



**HAL**  
open science

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

Felix Erdmann, Olivier Caumont, Eric Defer

► **To cite this version:**

Felix Erdmann, Olivier Caumont, Eric Defer. A geostationary lightning pseudo-observation generator utilizing low frequency ground-based lightning observations. *Journal of Atmospheric and Oceanic Technology*, In press, 39 (1), pp.3-30. 10.1175/JTECH-D-20-0160.1 . hal-03389546

**HAL Id: hal-03389546**

**<https://hal.science/hal-03389546v1>**

Submitted on 21 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3 Felix Erdmann\*

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

6 Olivier Caumont

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

8 Eric Defer

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

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

## ABSTRACT

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

## 30 1. Introduction

31 Lightning is defined as electrical discharges within the atmosphere, more particularly  
32 within and between clouds (intra- and intercloud, IC) or between clouds and the ground  
33 (CG). Transient lightning phenomena also occur between the cloud and the upper atmo-  
34 sphere, e.g., Sprites and Jets. While cloud electrification and lightning initiation are still  
35 subject of studies, it is widely accepted that cloud ice and graupel are necessary to sepa-  
36 rate charges within clouds (e.g., Luque et al. 2020; Emersic and Saunders 2020; Lyu et al.  
37 2019; Kolmasova et al. 2019; Takahashi et al. 2017; MacGorman and Rust 1998; Brooks  
38 et al. 1997). In particular, convection creates favorable conditions for lightning, and the  
39 updraft strength can be well correlated to the total lightning rate (e.g., Deierling and Pe-  
40 tersen 2008). Ávila et al. (2010) found a high correlation between the occurrence of deep  
41 convection and lightning over land at a global scale. Hence, lightning is an effective tracer  
42 of deep convection.

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

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

54 To assimilate new observation types in NWP models it is desired to develop an assimilation  
55 scheme prior to the instrument launch and data availability. The simulation of appropriate  
56 realistic pseudo-observations precedes the development of any assimilation scheme, espe-  
57 cially when the sensor is not yet in operation. Such synthetic observations can be derived  
58 from existing GEO sensors over other regions, i.e., GLM, and ground-based Lightning Lo-  
59 cating Systems (LLSs). In addition, Low Earth Orbit (LEO) missions such as the Lightning  
60 Imaging Sensor (LIS) on the Tropical Rainfall Measurement Mission (TRMM) satellites  
61 (e.g., Christian et al. 1999; Cecil et al. 2005) and on board the International Space Sta-  
62 tion (ISS) (Blakeslee and Koshak 2016; Blakeslee et al. 2020) provide space-based lightning  
63 observations. One can also use ground-based networks, e.g., the National Lightning De-  
64 tection Network (NLDN) (e.g., Cummins and Murphy 2009), Meteorage (e.g., Schulz et al.  
65 2016; Erdmann et al. 2020), and Lightning Mapping Arrays (LMAs) (e.g., Rison et al. 1999;  
66 Thomas et al. 2004; Coquillat et al. 2019). While the satellite sensors detect visible light  
67 of lightning at 777.4 nm, the ground-based networks are operated at different frequencies  
68 that match electromagnetic radiation emitted by different lightning processes. NLDN and  
69 Meteorage use low frequency (LF) sensors that are most sensitive to discharge processes  
70 such as return strokes for CG flashes. Most LF networks can distinguish CG and IC sig-  
71 nals. The CG flash detection (with return strokes) is usually reliable, whereas the IC flash  
72 detection efficiency (DE) increases within the network and for shorter baselines given one  
73 LF sensor type (personal communication, Stéphane Pedeboy, 2020/21). Global LF networks  
74 have lower DE and accuracy than national and regional LF networks (e.g., Nag et al. 2015).  
75 LMA stations sense very high frequency (VHF) signals of lightning leader propagation and

76 allow for 3-dimensional (3D) channel mapping (e.g., Rison et al. 1999). Their drawback is  
77 the limited range. A LMA network provides coverage within a radius of typically a few  
78 hundred kilometers (e.g., Thomas et al. 2004; Koshak et al. 2004; Chmielewski and Bruning  
79 2016; Coquillat et al. 2019).

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

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

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

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

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

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

## 136 **2. Instruments and Data**

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

### 141 *a. Geostationary Lightning Mapper (GLM)*

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



145 distinguish IC from CG signals, Koshak and Solakiewicz (2015) show that some ICs and  
146 CGs can be statistically differentiated. Especially due to the difficulty of the detection of  
147 daytime lightning against bright, sunlit clouds, thresholds and filters are applied to separate  
148 the lightning optical signal from background and other light sources. Lightning is detected  
149 in a narrow (1 nm) band centered at the 777.4 nm oxygen line in the near infrared. The  
150 wide field-of-view (FOV) image is focused on a high speed Charge Coupled Device (CCD)  
151 focal plane with a nearly hemispheric FOV coverage ( $1372 \times 1300$  pixels). The variable pitch  
152 pixel CCD allows for resulting pixels of about 8 km at nadir and only 14 km at the edge of  
153 the FOV (Goodman et al. 2013). Images are produced continuously and in time frames of  
154 2 ms.

155 NASA's GLM lightning data algorithm produces Level 2 data with lightning information  
156 as events, groups and flashes. The x,y-coordinates of the focal plane are transformed to  
157 latitude and longitude coordinates of an estimated cloud top ellipsoid (with a height of  
158 14 km at the equator and 6 km at the poles). Bruning et al. (2019) describes the effects of  
159 using this ellipsoid on GLM parallax with respect to any known ground-relative reference.  
160 GLM events are single illuminated pixels that pass the optical filters and are thus identified  
161 as lightning signals. Their location is defined as the center of the illuminated pixel. Adjacent  
162 events observed in the same 2 ms time frame are merged to form a group. Next, groups are  
163 combined into flashes. NASA's clustering algorithm uses a Weighted Euclidean Distance  
164 (WED) with limits of 16.5 km in latitude and longitude direction and 330 ms in time. Two  
165 groups with a WED of less than one are assigned to the same flash. The WED criterion is  
166 tested for pairs of events with one event in each group (Mach 2020).

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

170 *b. The National Lightning Detection Network (NLDN)*

171 The NLDN (Cummins and Murphy 2009) consists of more than 100 Vaisala LS7002 ground  
172 sensors in the contiguous US (CONUS). It detects LF electromagnetic signals generated by  
173 fast lightning discharges such as return strokes or by intracloud components. Due to a  
174 combination of magnetic direction finding and time-of-arrival techniques, only two sensors  
175 are needed to construct the horizontal location (latitude and longitude, no altitude) and  
176 time of a signal. NLDN locates total lightning, including CG and IC discharges. According  
177 to Vaisala (2013), up to 95 % and better than 50 % of all CG and IC lightning, respectively,  
178 is detected. Zhu et al. (2016) found that one third of 153 IC pulses were detected by NLDN,  
179 and 86 % were classified correctly. NLDN detected 92 % of 367 return strokes, and also 92 %  
180 were correctly classified as CG. The median location accuracy approaches 250 m for CG  
181 strokes in the interior of the network. Lightning can be located at long range (1500 km),  
182 but the location accuracy in the interior of the network is significantly higher than outside.  
183 NLDN measures also the peak amplitude of the LF source. NLDN data used in this study  
184 include time (resolved at 1 ms), the location as latitude and longitude, the peak amplitude  
185 [kA], the polarity, the type (CG or IC) of the LF source, and quality parameters, e.g., the  
186 location error ellipse axes. Although Vaisala merges strokes to flashes (within 10 km and  
187 1 s), this study retrieves NLDN flash level data using the algorithm developed by Erdmann  
188 et al. (2020) for Meteorage records in France. Hence, pulses/strokes are merged into a flash

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

191 *c. Database for the current study*

192 The general dataset consists of six months of GLM and NLDN records, from March 15,  
193 2018 to September 15, 2018. NLDN data were provided in a region between 30°N and 35°N  
194 and 95°W and 82°W. GLM data before 26 September 2018 need a time-of-flight (TOF)  
195 correction that takes into account the time lightning photons need to travel from the cloud  
196 tops (approximated at 10 km of altitude) to the GLM orbit. Our study applies a dynamical  
197 TOF correction with values ranging from 122.8 ms to 124.9 ms in the region of interest.

