1507 Zhu, Y., V. A. Rakov, M. D. Tran, and A. Nag, 2016: A study of national light-
1508 ning detection network responses to natural lightning based on ground truth data ac-
1509 quired at log with emphasis on cloud discharge activity. *Journal of Geophysical Re-*
1510 *search: Atmospheres*, **121 (24)**, 14,651–14,660, doi:10.1002/2016JD025574, URL https:
1511 //agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JD025574.

LIST OF TABLES

1561 TABLE 1. Study dates (year 2018) with the amounts of GLM and NLDN data. The three
1562 rightmost columns indicate whether the data are used for ML-based generator building (GB) or
1563 the generator test (GT) part, the time of most lightning activity in the region (D: local daytime,
1564 N: local nighttime), and the primary forcing (Trigger) for storm development and lightning.

Date	number of GLM events	number of GLM flashes	number of NLDN pulses, strokes	number of NLDN flashes	Usage	Time	Trigger
19 Mar	4,053,599	79,420	315,854	78,351	GB	D+N	cyclone, cold front
29 Mar	2,611,064	35,822	122,772	37,931	GB	D+N	stationary front, MCS
07 Apr	5,854,407	94,447	494,686	113,978	GT	D+N	short wave trough, front
14 Apr	8,610,567	142,587	729,622	169,181	GB	D+N	cold front
26 May	4,364,985	130,632	422,193	120,608	GT	D	thermal convection
03 Jun	6,103,693	204,295	825,601	188,330	GB	D+N	cold front
21 Jul	5,541,425	150,363	943,644	142,023	GB	D+N	squall line, outflow boundary
31 Jul	4,885,532	114,133	391,602	106,142	GT	D	dry line, thermal convection
07 Aug	5,283,358	153,671	472,369	137,963	GB	D	thermal convection
13 Sep	1,015,483	28,301	61,124	21,168	GB	D	Hurricane Florence

1565 TABLE 2. Comparison of distribution statistics for observed GLM data and the best generator
1566 for each target during the full test period. The recommended *linSVR*-based generator is shown in
1567 bold. Details about the target generator names are provided in the Appendix B.b.

Generator	Mean	Median	Minimum	Maximum	RMSE per flash	KS	CvM
GLM flash duration [s]							
Observed	0.43	0.31	0.00	16.44	0.00	0.00	0.0
linSVR num ext raw	0.46	0.30	0.00	9.41	0.77	0.15	656.5
linSVR num ext(a) plus	0.41	0.31	0.02	10.0	0.68	0.32	3770.3
GLM event number per flash [-]							
Observed	49.3	25.0	2	1395	0.0	0.00	0.0
linSVR num ext(a) raw plus	35.3	30.0	2	411	79.8	0.38	6687.6
linSVR num ext(a) plus	34.7	30.0	14	457	80.0	0.40	6989.5
GLM flash extent [km]							
Observed	32.9	27.5	0.0	277.0	0.0	0.00	0.0
linSVR num ext(a2)	30.1	26.3	0.0	157.3	30.9	0.24	3479.2
linSVR num ext(a) plus	29.0	26.2	8.5	305.5	31.5	0.24	2738.4

TABLE 3. Comparison of D_{real} and D_{abs} in percent of observed value for the FED sum during the full test period. The best 20 and the worst 5 of the 196 generators (ranked by D_{abs}) are included. The recommended *linSVR*-based generator is shown in bold. Details about the target generator names are provided in the Appendix B.b.

Generator	D_{abs} [%]	D_{real} [%]
linSVR num ext(a2) raw	22.2	2.3
linSVR num ext(a) raw	22.7	-22.5
linSVR num ext(a)	22.9	3.6
linSVR num ext(a) plus	24.9	9.8
linSVR num ext raw	25.7	11.4
linSVR num ext default	26.5	11.0
linSVR num ext(a2)	27.4	12.6
linSVR num ext(a) raw plus	28.8	15.4
linSVR plus num ext	29.6	17.0
linSVR num(a) raw	31.0	18.5
linSVR num ext(a2) plus	32.3	20.9
linSVR(a) raw	34.4	23.2
linSVR raw	35.1	24.2
linSVR num ext(a2) raw plus	35.6	24.9
linSVR num raw	36.0	25.4
linSVR default	36.1	25.3
linSVR num(a) raw plus	36.2	26.0
linSVR num(a)	36.4	25.6
linSVR(a)	36.7	26.0
linSVR num default	36.9	26.3
MLP alpha8 num ext(a) raw plus	95.0	93.6
ETR num ext(a) raw	95.3	93.8
MLP alpha8 num ext(a2) raw	96.7	95.6
ETR num ext(a)	97.8	96.3
MLP alpha8 num ext(a) raw	107.6	106.7

TABLE 4. Comparison of D_{real} and D_{abs} in percent of observed value for the electrified area during the full test period. The best 20 and the worst 5 of the 196 generators (ranked by D_{abs}) are included. In addition, the recommended *linSVR*-based generator is shown in bold. Details about the target generator names are provided in the Appendix B.b.

Generator	D_{abs} [%]	D_{real} [%]
HGBR num ext(a2) plus	7.4	-2.3
HGBR num ext(a2) raw plus	7.4	-2.4
HGBR num ext(a2) raw	7.4	-3.0
HGBR num ext(a2)	7.5	-2.8
Poly plus num ext	7.5	-2.5
Poly num ext raw plus	7.5	-2.5
BAGR KNN dist raw plus	7.6	-1.4
BAGR KNN dist num raw plus	7.6	-1.5
BAGR KNN dist num(a) raw plus	7.6	-1.4
BAGR KNN dist(a) raw plus	7.6	-1.3
ETR plus num	7.7	-3.7
HGBR num ext default	7.7	-1.4
ETR plus	7.7	-3.7
ETR(a) plus	7.7	-3.5
MLP alpha8 num ext plus	7.7	-1.8
ETR raw plus	7.7	-3.9
ETR num(a) plus	7.7	-3.6
ETR num raw plus	7.7	-3.9
BAGR KNN dist num(a)	7.8	-2.1
MLP alpha8 num ext default	7.8	-4.0
linSVR num ext(a) plus	21.3	-21.3
Poly num ext raw	26.0	10.4
Poly num ext default	26.1	10.6
linSVR num ext(a2) raw	27.4	-27.4
linSVR num ext(a)	28.0	-28.0
linSVR num ext(a) raw	35.4	-35.4

Table B1. ML model types with abbreviation.