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

211 and 1,115,585 NLDN flashes due to possible<sup>1</sup> short periods of instrument inactivity. Hence,  
212 the difference in the number of observed flashes is less than 2% of the flash counts, and both  
213 instruments operated continuously during the selected days. As the effect of downtimes of  
214 an instrument can be disregarded, the following analysis uses all available data. Three  
215 among the ten days are chosen to test the generators with uncorrelated data and to assess  
216 the variability in the results (test days). The test days (07 April 2018, 26 May 2018, and  
217 31 July 2018) feature both thermally driven convection and dynamic forcing at air mass  
218 boundaries. In the following, the different weather regimes with different lightning activity  
219 are briefly described for the test days as the final FED product is in fact only analyzed for  
220 these three days.

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

226 26 May 2018 was characterized by relatively warm surface temperatures with slightly  
227 decreasing temperatures from west to east within the region. Moisture was induced into  
228 the region by a weak tropical depression over Cuba and later southern Florida. Convection  
229 occurred mainly in the local afternoon as a result of surface heating. Well defined cells  
230 formed and propagated slowly southward in the cyclonic flow.

---

<sup>1</sup>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.

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

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

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

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

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

### 276 **3. Methods**

277 This section defines the concepts and the strategy to generate GEO pseudo lightning  
278 observations. The methods are designed to use NLDN data and evaluated using real GLM  
279 observations. MTG-LI will provide total lightning observations with similar data structure  
280 as GLM observations. It will also consist of events, groups, and flashes. Although MTG-  
281 LI's spatial resolution (4.5 km at nadir versus 8.0 km at nadir) and the temporal resolution  
282 (1 ms versus 2 ms) will be higher than those of GLM, the methods presented here can still  
283 be applied to simulate MTG-LI observations. A comparison of ISS-LIS records over the  
284 domain of this study (USA) and the target region (France) revealed statistically similar  
285 flash characteristics (Erdmann 2020, thesis, Chapter II.1-2). In addition, for both regions  
286 of interest, statistics on NLDN and Meteorage LF lightning observations relative to ISS-LIS  
287 records were consistent. The FED as explained in the following section is simulated on a  
288  $5 \text{ km} \times 5 \text{ km}$  resolution grid approximating the MTG-LI grid.

#### 289 *a. Definition of the Flash Extent Density (FED)*

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

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

300 The FED in this study is calculated on a regular latitude longitude grid with an average  
301 pixel size of  $5 \text{ km} \times 5 \text{ km}$ . To obtain the regular latitude longitude grid, the distance of  
302  $5 \text{ km}$  is transformed to latitudinal and longitudinal distance as of the pixel at the center  
303 of the study region. Appendix A describes the details to transform GEO pixel grid to the  
304 regular FED grid. In the present study, FEDs are analyzed per 60 min time integration  
305 periods. The 1-hour period maintains information to locate tracks of convective cores and  
306 most electrified regions while it is also long enough to capture several storms distributed  
307 within the full domain. There might be, however, multiple storms at one location during  
308 60 min. The FED integration period can be changed as needed since our GEO lightning  
309 pseudo-observation generator simulates data at the flash level. The sum of multiple short  
310 FED periods is equal to the FED of a corresponding long period, but the computation of one  
311 long period is more efficient. Hence, this work simulates FED per hour for computational  
312 reasons. It should be mentioned, however, that other FED time integration periods are  
313 currently under investigation, and the assimilation of MTG-LI will use a shorter FED time  
314 integration period.

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

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



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

## 326 1) SIMULATE PSEUDO-GLM FLASH CHARACTERISTICS

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

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

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

359 Building the GEO lightning pseudo-observation generator requires independent generator  
360 building (GB) and generator testing (GT) data for the generator design and for the veri-  
361 fication of the generated product, i.e., the FED, respectively. The split of our dataset is  
362 illustrated in Figure 2. The GB data consist of seven days and the GT data of the remaining  
363 3 days (test days) of the full dataset (see Section 2.c and Table 1). The GB includes an  
364 ML part. Here, only matched flashes are considered in order to compare feature and target  
365 values (see Figure 2). Features (input data) of the ML are the six NLDN characteristics,  
366 and targets (output data) are GLM flash duration, event number per flash, and flash extent.  
367 Feature and target sample sizes are given as the number of matched flashes detected by

368 GLM and NLDN, respectively, and are not equal in general (Section 2.d). Since training  
369 the ML models requires the same sample size for the features and targets, two (or more)  
370 flashes matched to the same flash of the other LLS are merged, and characteristics of the  
371 merged flashes are combined. The resulting ML data (dark orange in Figure 2) consist of  
372 672,794 flashes, each sample with six NLDN features and three GLM targets. The ML part  
373 further splits this set of ML data randomly into independent ML training and ML validation  
374 data at a ratio of 90 % to 10 %. The ML models are thus trained with 605,515 flashes. The  
375 ML validation data serves to calculate goodness-of-fit scores for each applied ML technique.  
376 Then, the different ML models are compared and the model parameters (e.g., the number  
377 of trees or the number of neural network layers, see Appendix B.a) are tuned based on the  
378 scores. The 3-day GT dataset is used to evaluate each generator as a whole including the ML  
379 and event generation parts. The test exercise exploits both observed GLM and generated,  
380 NLDN-based pseudo-GLM datasets as two independent populations.

381 The generator simulates one pseudo-GLM flash for each observed NLDN flash. Thereby,  
382 it is assumed that flashes detected by GLM only and detected by NLDN only compensate  
383 each other. The assumption was justified as (i) GLM and NLDN feature flash DEs on the  
384 same order, (ii) both GLM-only and NLDN-only flashes were smaller in extent and shorter  
385 in duration than the flashes with coincident observations (see also e.g., Zhang and Cummins  
386 2020; Erdmann et al. 2020), and (iii) most GLM-only and NLDN-only flashes were found in  
387 the same regions in proximity to convective cores where high flash rates were observed (as  
388 in Zhang and Cummins 2020). The overall GLM and NLDN flash numbers (see Table 1)  
389 vary by only a few percent. However, it can be seen that there are days and cases where  
390 NLDN detects more flashes than GLM, i.e., 07 Apr. and 14 Apr. and other days where  
391 GLM detected more flashes than NLDN, i.e., 26 May, 03 Jun., and 07 Aug.

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

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

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

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

## 424 2) ML-BASED TARGET GENERATORS RELATING NLDN FLASH CHARACTERISTICS TO 425 GLM FLASH CHARACTERISTICS

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

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

### 438 3) MULTI-STEP APPROACH

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

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

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

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

#### 466 4) APPLIED SCALING METHODS

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

$$X_N = \frac{X - \min X}{\max X - \min X} \quad (1)$$

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

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

$$Z = \frac{X - \text{mean } X}{\text{std } X} \quad (2)$$

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

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

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

## 480 5) GENERATE PSEUDO-GLM EVENTS

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

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

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



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

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

## 527 4. Results

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

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

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

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

549 2 and 1395 events. The test data features GLM flash extent between 0 km and 277 km. The  
550 target generators should handle these ranges of values and predict target statistics similar  
551 to the statistics of observed GLM flashes.

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

559 This analysis was also conducted for each test day. The results are presented in Ap-  
560 pendix D.a.

561 The majority of the target generators features mean values similar to the observed means  
562 for all three target characteristics. The simulated medians, however, exceed the observed  
563 medians in most cases, especially for the number of events per flash, suggesting a tendency  
564 to overestimate the target values. The previously described behavior is true for all but the  
565 *linSVR*-based generators. *linSVR* filters the dataset in advance to build the prediction on the  
566 support vectors (Appendix B.a.(vi)). That results (in this study) in lower differences between  
567 the simulated and observed median values as compared to using the other ML model types.  
568 The mean values of *linSVR*-based predictions, are, however, often smaller than the observed  
569 mean, especially for the event number per flash. Table 2 demonstrates this behavior of  
570 *linSVR*-based generators. To detail one example, the recommended *linSVR*-based generator  
571 (see Section 4.b, bold in Table 2) underestimates median and mean flash extent by about  
572 4.5% and 11.7%, respectively. The mean event number per flash is also underestimated

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

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

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

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

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

---

<sup>2</sup>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*.

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

627 Overall NDAs for all three targets range from 0.35 for the *linSVR num ext raw* generator to  
628 0.87 for the *MLP num ext(a) raw* generator. The best (i.e., lowest NDA) 24 target generators  
629 all use a *linSVR*, and the performance of the best target generators varies only within the  
630 range of uncertainty given in Section 1. For example, the difference between the 1st and 10th  
631 ranked target generator is only 0.04 NDA. The NDA ranking of target generators reveals  
632 a clustering explained by the ML model type, with *linSVR*-based generators performing  
633 the best, followed by *BAGR KNN dist*-based, *ETR*-based, and polynomial regression-based  
634 generators. MLP- and HGBR-based generators exhibit the highest NDAs.

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

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

## 657 1) CONFIDENCE IN THE RESULTS

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

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

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



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

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

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

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

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

729 The results are further investigated for the 3-day test period by comparing pixel-to-pixel  
730 simulated and observed hourly FED. Figure 8 shows the 2D-histograms, computed for the  
731 entire 3-day test dataset, for the same *linSVR* (a), *MLP* (b), and *BAGR KNN dist* (c) based  
732 generators as used in Figure 7. In general, the Pearson correlation coefficients  $R$  of 0.91 to  
733 0.92 indicate well correlated distributions of observed and simulated FED. Figure 8 also

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

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

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

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

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

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

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

762 As mentioned, the best 28 generators for the FED sum are all of type *linSVR*. The best  
763 10 generators use the multi-step approach (*num* and *num ext*, Table 3). The use of mean  
764 LF amplitude and CG fraction (*plus*) as additional NLDN features has a minor effect on  
765 the simulation of FED sum.

766 Results for the electrified area are in general closer to the observation than the FED sum.  
767 They are shown in Table 4 for the best 20 and worst 5 generators as ranked by  $D_{abs}$  of the  
768 electrified area. The generators with the lowest  $D_{abs}$ , *HGBR* type, differ absolutely by about  
769  $7.5\%$  from the observed electrified area. The vast majority of all tested target generators  
770 underestimate the electrified area (negative  $D_{real}$ ). Multiple generators of various types  
771 feature  $D_{abs}$  of less than  $10\%$ , e.g., using *HGBR*, *Poly*, *BAGR KNN dist*, *ETR*, or *MLP*  
772 models. The *linSVR*-based generators, that performed best for the FED sum, exhibit the  
773 highest differences to the observation here with  $D_{abs}$  from  $15\%$  to  $35\%$  (all with negative  
774  $D_{real}$ ). For example, the best performer for the FED sum is ranked as third worst for the  
775 electrified area with a high underestimation of the area.

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

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

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

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

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

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

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

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

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

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

## 850 5. Summary

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

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

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

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



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

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

885 In general, generators that perform well for the FED sum exhibit high  $D_{abs}$  for the electri-  
886 fied area and vice versa. For example, the best generator for the electrified area with  $D_{abs}$   
887 ( $D_{real}$ ) of 7 % (-2 %) highly overestimates the FED in most cases with  $D_{abs}$  and  $D_{real}$  of  
888 75 % and 72 %, respectively. On the other hand, the best generator for the FED sum with  
889  $D_{abs}$  ( $D_{real}$ ) of 22 % (2 %) always underestimates the electrified area with  $D_{abs}$  ( $D_{real}$ ) of  
890 27 % (-27 %). Figure 9 illustrates this finding on the example of test day 31 July 2020.

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

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

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

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

926 The studied dataset is limited to a region in the SE USA and for the months of March  
927 to September. GLM features high flash DE (e.g., Marchand et al. 2019; Murphy and Said  
928 2020) in this region satisfying our objective to build a high-fidelity generator to simulate GEO  
929 lightning data. However, the limited dataset lacks winter storms that may have different  
930 characteristics. For the application of our generator in Europe, this should be a minor  
931 limitation as winter storms rarely occur here. Taszarek et al. (2020) found that 3.6%  
932 of flashes over Europe occurred during the European winter. Wintertime flashes might  
933 be important over SE Europe and the Mediterranean Sea. The performance of the data  
934 generator will depend on the LF network performance, e.g., flash DE. Realistic data can  
935 only be expected in regions where the LF network provides good coverage. The simulated  
936 data are, thus, restricted by the quality and range of the LF data input. The SE US region  
937 features mostly normal polarity storms while storms with different charge structure occur  
938 more often in other parts of the US. For example, Rutledge et al. (2020) show that flash  
939 characteristics and GLM flash DE are altered for storms with anomalous charge structure.  
940 In addition, the data used to train our GEO lightning data generator were recorded in  
941 this region well covered and far from the edges of the GLM's (on GOES-16) field of view.  
942 Simulating data of a GEO LLS near the edges of the field of view needs caution regarding  
943 parallax effects and an increase in the area one event covers.

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

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

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

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

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

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

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

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

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

983 The authors declare that there is no conflict of interests.

## 984 APPENDIX A

### 985 **GEO pixel slicing for FED**

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

989 of the full half-year dataset. This large dataset was used to ensure that the reconstruction  
990 of the GLM grid would be complete, i.e., there was at least one event at each GLM pixel. A  
991 time invariant real GLM grid is assumed. As individual pixels appear to wobble locally with  
992 time and do not appear on a regular grid; due to micro-vibrations of the satellite platform,  
993 spacecraft jitter, and variable pitch CCD; a k-means-clustering analysis is performed to  
994 identify the statistical mean location of each pixel center. Corner points of pixels are then  
995 defined as the mean locations between the centers of the 4 pixels adjacent to each point. It  
996 is assumed that corner points can be connected by straight lines in order to represent the  
997 pixel shapes. This assumption is not entirely true, as the regular CCD grid is projected on  
998 the Earth (more precisely, on the cloud top ellipsoid, Section 2.a), however, the FED should  
999 be less impacted by this assumption than by assuming a regular GLM grid. Shapes of GLM  
1000 events do usually not match the FED grid pixel shapes. One GLM event with average side  
1001 length of 8.7 km can overlap multiple FED pixels with side length of 5 km to some part.  
1002 The fractions of the GLM event within each pixel of the FED grid are summed up while  
1003 integrating over the time period. This slicing of GLM events reduces the effect of producing  
1004 gaps or double counts of GLM pixel when transformed to the regular FED grid, as recently  
1005 described by Bruning et al. (2019).