ML model type	Abbreviation
Multivariate Linear Regression	*LinReg*
Multivariate Polynomial Regression	*Poly*
Random Forest Regressor[1]	*RF*
Random Forest Extra Trees Regressor	*ETR*
Bagging Regressor with k-Nearest Neighbor Regressor (distance weighting)	*BAGN KNN dist*
Multilayer Perceptron Neural Network	*MLP*
(Linear) Support Vector Regressor	*SVR* (*linSVR*)
Histogram-based Gradient Boosting Regression Tree	*HGBR*

[1]This study uses ETR as a special form of RF.

Table B2. Naming conventions of used target generator configurations. The name extensions as of column 1 are used following the ML model type. The three columns indicate the utilized features during the ML training for each of the three targets GLM flash duration (Flash Duration), number of events per flash (Event Number), and GLM flash extent (Flash Extent). NLDN indicates that NLDN flash duration, the number of pulses/strokes per flash, NLDN flash extent, and the maximum LF amplitude are used as features. The GLM pseudo-features Flash Duration, Flash Extent and/or Event Number can complement the NLDN features for some configurations. Feature set selections define how one target (header) is generated, i.e., ST or multi-step approach. The attributes can or cannot be applied and may replace *default* in the generator name. Combinations of a feature set selection with 0, 1, or 2 attributes are possible.

Name extension	Flash Duration	Event Number	Flash Extent
Feature set selections			
default	NLDN	NLDN	NLDN
(a)	Training with default configuration. Model predictions within the training uncertainty of the default model (Initial training step for applying the multi-step approach)		
num	NLDN	NLDN + Flash Duration + Flash Extent	NLDN
num(a)	NLDN	NLDN + Flash Duration	NLDN
num ext	NLDN	NLDN + Flash Duration + Flash Extent	NLDN + Flash Duration + Event Number
num ext(a)	NLDN	NLDN + Flash Duration	NLDN + Flash Duration + Event Number
num ext(a2)	NLDN	NLDN + Flash Duration	NLDN + Flash Duration
Attributes			
plus	NLDN with mean LF amplitude and CG fraction	NLDN with mean LF amplitude and CG fraction	NLDN with mean LF amplitude and CG fraction
raw	features and target not scaled	features and target not scaled	features and target not scaled

Table D1. Comparison of distribution statistics for observed GLM data and the best generator for each target on 07 April 2018. The recommended *linSVR*-based generator is shown in bold. Details about the target generator names are provided in the Appendix B.b.

Generator	Mean	Median	Minimum	Maximum	RMSE per flash	KS	CvM
GLM flash duration [s]							
Observed	0.62	0.45	0.00	16.44	0.00	0.00	0.0
BAGR KNN dist num ext default	0.57	0.46	0.01	10.56	0.92	0.25	949.2
linSVR num ext(a) plus	0.46	0.34	0.02	10.01	0.93	0.21	810.3
GLM event number per flash [-]							
Observed	73.5	46.0	2	1395	0.0	0.00	0.0
BAGR KNN dist num raw	57.8	48.0	3	467	99.7	0.24	996.6
linSVR num ext(a) plus	37.9	33.0	15	457	103.6	0.32	2066.6
GLM flash extent [km]							
Observed	38.5	34.8	0.0	277.0	0.0	0.00	0.0
BAGR KNN dist num ext(a2) raw plus	41.24	35.1	0.0	166.0	33.7	0.24	582.7
linSVR num ext(a) plus	31.1	26.3	8.7	182.6	34.6	0.27	1068.3

1591 Table D2. Comparison of distribution statistics for observed GLM data and the best generator
1592 for each target on 26 May 2018. The recommended *linSVR*-based generator is shown in bold.
1593 Details about the target generator names are provided in the Appendix B.b.

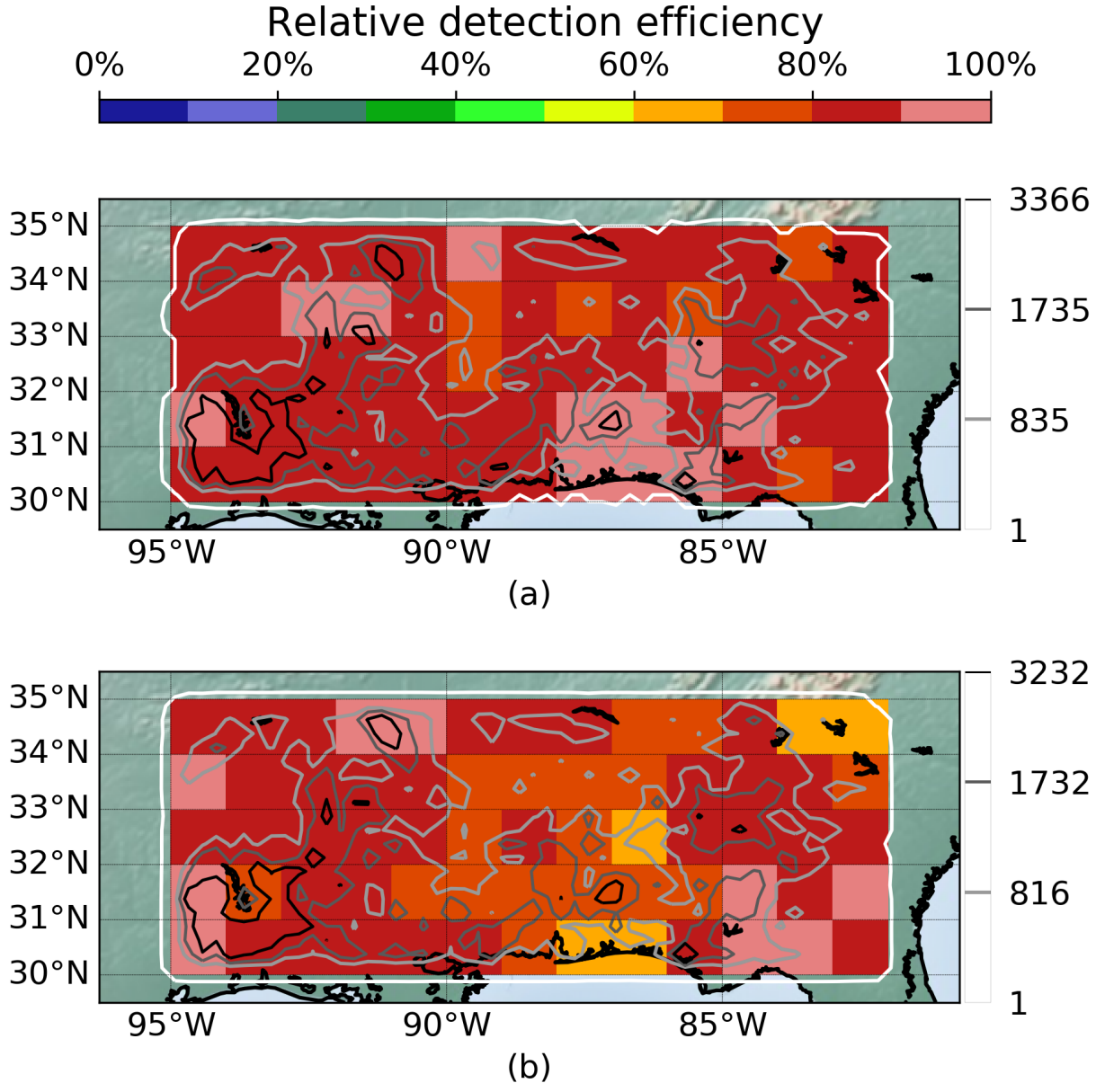
Generator	Mean	Median	Minimum	Maximum	RMSE per flash	KS	CvM
GLM flash duration [s]							
Observed	0.34	0.26	0.00	7.42	0.00	0.00	0.0
linSVR raw	0.41	0.25	0.01	4.73	0.56	0.21	599.9
linSVR num ext(a) plus	0.38	0.31	0.02	5.44	0.46	0.41	2451.3
GLM event number per flash [-]							
Observed	34.8	19.0	2	775	0.0	0.00	0.0
linSVR num ext(a) raw plus	32.2	29.0	3	341	54.0	0.45	3593.1
linSVR num ext(a) plus	31.8	28.0	14	249	54.1	0.48	3670.5
GLM flash extent [km]							
Observed	28.0	20.2	0.0	218.6	0.0	0.00	0.0
linSVR num ext raw	27.4	26.1	0.0	154.8	27.8	0.28	1029.2
linSVR num ext(a) plus	26.7	26.1	8.5	157.6	27.2	0.29	1130.4

1594

Table D3. Comparison of distribution statistics for observed GLM data and the best generator for each target on 31 July 2018. The recommended *linSVR*-based generator is shown in bold. Details about the target generator names are provided in the Appendix B.b.

Generator	Mean	Median	Minimum	Maximum	RMSE per flash	KS	CvM
GLM flash duration [s]							
Observed	0.39	0.29	0.00	9.21	0.00	0.00	0.0
linSVR num ext(a) plus	0.39	0.30	0.02	5.34	0.58	0.32	1307.9
GLM event number per flash [-]							
Observed	45.7	23.0	2	883	0.0	0.00	0.0
linSVR num ext(a) raw plus	33.9	30.0	2	316	76.4	0.40	2441.8
linSVR num ext(a) plus	34.4	30.0	16	315	76.1	0.42	2505.0
GLM flash extent [km]							
Observed	33.8	27.5	0.0	242.2	0.0	0.00	0.0
linSVR raw	29.6	26.2	0.0	180.6	33.5	0.22	745.2
linSVR num ext(a) plus	29.5	26.2	8.5	305.5	32.6	0.24	871.4

LIST OF FIGURES



1670 FIG. 1. Relative flash detection efficiency per $1° \times 1°$ pixel (color) for the full 10-day dataset. In
1671 (a) for GLM and in (b) for NLDN. Grey-scale lines contour the flash number at the 0th (1 flash),
1672 50th, 80th, and 95th percentile of the flash number distribution per $0.25° \times 0.25°$ pixel (only for
1673 pixels with flash activity). The corresponding percentile values are stated on the right-side color
1674 bar.