## 1006 APPENDIX B

### 1007 **Definitions of the machine learning (ML) algorithms**

#### 1008 *a. ML model types*

1009 This section defines the seven ML model types that are trained in the study. The basic  
1010 idea of each ML model type is introduced, and specifications and important parameters

1011 for their tuning are briefly described. As mentioned in Section 2, Python’s sklearn-package  
1012 is used. Model names are given as they appear in the sklearn library and documentation  
1013 (<https://scikit-learn.org/stable/>) that provides further details.

1014 (i) *Multivariate Linear Regression.* The first approach is the most commonly used linear  
1015 regression `sklearn.linear_model.LinearRegression`. It is applied simultaneously to all features  
1016 and targets and is, thus, a multivariate linear regression (*LinReg*). The algorithm seeks for  
1017 the minimum sum of squared errors between the features and the targets by using linear  
1018 functions. It is an ordinary least square fit in a space with dimensions equal to the number  
1019 of features times the number of targets.

1020 (ii) *Multivariate Polynomial Regression.* The Polynomial Regression (*Poly*) is an adjust-  
1021 ment to the multivariate linear regression. It fits a polynomial of degree 3 (rather than  
1022 a linear function) to minimize the sum of squares between predicted targets and the cor-  
1023 responding observations in the validation dataset. The cubic polynomial model is chosen  
1024 based on the initial correlation analysis with relations between any one feature and one  
1025 target. The low polynomial degree allows fast computation.

1026 (iii) *Random Forest Regressor.* A Random Forest (*RF*) is a ML algorithm using boot-  
1027 strapping and applying single decision trees to each bootstrap sample. The overall result  
1028 is the average of the outcomes of all the decision trees. The minimum leaf size defines the  
1029 minimum size at the end of the decision tree. A specific form of the *RF* is called Extra Trees  
1030 `sklearn.ensemble.ExtraTreesRegressor` (*ETR*, Geurts et al. 2006). *ETR* enforces randomness  
1031 by not only selected random features in each subset but also splits depending on the best  
1032 randomly produced thresholds instead of looking for the most distinctive threshold (as in  
1033 *RF*). *ETR* usually reduces the variance and increases the bias of the model compared to

1034 *RF*. In general, a higher number of trees improves the performance but also the computation  
1035 time. Our *RF* implementation uses a *ETR* model with 50 decision trees. The number of  
1036 decision trees results from a sensitivity test (*ETRs* with 5, 10, 20, 50, 100, and 500 trees were  
1037 tested) between performance as  $R^2$ -score (see sklearn documentation) and computational  
1038 effort. Here, the GB dataset with independent ML training and validation (i.e., calculating  
1039 the  $R^2$ -scores) data (see Section 1) is used. The minimum leaf size is set to two, i.e., a  
1040 remaining sample of two data points defines the end of the branch. Single point leaf size  
1041 would increase the variability of the trees and would lead to a higher likelihood of overfitting.

1042 *(iv) Bagging Regressor with k-Nearest Neighbor Regressor.* Bootstrap Aggregation, short  
1043 Bagging (Breiman 1996), uses subsamples drawn by bootstrapping from the entire dataset.  
1044 This step is similar to the *RF* regressor. The algorithm used to treat the subsamples can,  
1045 however, be chosen (not always a decision tree). This paper applies the bagging regressor  
1046 `sklearn.ensemble.BaggingRegressor` combined with a K-Nearest Neighbor (*KNN*) regressor  
1047 (e.g., Altman 1992) `sklearn.neighbors.KNeighborsRegressor` on each of 50 subsamples. The  
1048 number of neighbors to use by default is set to the 5 closest points and distance weighting is  
1049 applied for Euclidean distances. The *KNN* finds closest neighbors with a K-dimensional tree  
1050 (KD tree) method (Bentley 1975). It reduces the number of distance calculations compared  
1051 to a brute-force approach calculating distances between all data points. The *KNN* regressor  
1052 in combination with distance weighting should represent the actual range of the subsample  
1053 training data better than a decision tree (as used in *RF* and *ETR*). The expense might be  
1054 an increase in overfitting of the data.

1055 *(v) Multilayer Perceptron Neural Network.* MultiLayer Perceptrons (*MLPs*) are a form  
1056 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

1080 *SVR* with non-linear kernel-functions. Non-linear *SVR* provides usually better separation  
1081 of different clusters in the data and thus a higher score than linear *SVR*. The distances of  
1082 the nearest data points to the hyper-planes (so-called functional margins) are maximized.  
1083 Points with a larger functional margin lead to less uncertainty for the prediction than data  
1084 close to the hyper-planes. *SVM* in general analyzes all data while the cost function (L1  
1085 loss) depends on a subset of the training data, referred to as support vectors. Support  
1086 vectors are a set of data points with some distance from the target values that still allow  
1087 the correct prediction. The systematic reduction of the training data makes this model type  
1088 fundamentally different from the remaining model types of this study. Further information  
1089 is also provided by Smola and Schölkopf (2004).

1090 Due to our large sample size (672,794 flashes), only the *linSVR sklearn.svm.LinearSVR* is  
1091 used in this study in its default configuration. As for the *MLP*, an early stopping criterion  
1092 is used for a lack of improvement between consecutive iterations.

1093 (vii) *Histogram-based Gradient Boosting Regression Tree*. Boosting is, besides bagging,  
1094 another approach to reduce overfitting of ML models. It combines an ensemble of  
1095 weak learners to one strong learner. The Histogram Gradient Boosting Regression  
1096 *sklearn.ensemble.HistGradientBoostingRegressor* (*HGBR*) is much faster than regular Gra-  
1097 dient Boosting Regressors. Data is first binned into 256 integer-valued bins. The algorithm  
1098 can then leverage histograms instead of relying on sorted continuous values when building  
1099 the decision trees. The number of splitting points is reduced and the algorithm becomes  
1100 time efficient, inspired by LightGBM (Ke et al. 2017). The first step of the *HGBR* averages  
1101 the target values and calculates residuals (average difference of observation to prediction)  
1102 with a least-squares loss function. Based on these residuals, a small decision tree is built,

1103 along with a learning rate. The learning rate limits the influence of a single small decision  
1104 tree in the final ensemble to avoid overfitting. Then, new predictions are computed using  
1105 the averages and the decision tree for residuals. Based on new predictions, new residuals  
1106 are calculated and a new decision tree is created. The final model combines several of these  
1107 decision trees to pull the target averages towards the observations. The used maximum  
1108 number of iteration is 500 and the early stopping criteria kicks in after 50 iterations without  
1109 significant improvement of the loss value.

#### 1110 *b. Naming convention for the GEO lightning pseudo-observation generator configurations*

1111 This section defines the meaning of names given to different configurations of a target  
1112 generator. The names and abbreviations of the ML model types can be found in Table B1.  
1113 The given ML model types are used in the first part of the GEO lightning pseudo-observation  
1114 generator referred to as target generator<sup>3</sup>. Table B2 summarizes the feature usage that is  
1115 available for each ML model type available for the target generator. The feature set selections  
1116 indicate whether a single-target or multi-step approach is used. The feature set selection  
1117 called *NLDN* is the default configuration as described. Generators with extension of only  
1118 *default*, *plus*, *raw*, or *raw plus* are single-step approaches, i.e., using 3 times the model of class  
1119 (1) in Figure 3. Multi-step simulations always simulate the GLM flash duration in the first  
1120 step here. The order of the targets number of events per flash and the GLM flash extent is  
1121 not fixed. The extension *num* indicates one additional step only for the pseudo-GLM event  
1122 number per flash using the pseudo-GLM flash duration as pseudo-feature. GEO lightning  
1123 pseudo-observation generator configurations with extension *num ext* and *num ext(a)* have  
1124 two additional steps using different pseudo-features as shown in Table B2. The *num ext(a2)*

---

<sup>3</sup>RF is included in the table for completeness. Only ETR as a special RF model is used in the study.