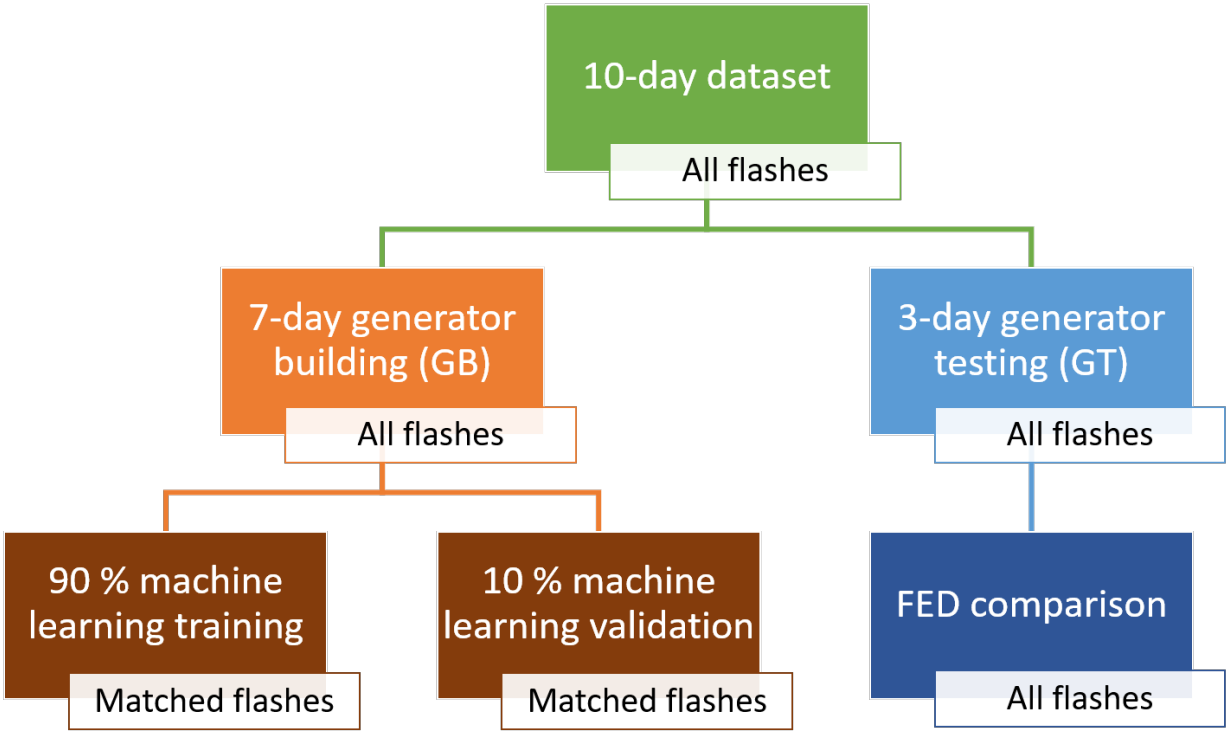
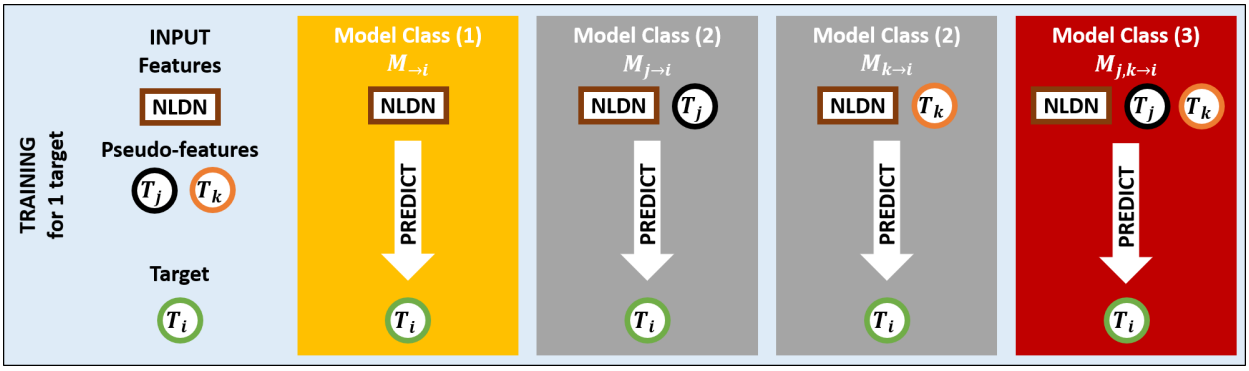
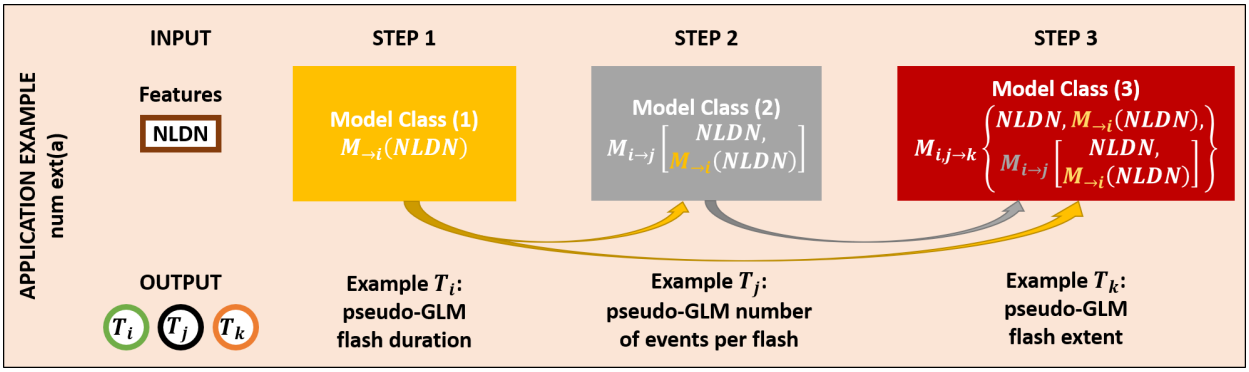


FIG. 2. Illustration of splitting the 10-day dataset with GLM and NLDN flashes in 7-day generator building (GB) and 3-day generator testing (GT) data. The GB data is further processed for the machine learning (ML) part.

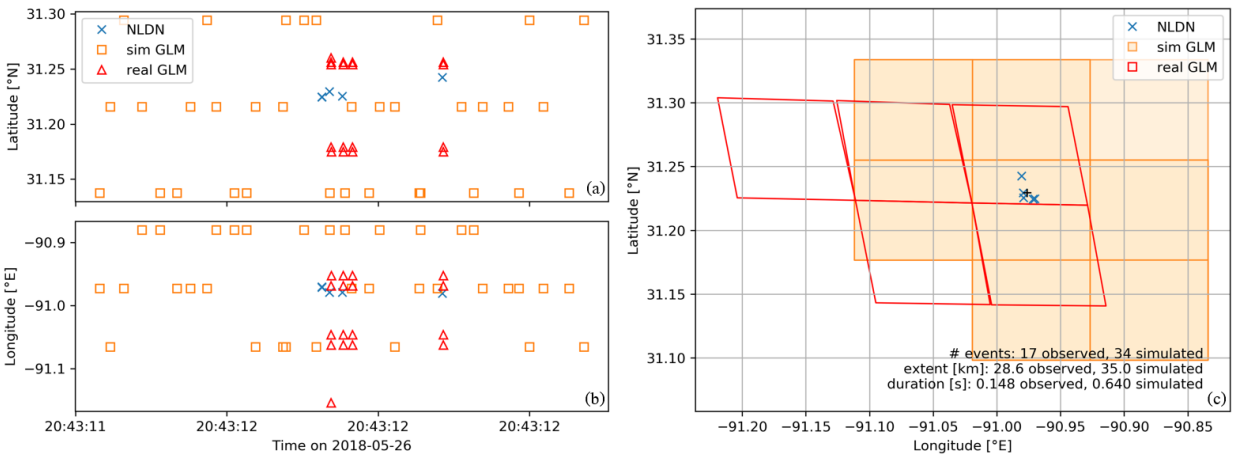


(a)

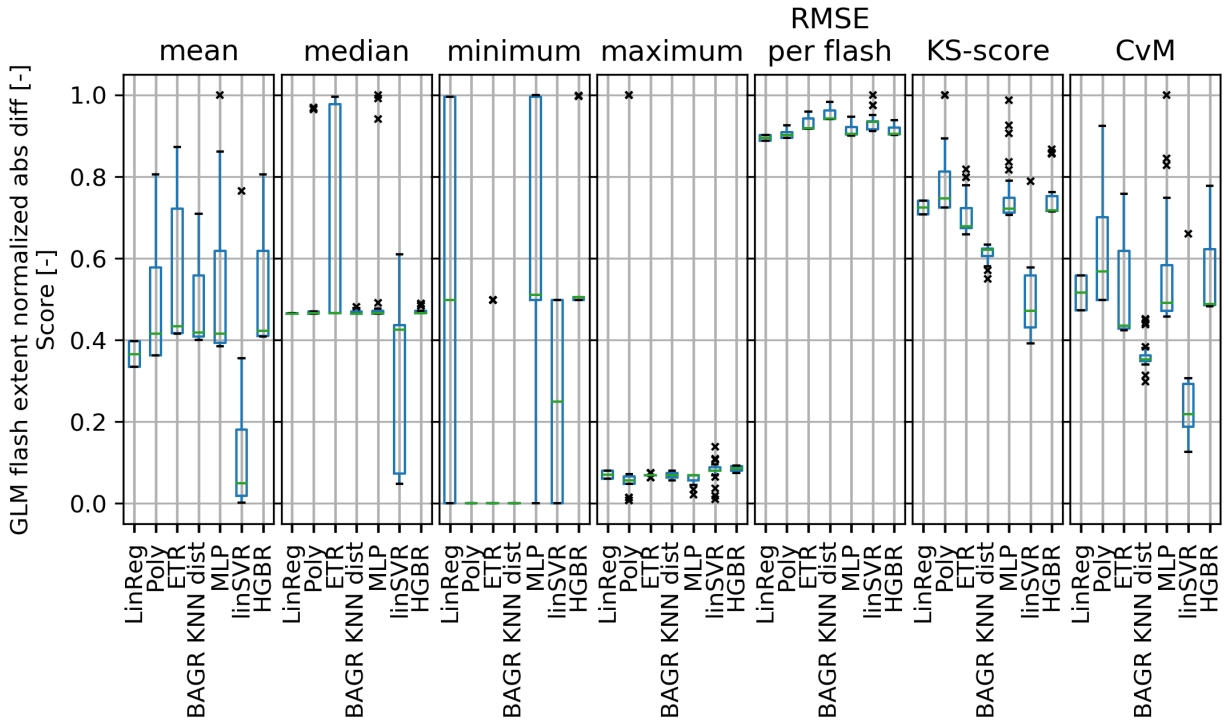


(b)

FIG. 3. Flow chart of the multi-step approach illustrating the possible predictions of a given target using different combinations of features and pseudo-features (a, Training). The Application (b) shows the example of the *num ext(a)* configuration (Appendix B.b).



1681 FIG. 4. An example of one simulated flash with corresponding GLM and NLDN observations on
1682 26 May 2018. The final GEO lightning pseudo-observation generator is used including a linear SVR
1683 model, i.e., *linSVR num ext(a) plus*. Time series of latitudes (a), longitudes (b), and a map (c).
1684 The map (c) includes characteristics of the observed and simulated GLM flash. The time interval
1685 shown matches the simulated flash duration of 640 ms.



1686 FIG. 5. Normalized absolute difference of statistics and scores (titles) between distributions of
1687 observed and simulated GLM flash extent (0 means equal to observation, 1 represents the worst
1688 simulation). The boxplots represent the distributions of 28 target generator results per ML type
1689 (x-axis) including the Inter-Quartile-Range (IQR, blue box), 1.5 times the IQR (whiskers), and
1690 outliers (black cross). The horizontal green line give the median. Results for the full test dataset.
1691 ML type abbreviations provided in Table B1.