1125 generators use only the GLM flash duration as pseudo-feature, thus two models of class (2)  
1126 as of Figure 3.

1127 The attributes define a modification of the feature set selections with binary character.  
1128 The *plus* attribute indicates that NLDN LF amplitude and CG fraction are added to the list  
1129 of features. Attribute *raw* means that no feature and target scaling were used. Combinations  
1130 of the given feature set selections and attributes are possible, e.g., an unscaled model with  
1131 NLDN mean LF amplitude and CG stroke ratio as additional features that uses the GLM  
1132 flash duration as pseudo-feature for the event number per flash gets the extension  $num(a)$   
1133 *raw plus*. The total number of generator configurations equals 196: There are 7 ML model  
1134 types (Table B1 except *RF*). For each ML model type there are 7 feature set selections  
1135 resulting from the single and multi-step approaches, and for each combination of ML model  
1136 and feature set selection again 4 different attribute usages (Table B2), i.e., none, *plus*, *raw*,  
1137 or *raw plus*. The 196 generator configurations (28 for each ML model type) define the base  
1138 for the statistical results presented in Section 4.

## 1139 APPENDIX C

### 1140 The multi-target multi-step approach

1141 This section describes a multi-target regression that simplifies the idea of the stacked single  
1142 target (SST) approach (Spyromitros-Xioufis et al. 2016). In this study, there are six NLDN  
1143 features (as physical input) and three GLM targets (as physical simulated variables) per  
1144 sample, i.e., per flash. The three GLM targets are denoted  $T_i$ ,  $T_j$ , and  $T_k$ .  $T_i$  can represent  
1145 any of the 3 targets. The indexes  $i, j, k$  indicate the order of obtaining the final targets.  
1146 Targets that are used like features are referred to as pseudo-features, i.e.,  $T_j$  and  $T_k$  in  
1147 Figure 3a. With this dataset, there are in general 4 different ways to simulate the target  $T_i$ .

1148 The 4 ST models are shown as the training part in Figure 3a. There are 3 classes of models:  
 1149 Yellow is the model class (1) without pseudo-features, gray indicates model class (2) using  
 1150 1 pseudo-feature, and the red for model class (3) uses 2 pseudo-features. The model  $M_{\rightarrow i}$   
 1151 constitutes the common ML model, i.e., class (1), with only the NLDN features as input.  
 1152 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  
 1153 pseudo-features in order to simulate the target  $T_i$ . The resulting models  $M_{j\rightarrow i}$  (using  $T_j$  as  
 1154 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  
 1155 features) may indeed take advantage of correlations between the predicted target and the  
 1156 targets that are used as pseudo-features.

1157 The application case only uses the NLDN features as first input. Therefore, a multi-step  
 1158 approach is required. Figure 3b presents the example application for a 3-step approach  
 1159 that first predicts the pseudo-GLM flash duration, then the pseudo-GLM event number  
 1160 per flash, and finally the pseudo-GLM flash extent. This configuration is denoted *num*  
 1161 *ext(a)* (see Table B2 for details on the configuration naming). The first step,  $M_{\rightarrow i}$ , uses  
 1162 the NLDN features and predicts the first pseudo-GLM characteristic  $M_{\rightarrow i}(NLDN)$ , i.e.,  
 1163 pseudo-GLM flash duration. The second step,  $M_{i\rightarrow j}$ , uses the NLDN features and the result  
 1164 of the first step,  $M_{\rightarrow i}(NLDN)$ , i.e., the pseudo-feature GLM flash duration. This model of  
 1165 class (2) predicts the second pseudo-GLM characteristic  $M_{i\rightarrow j}[NLDN, M_{\rightarrow i}(NLDN)]$ , i.e.,  
 1166 the pseudo-GLM event number per flash. Both predicted pseudo-GLM characteristics (i.e.,  
 1167 GLM flash duration and event number per flash) can then be used as pseudo-features to  
 1168 predict the third target with the class (3) model  $M_{i,j\rightarrow k}$ . Hence, the final target prediction  
 1169  $M_{i,j\rightarrow k}\langle NLDN, M_{i\rightarrow j}[NLDN, M_{\rightarrow i}(NLDN)]\rangle$  depends on the NLDN features and both  
 1170 previous predictions for this configuration. In general, a model of class (3) can also use two  
 1171 pseudo-features produced by two models of class (1). Also, 2 models of class (2) could be

1172 used to simulate the remaining 2 targets after the first step. Utilizing 3 times a model of  
1173 class (1) is equal to the common ML ST approach. Hence, several combinations of models  
1174 of different classes are possible and have been investigated here.

1175 The ML training for the multi-step approach can be performed in parallel for the models  
1176  $M_{\rightarrow i}$ ,  $M_{j \rightarrow i}$ , and  $M_{j,k \rightarrow i}$ . The approach can use all ML model types as the training creates  
1177 independent learners. Our multi-step approach adapts the idea of the SST, but uses GLM  
1178 observations instead of simulated pseudo-GLM targets during the ML training. A trained  
1179 generator can be applied even if the observations are not available using the corresponding  
1180 pseudo-observation in their place. This method assumes similarity between observations and  
1181 pseudo-observations, however, the pseudo-observations only approximate the real observa-  
1182 tions. Our approach does not propagate errors in successive steps. However, the training is  
1183 more efficient than for an SST approach as all generator parts can be trained simultaneously  
1184 rather than waiting for the pseudo-observations to be created. Computational efficiency was  
1185 necessary due to the large number of generators tested in this paper and in the perspective  
1186 of an operational-like application. The results (Section 4) showed that our multi-step ap-  
1187 proach aids in simulating realistic pseudo-GLM observations and the performance is often  
1188 better than with using common ST models without pseudo-features.

1189 Although the correlations between the NLDN features and both GLM flash extent and  
1190 event number per flash are relatively weak, the NLDN features improve the prediction during  
1191 each step as seen through feature drop tests (not shown). Indeed, all features have a positive  
1192 effect on the model score. Due to strong correlations between GLM flash duration and NLDN  
1193 features flash duration and pulse/stroke number, and to reduce the number of ML-based  
1194 target generators, only the multi-step approaches which predict the GLM flash duration in  
1195 the first step ( $M_{\rightarrow i}$ ) are considered. There remains only one model of class (2) in Figure 3a

1196 and three ways to simulate a target  $T_i$ . The GLM flash duration is also weakly correlated  
1197 to both GLM flash extent and event number ( $R$  of about 0.10 and 0.17), and GLM flash  
1198 extent and event number per flash are well correlated ( $R$  of about 0.74). Thus, the first  
1199 step always provides the pseudo-GLM flash duration from the NLDN flash characteristics  
1200 as features. The second step uses the simulated flash duration in addition to the NLDN  
1201 features to simulate one or both of pseudo-GLM flash extent and event number per flash.  
1202 The pseudo-feature used in model class (2) (Figure 3) is fixed in this paper to be flash  
1203 duration leaving only one realization of model class (2) to simulate a second target. To  
1204 further reduce the number of multi-step configurations, the approaches that simulate the  
1205 flash extent but not the event number per flash through a multi-step process are not further  
1206 considered since (i) GLM flash duration shows weaker correlation with GLM flash extent  
1207 than with the event number per flash and (ii) the ST approach for event number per flash  
1208 from NLDN features exhibits the lowest skill of the three targets. A potential third step may  
1209 simulate the last GLM target based on NLDN features and the two remaining simulated  
1210 pseudo-GLM characteristics as additional pseudo-features. The paper describes generator  
1211 configurations using only the GLM duration (strongest correlations) or using GLM duration  
1212 and a second target as additional pseudo-features to simulate the remaining target (GLM  
1213 flash extent or event number per flash).