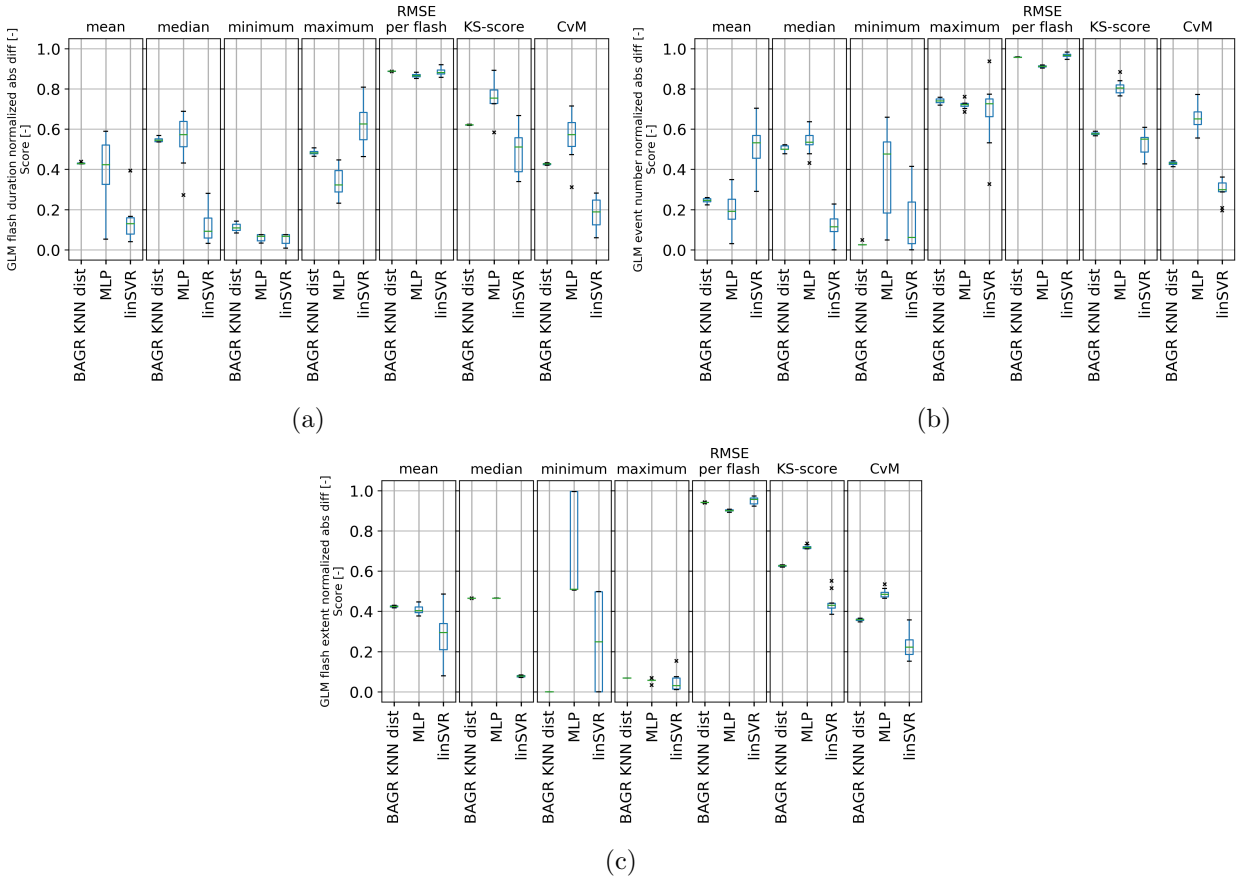
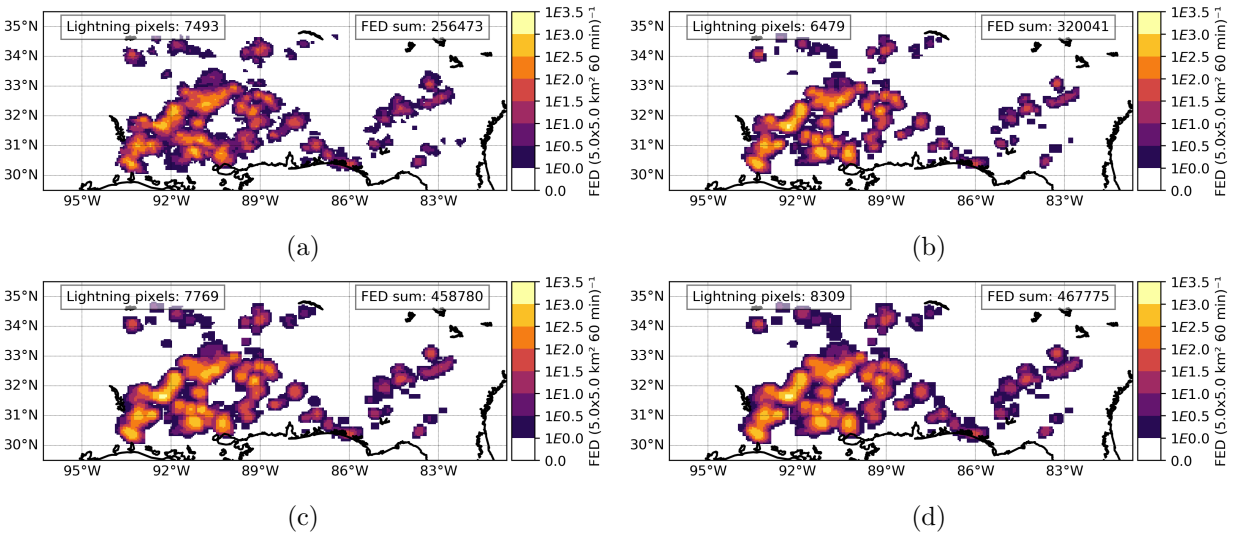


FIG. 6. Normalized absolute difference of statistics and scores (titles) between distributions of observed and simulated GLM flash duration (a), event number per flash (b), and flash extent (c); y-range (0 to 1) as in Figures D1 (for a), D2 (for b), and 5 (for c). Boxplots (as in 5) represent the distribution for training the same model (x-axis) 10 times during the first step of the simulation. ML type abbreviations provided in Table B1.



1697 FIG. 7. Observed (a) and simulated hourly FED using *linSVR num ext(a) plus* (b), *MLP alpha8*
1698 *num raw* (c), and *BAGR KNN dist num plus* (d) generator on 26 May 2018, 20:00 to 21:00 UTC.
1699 The FED grid uses pixels of 5 km × 5 km. ML type abbreviations provided in Table B1.