1214 Our multi-step approach aims at producing more realistic pseudo-GLM flash extent and  
1215 event number per flash than using the NLDN features alone. The NLDN features also remain  
1216 important as the correlations between some targets are weak.

## 1217 APPENDIX D

### 1218 **Supplementary results for each test day and target distribution statistic**

1219 This section contains detailed results for each test day that are presented in the main  
1220 paper for the combined 3-day period. The second part includes the Figures and analysis  
1221 of the normalized difference between distribution statistics of observed and simulated GLM  
1222 flash duration and event per flash. The results are presented in a similar way as for the flash  
1223 extent statistics in the main paper.

1224 *a. Target generator results for each test day*

1225 Tables D1, D2, and D3 present the results for 07 April 2018, 26 May 2018, and 31 July  
1226 2018, respectively. As explained for Table 2 with results for the 3 days combined, the  
1227 tables show the observed statistics for each target distribution, the outcomes using the  
1228 best performing generator, and statistics of data simulated with the *linSVR num ext(a)*  
1229 *plus* recommended generator. The most common behavior of the target generators exhibits  
1230 simulated mean values close to the observation statistics of the three target distribution.  
1231 Median values are usually underestimated by the target generators as seen in Tables D2 and  
1232 D3. Results for the 07 April 2018 test case differ from the general behavior (Table D1). That  
1233 day saw exceptionally large flashes with high event numbers per flash that likely occurred  
1234 within the MCS and the squall line. As a consequence, the ML-based target generators  
1235 underestimated the means of the observed flash characteristics for that test case, but the  
1236 medians of simulated and observed targets are similar.

1237 The results for each test day resample the results for the combined 3-day test data (see  
1238 Table 2) overall. Minimum values are often only slightly overestimated for the three targets,  
1239 while the simulated maxima cannot reach the observed maxima for none of the targets and  
1240 on none of the 3 test days (Tables D1 to D3). *linSVR*-based target generator outperform  
1241 all other generator types on 26 May and on 31 July 2020. On April 07 with extensively



1242 large flashes, different *BAGR KNN dist*-based generators are found as best performers for  
1243 all three targets (Table D1). The choice of the most suitable generator appears to be  
1244 situational, i.e., there is no generator that performs better than all other generators in all  
1245 cases. The recommended *linSVR num ext(a) plus* (bold in Tables D1 to D3) performs on  
1246 one level with best generator for the event number per flash and flash extent on 26 May  
1247 an on 31 July 2020. The event number per flash is significantly underestimated by this  
1248 *linSVR*-based generator on 07 April 2020 for the mentioned reason. The flash extent, as  
1249 most important target for the FED, is also underestimated on that day, however, the CvM  
1250 is only about half of the CvM for event number per flash meaning a more realistic simulation  
1251 of the flash extent distribution than the event number per flash distribution.

1252 *b. Normalized statistics for difference between observation and simulation for GLM flash*  
1253 *duration and event number per flash*

1254 Figures D1 and D2 group the results for each statistic by the seven ML model types. As  
1255 explained in Section 4.a, each distribution contains the results of 28 generators (see also  
1256 Appendix B.b, Table B2).

1257 Figure D1 shows the normalized differences and scores of different target generators for  
1258 the GLM flash duration for the 3 test days combined. The GLM flash duration distribution  
1259 is equally well simulated by a variety of ML-based target generators as the narrow spread  
1260 of the medians (green line) indicates. In detail, a *linSVR*-based generator and a *MLP*-  
1261 based generator can predict the mean well, a *MLP*-based generator and an *ETR*-based  
1262 generator are best for the maximum, while a *linSVR*-based generator exhibits the lowest  
1263 differences for the median as well as both KS and CvM scores. In total, a *linSVR*-based

1264 target generator (*linSVR num ext raw*) best approximates the observed distribution of the  
1265 GLM flash duration in this comparison with an NDA value of 0.30.

1266 For the GLM event number per flash in Figure D2, *linSVR* and *BAGR KNN dist* models  
1267 make the best target generators. The lowest NDA of 0.45 is obtained for several *linSVR*  
1268 and *BAGR KNN dist*-based generators, e.g., *linSVR num ext(a) raw plus* and *BAGR KNN*  
1269 *dist num ext raw*. For test day 07 April 2018 (a dominant mesoscale system with above-  
1270 average mean and median GLM event numbers per flash), all generators underestimated the  
1271 event number per flash. As generators using a *linSVR* usually predict lower values than the  
1272 other generators, they underestimate the observed statistics even more on 07 April 2018.  
1273 Nevertheless, for the full test data, there are *linSVR*-based generators that predict the mean  
1274 event number equally well as the best target generator, i.e., *MLP*-based, as demonstrated  
1275 by the lower whiskers in Figure D2. *LinSVR*-based generators are again most suitable to  
1276 predict the event number distribution median.

## 1277 **References**

- 1278 Aguiar, G. J., E. J. Santana, S. M. Mastelini, R. G. Mantovani, and S. Barbon, Jr, 2019:  
1279 Towards meta-learning for multi-target regression problems. URL [https://arxiv.org/abs/](https://arxiv.org/abs/1907.11277)  
1280 1907.11277.
- 1281 Allen, B. J., E. R. Mansell, D. C. Dowell, and W. Deierling, 2016: Assimilation of pseudo-  
1282 GLM data using the ensemble Kalman filter. *Monthly Weather Review*, **144**, 3465–3486,  
1283 doi:10.1175/MWR-D-16-0117.1.
- 1284 Altman, N. S., 1992: An introduction to kernel and nearest-neighbor nonparametric regres-  
1285 sion. *The American Statistician*, **46** (3), 175–185, doi:10.1080/00031305.1992.10475879,

1286 URL <https://amstat.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879>.

1287 Anderson, T. W., 1962: On the distribution of the two-sample Cramér-von Mises criterion.

1288 *The Annals of Mathematical Statistics*, **33** (3), 1148–1159, doi:10.1214/aoms/1177704477.

1289 Ávila, E. E., R. E. Bürgesser, N. E. Castellano, A. B. Collier, R. H. Compagnucci, and

1290 A. R. Hughes, 2010: Correlations between deep convection and lightning activity on

1291 a global scale. *Journal of Atmospheric and Solar-Terrestrial Physics*, **72** (14), 1114

1292 – 1121, doi:<https://doi.org/10.1016/j.jastp.2010.07.019>, URL <http://www.sciencedirect.com/science/article/pii/S1364682610002154>.

1293

1294 Bateman, M., 2013: A high-fidelity proxy dataset for the geostationary lightning map-

1295 per (GLM). *AMS Sixth Conference on the Meteorological Application of Lightning Data*,

1296 Austin, TX, USA, 725.

1297 Bateman, M., and D. Mach, 2020: Preliminary detection efficiency and false alarm rate

1298 assessment of the Geostationary Lightning Mapper on the GOES-16 satellite. *Journal*

1299 *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>.

1300

1301 Bateman, M., D. Mach, and M. Stock, 2021: Further investigation into detection ef-

1302 ficiency and false alarm rate for the geostationary lightning mappers aboard GOES-

1303 16 and GOES-17. *Earth and Space Science*, **8** (2), e2020EA001237, doi:<https://doi.org/10.1029/2020EA001237>,

1304 URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020EA001237>,

1305 [https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/](https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020EA001237)

1306 2020EA001237.

- 1307 Bentley, J. L., 1975: Multidimensional binary search trees used for associative searching.  
1308 *Communications of the ACM*, **18 (9)**, 509–517, doi:10.1145/361002.361007, URL https:  
1309 //doi.org/10.1145/361002.361007.
- 1310 Betz, H. D., K. Schmidt, P. Laroche, P. Blanchet, W. P. Oettinger, E. Defer, Z. Dziewit,  
1311 and J. Konarski, 2009: LINET—an international lightning detection network in eu-  
1312 rope. *Atmospheric Research*, **91 (2)**, 564 – 573, doi:10.1016/j.atmosres.2008.06.012,  
1313 URL <http://www.sciencedirect.com/science/article/pii/S0169809508002305>, 13th Inter-  
1314 national Conference on Atmospheric Electricity.
- 1315 Biron, D., L. D. Leonibus, P. Laquale, D. Labate, F. Zauli, and D. Melfi, 2008: Simulation of  
1316 Meteosat Third Generation-Lightning Imager through tropical rainfall measuring mission:  
1317 Lightning Imaging Sensor data. *Remote Sensing System Engineering*, P. E. Ardanuy, and  
1318 J. J. Puschell, Eds., SPIE, International Society for Optics and Photonics, Vol. 7087, 77  
1319 – 88, doi:10.1117/12.794764.
- 1320 Blakeslee, R., and W. Koshak, 2016: LIS on ISS: Expanded global coverage and enhanced  
1321 applications. *The Earth Observer*, **28**, 4–14.
- 1322 Blakeslee, R. J., and Coauthors, 2020: Three years of the lightning imaging sensor on-  
1323 board the international space station: Expanded global coverage and enhanced applica-  
1324 tions. *Journal of Geophysical Research: Atmospheres*, **125 (16)**, e2020JD032918, doi:  
1325 10.1029/2020JD032918, URL [https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD032918)  
1326 [2020JD032918](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD032918).
- 1327 Borchani, H., G. Varando, C. Bielza, and P. Larrañaga, 2015: A survey on multi-output  
1328 regression. *WIREs Data Mining and Knowledge Discovery*, **5 (5)**, 216–233, doi:10.1002/  
1329 widm.1157, URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1157>.

- 1330 Breiman, L., 1996: Bagging predictors. *Machine Learning*, **24**(2), 123–140, doi:10.1023/A:  
1331 1018054314350.
- 1332 Brooks, I. M., C. P. R. Saunders, R. P. Mitzewa, and S. L. Peck, 1997: The effect on  
1333 thunderstorm charging of the rate of rime accretion by graupel. *Atmospheric Research*,  
1334 **43** (3), 277 – 295, doi:10.1016/S0169-8095(96)00043-9, URL <http://www.sciencedirect.com/science/article/pii/S0169809596000439>.
- 1336 Bruning, E. C., and Coauthors, 2019: Meteorological imagery for the geostationary lightning  
1337 mapper. *Journal of Geophysical Research: Atmospheres*, **124** (24), 14 285–14 309, doi:  
1338 10.1029/2019JD030874.
- 1339 Cecil, D. J., S. J. Goodman, D. J. Boccippio, E. J. Zipser, and S. W. Nesbitt, 2005: Three  
1340 years of TRMM precipitation features. Part I: Radar, radiometric, and lightning charac-  
1341 teristics. *Monthly Weather Review*, **133**, 543–566, doi:10.1175/MWR-2876.1.
- 1342 Chmielewski, V. C., and E. C. Bruning, 2016: Lightning mapping array flash detection per-  
1343 formance with variable receiver thresholds. *Journal of Geophysical Research: Atmospheres*,  
1344 **121** (14), 8600–8614, doi:10.1002/2016JD025159, URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JD025159>.
- 1346 Christian, H. J., and Coauthors, 1999: The lightning imaging sensor. *11th International*  
1347 *Conference on Atmospheric Electricity*, 746–749, NASA Conf. Publ. NASA/CP-1999-  
1348 209261a.
- 1349 Coquillat, S., and Coauthors, 2019: SAETTA: high-resolution 3-D mapping of the total  
1350 lightning activity in the Mediterranean Basin over Corsica, with a focus on a mesoscale

1351 convective system event. *Atmospheric Measurement Techniques*, **12** (11), 5765–5790, doi:  
1352 10.5194/amt-12-5765-2019, URL <https://amt.copernicus.org/articles/12/5765/2019/>.

1353 Cummins, K. L., and M. J. Murphy, 2009: An overview of lightning locating systems:  
1354 History, techniques, and uses, with an in-depth look at the U.S. NLDN. *IEEE Trans.*  
1355 *Electromag. Compat.*, **51**(3), 499–518, doi:10.1109/TEMC.2009.2023450.

1356 Deierling, W., and W. A. Petersen, 2008: Total lightning activity as an indicator of  
1357 updraft characteristics. *Journal of Geophysical Research*, **113**, D16 210, doi:10.1029/  
1358 2007JD009598.

1359 Dobber, M., and J. Grandell, 2014: Meteosat Third Generation (MTG) Lightning Imager  
1360 (LI) instrument performance and calibration from user perspective. *Proceedings of the*  
1361 *23rd Conference on Characterization and Radiometric Calibration for Remote Sensing*  
1362 *(CALCON)*, 11-14 August 2014, Utah State University, Logan, Utah, USA. 13 pages.

1363 Emersic, C., and C. Saunders, 2020: The influence of supersaturation at low rime accretion  
1364 rates on thunderstorm electrification from field-independent graupel-ice crystal collisions.  
1365 *Atmospheric Research*, **242**, 104 962, doi:10.1016/j.atmosres.2020.104962, URL [http://](http://www.sciencedirect.com/science/article/pii/S0169809519312803)  
1366 [www.sciencedirect.com/science/article/pii/S0169809519312803](http://www.sciencedirect.com/science/article/pii/S0169809519312803).

1367 Erdmann, F., 2020: Préparation à l’utilisation des observations de l’imageur d’éclairs de  
1368 météosat troisième génération pour la prévision numérique à courte échéance (Prepara-  
1369 tion for the use of Meteosat Third Generation Lightning Imager observations in short-  
1370 term numerical weather prediction). Ph.D. thesis, Université Toulouse 3 – Paul Sabatier,  
1371 Toulouse, France.

- 1372 Erdmann, F., E. Defer, O. Caumont, R. J. Blakeslee, S. Pédeboy, and S. Coquillat,  
1373 2020: Concurrent satellite and ground-based lightning observations from the optical light-  
1374 ning imaging sensor (ISS-LIS), the low-frequency network meteorage and the SAETTA  
1375 lightning mapping array (LMA) in the northwestern Mediterranean region. *Atmospheric*  
1376 *Measurement Techniques*, **13** (2), 853–875, doi:10.5194/amt-13-853-2020, URL <https://www.atmos-meas-tech.net/13/853/2020/>.  
1377
- 1378 Fierro, A. O., Y. Wang, J. Gao, and E. R. Mansell, 2019: Variational assimilation  
1379 of radar data and GLM lightning-derived water vapor for the short-term forecasts of  
1380 high-impact convective events. *Monthly Weather Review*, **147**, 4045–4069, doi:10.1175/  
1381 MWR-D-18-0421.1.
- 1382 Geurts, P., D. Ernst, and L. Wehenkel, 2006: Extremely randomized trees. *Machine Learn-*  
1383 *ing*, **63**, 3–42, doi:10.1007/s10994-006-6226-1.
- 1384 Glorot, X., and Y. Bengio, 2010: Understanding the difficulty of training deep feedforward  
1385 neural networks. *JMLR W&CP: 13th International Conference on Artificial Intelligence*  
1386 *and Statistics*, Sardinia, Italy, Vol. 9, 249–256.
- 1387 Goodman, S., D. Mach, W. Koshak, and R. Blakeslee, 2012: GLM lightning cluster-filter  
1388 algorithm. Algorithm theoretical basis document, NOAA NESDIS Center for Satellite  
1389 Application and Research.
- 1390 Goodman, S. J., and Coauthors, 2013: The GOES-R Geostationary Lightning Mapper  
1391 (GLM). *Atmospheric Research*, **125-126**, 34 – 49, doi:10.1016/j.atmosres.2013.01.006,  
1392 URL <http://www.sciencedirect.com/science/article/pii/S0169809513000434>.