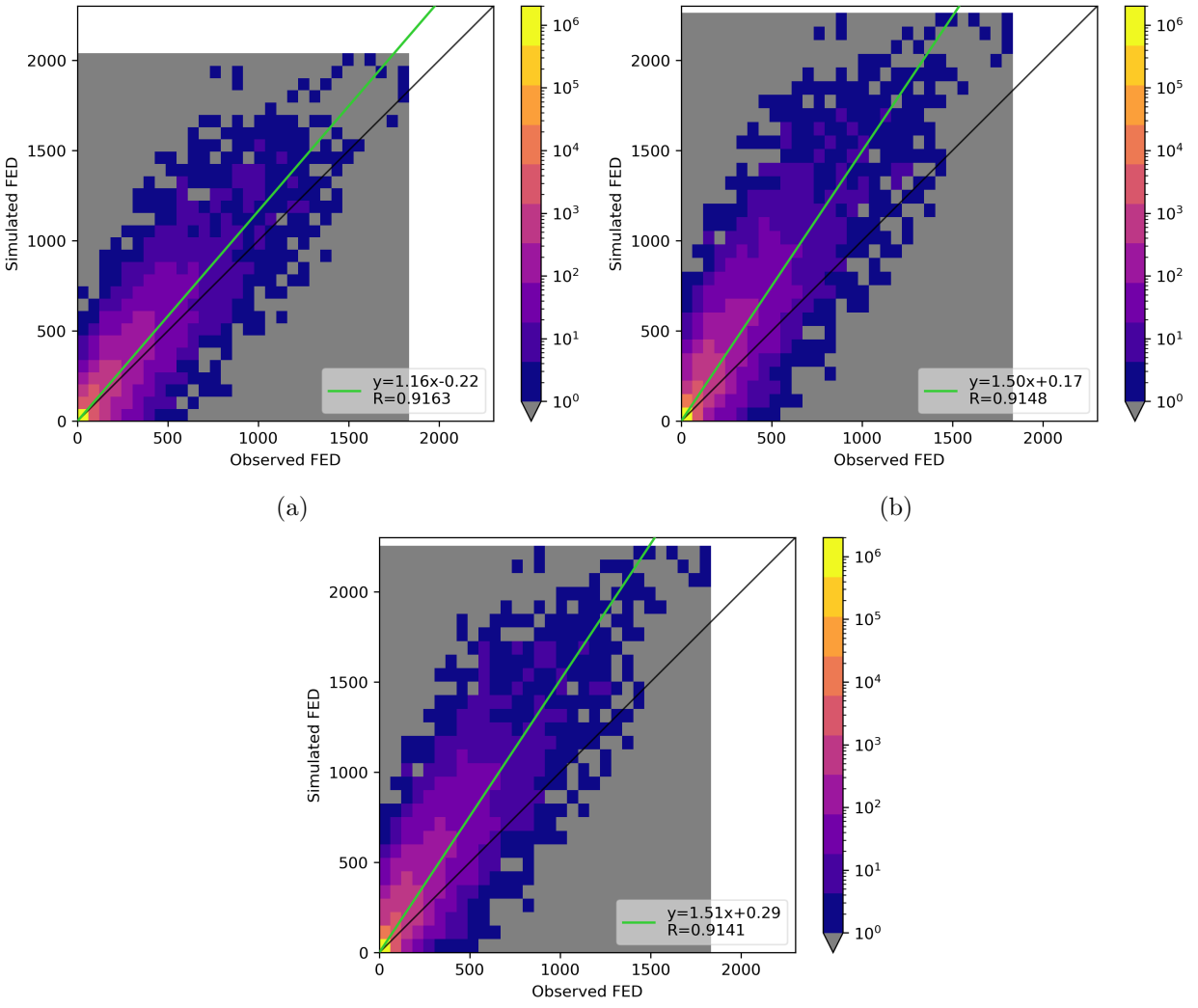
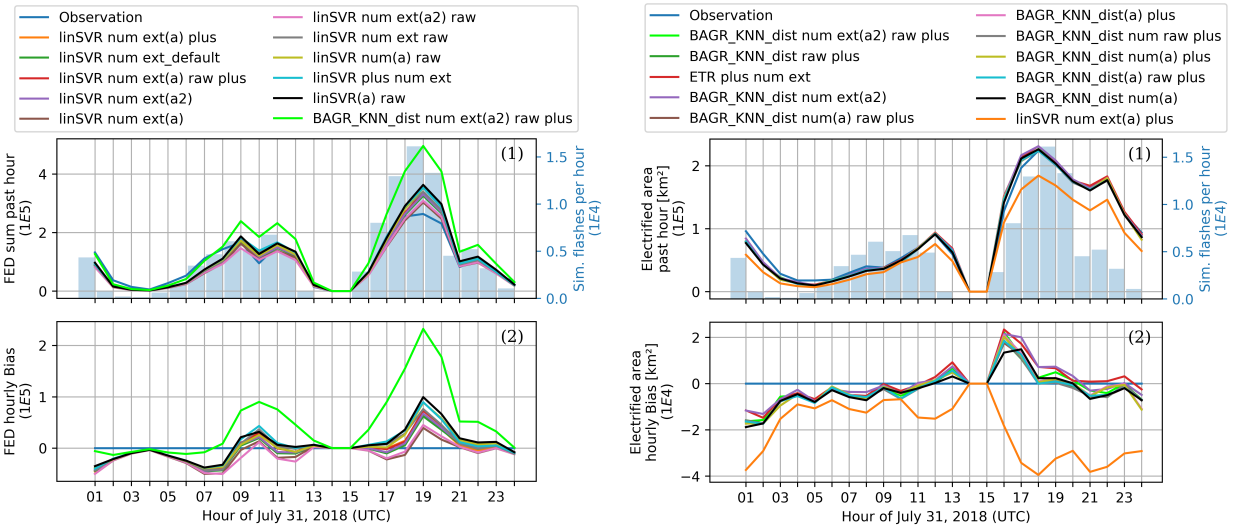
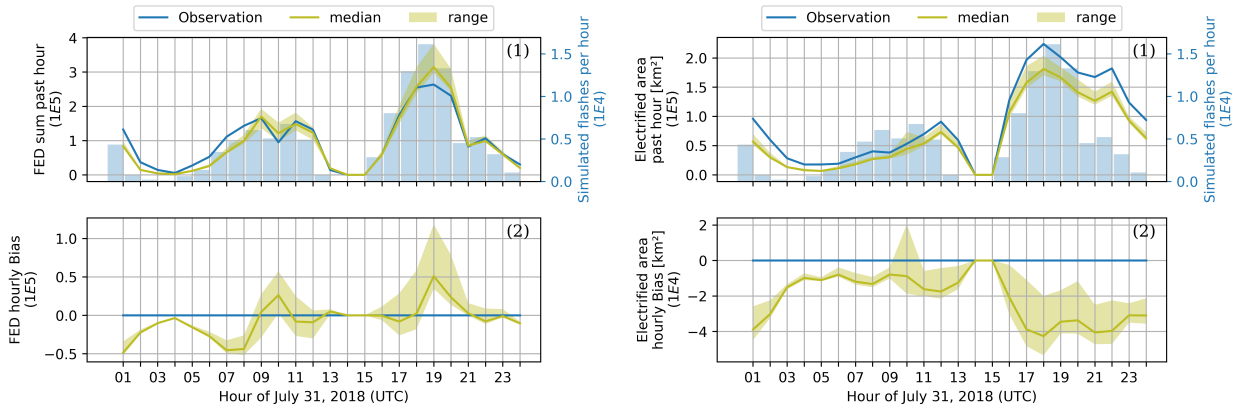


FIG. 8. Pixel-to-pixel (5 km × 5 km) simulated versus observed hourly FED for the 3-day test period using the same *linSVR* (a), *MLP* (b), and *BAGR KNN dist* (c) based generators as in Figure 7. The gray box and white margins indicate the upper limits of distributions on each axis. ML type abbreviations provided in Table B1.



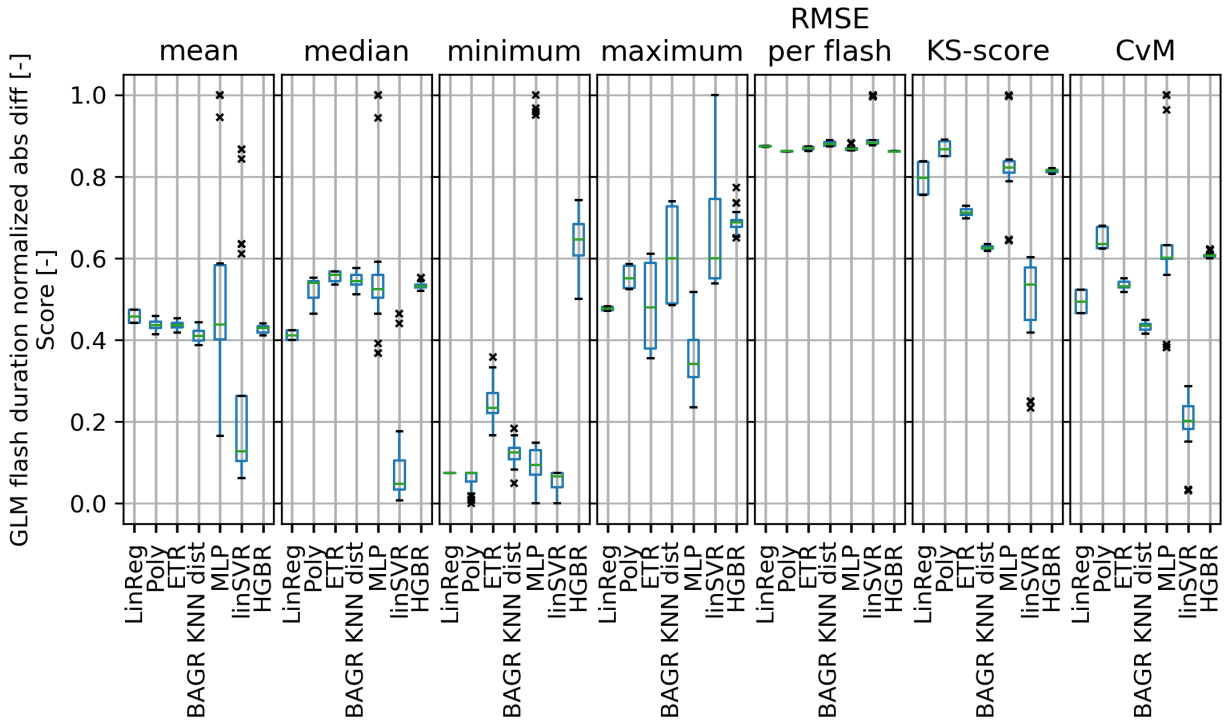
(a) (b)

1704 FIG. 9. Hourly sum of FED (a) and hourly electrified area (b) within the region of interest.
1705 Top (1): Absolute values and number of simulated flashes per hour. Bottom (2): Difference of
1706 simulation minus observation. The observation is plotted in blue, the remaining colors represent
1707 the 10 best generators for FED sum (a) and electrified area (b), respectively. The best generator
1708 of FED sum (a) is also included in (b) [orange], and the best generator from (b) is included in (a)
1709 [lime]. Results for 31 July 2018. Details about the generator names are provided in Appendix B.b.



(a) (b)

1710 FIG. 10. As Figure 9 with 10 repetitions of the recommended *linSVR num ext(a) plus* generator.
1711 Median (line) and range (shaded) of 10 generator repetitions for 31 July 2018.



1712 FIG. D1. Normalized absolute difference of statistics and scores (titles) between distributions of
1713 observed and simulated GLM flash duration (0 means equal to observation, 1 represents the worst
1714 simulation). The boxplots represent the distributions of 28 target generator results per ML type
1715 (x-axis) including the Inter-Quartile-Range (IQR, blue box), 1.5 times the IQR (whiskers), and
1716 outliers (black cross). The horizontal green line give the median. Results for the full test dataset.
1717 ML type abbreviations provided in Table B1.

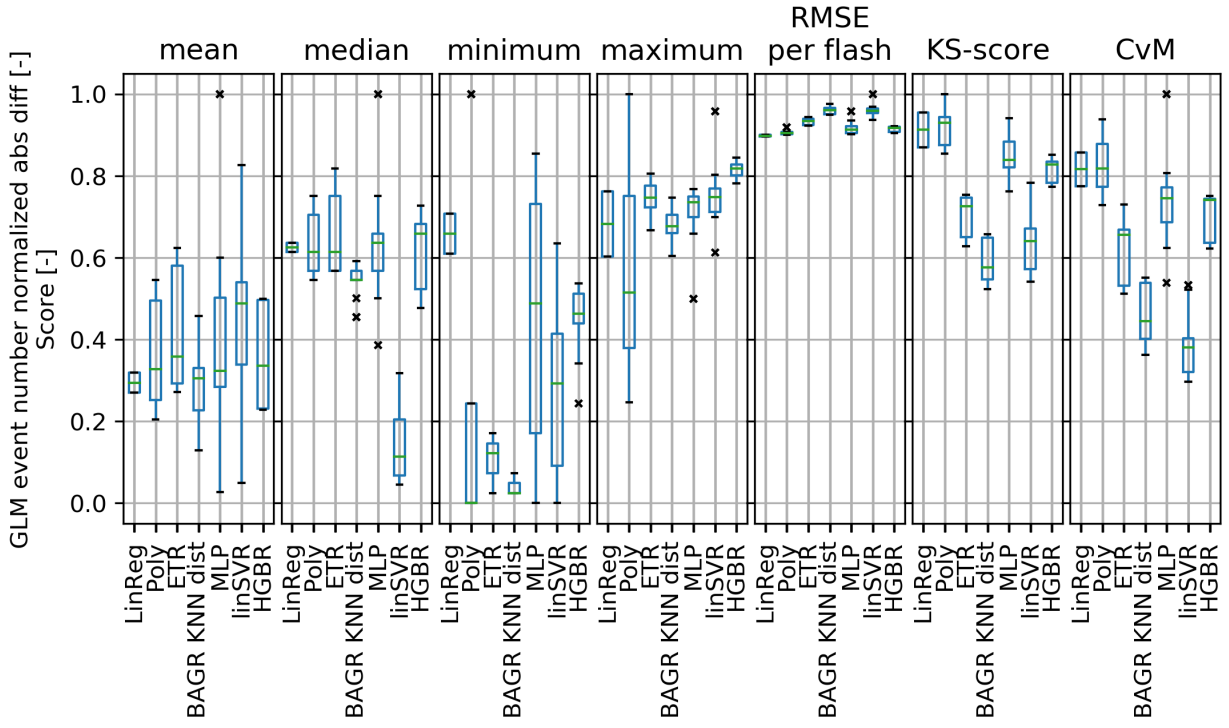


FIG. D2. As Figure D1 but for the normalized absolute difference of statistics and scores (titles) between distributions of observed and simulated GLM event number per flash (0 means equal to observation, 1 represents the worst simulation).