- 1393 Höller, H., and H.-D. Betz, 2010: Study on inter-comparison of LIS and ground-based  
1394 lightning location system observations. Report ITT No. 09/996, EUMETSAT, Deutsches  
1395 Zentrum für Luft- und Raumfahrt.
- 1396 Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and  
1397 T.-Y. Liu, 2017: LightGBM: A highly efficient gradient boosting decision  
1398 tree. *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V.  
1399 Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Gar-  
1400 nett, Eds., Curran Associates, Inc., 3146–3154, URL [http://papers.nips.cc/paper/  
1401 6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf](http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf).
- 1402 Kingma, D. P., and J. Ba, 2014: Adam: A method for stochastic optimization. URL [https:  
1403 //arxiv.org/abs/1412.6980](https://arxiv.org/abs/1412.6980).
- 1404 Kolmasova, I., T. Marshall, S. Bandara, S. Karunarathne, M. Stolzenburg, N. Karunarathne,  
1405 and R. Siedlecki, 2019: Initial breakdown pulses accompanied by VHF pulses during  
1406 negative cloud-to-ground lightning flashes. *Geophysical Research Letters*, **46**, 5592–5600,  
1407 doi:10.1029/2019GL082488.
- 1408 Koshak, W., D. Mach, M. Bateman, P. Armstrong, and K. Virts, 2010: GOES-16 GLM  
1409 level 2 data full validation data quality - product performance guide for data users. Guide,  
1410 NASA Marshall Space Flight Center.
- 1411 Koshak, W. J., and R. J. Solakiewicz, 2015: A method for retrieving the ground flash fraction  
1412 and flash type from satellite lightning mapper observations. *Journal of Atmospheric and  
1413 Oceanic Technology*, **32** (1), 79 – 96, doi:10.1175/JTECH-D-14-00085.1, URL [https://  
1414 journals.ametsoc.org/view/journals/atot/32/1/jtech-d-14-00085\\_1.xml](https://journals.ametsoc.org/view/journals/atot/32/1/jtech-d-14-00085_1.xml).



- 1415 Koshak, W. J., and Coauthors, 2004: North Alabama Lightning Mapping Array (LMA):  
1416 VHF source retrieval algorithm and error analyses. *Journal of Atmospheric and Oceanic*  
1417 *Technology*, **21** (4), 543 – 558, doi:10.1175/1520-0426(2004)021<0543:NALMAL>2.0.  
1418 CO;2, URL [https://journals.ametsoc.org/view/journals/atot/21/4/1520-0426\\_2004\\_021\\_](https://journals.ametsoc.org/view/journals/atot/21/4/1520-0426_2004_021_0543_nalmal_2_0_co_2.xml)  
1419 [0543\\_nalmal\\_2\\_0\\_co\\_2.xml](https://journals.ametsoc.org/view/journals/atot/21/4/1520-0426_2004_021_0543_nalmal_2_0_co_2.xml).
- 1420 Luque, M. Y., F. Nollas, R. G. Pereyra, R. E. Bürgesser, and E. E. Ávila, 2020:  
1421 Charge separation in collisions between ice crystals and a spherical simulated graupel  
1422 of centimeter size. *Journal of Geophysical Research: Atmospheres*, **125** (3),  
1423 e2019JD030941, doi:10.1029/2019JD030941, URL [https://agupubs.onlinelibrary.wiley.](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD030941)  
1424 [com/doi/abs/10.1029/2019JD030941](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD030941).
- 1425 Lyu, F., S. A. Cummer, Z. Qin, and M. Chen, 2019: Lightning initiation processes imaged  
1426 with very high frequency broadband interferometry. *Journal of Geophysical Research:*  
1427 *Atmospheres*, **124**, 2994–3004, doi:10.1029/2018JD029817.
- 1428 MacGorman, D. R., and W. D. Rust, 1998: *The electrical nature of storms*. 1st ed., Oxford  
1429 University Press, 198 Madison Avenue, New York, New York 10016.
- 1430 Mach, D. M., 2020: Geostationary lightning mapper clustering algorithm stabil-  
1431 ity. *Journal of Geophysical Research: Atmospheres*, **125** (5), e2019JD031900,  
1432 doi:10.1029/2019JD031900, URL [https://agupubs.onlinelibrary.wiley.com/doi/abs/10.](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD031900)  
1433 [1029/2019JD031900](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD031900).
- 1434 Marchand, M., K. Hilburn, and S. D. Miller, 2019: Geostationary lightning mapper and  
1435 earth networks lightning detection over the contiguous United States and dependence on  
1436 flash characteristics. *Journal of Geophysical Research: Atmospheres*, **124** (21), 11 552–

- 1437 11 567, doi:10.1029/2019JD031039, URL [https://agupubs-onlinelibrary-wiley-com-s-](https://agupubs-onlinelibrary-wiley-com-s)  
1438 [docadis.ups-tlse.fr/doi/abs/10.1029/2019JD031039](https://agupubs-onlinelibrary-wiley-com-s-docadis.ups-tlse.fr/doi/abs/10.1029/2019JD031039).
- 1439 Massey, F. J., 1951: The Kolmogorov-Smirnov test for goodness of fit. *Journal of the Ameri-*  
1440 *can Statistical Association*, **46 (253)**, 68–78, URL <http://www.tandfonline.com/doi/abs/>  
1441 [10.1080/01621459.1951.10500769](http://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500769).
- 1442 Mastelini, S. M., V. G. T. da Costa, E. J. Santana, F. K. Nakano, R. C. Guido, R. Cerri, and  
1443 S. Barbon, 2019: Multi-output tree chaining: An interpretative modelling and lightweight  
1444 multi-target approach. *Journal of Signal Processing Systems*, **91 (2)**, 191–215, doi:10.  
1445 [1007/s11265-018-1376-5](https://doi.org/10.1007/s11265-018-1376-5), URL <https://doi.org/10.1007/s11265-018-1376-5>.
- 1446 Murphy, M. J., and R. K. Said, 2020: Comparisons of lightning rates and properties from the  
1447 U.S. National Lightning Detection Network (NLDN) and GLD360 with GOES-16 Geo-  
1448 stationary Lightning Mapper and Advanced Baseline Imager data. *Journal of Geophysical*  
1449 *Research: Atmospheres*, **125 (5)**, e2019JD031172, doi:10.1029/2019JD031172.
- 1450 Nag, A., M. J. Murphy, W. Schulz, and K. L. Cummins, 2015: Lightning locating systems:  
1451 Insights on characteristics and validation techniques. *Earth and Space Science*, **2**, 65–93,  
1452 doi:10.1002/2014EA000051.
- 1453 Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *Journal of*  
1454 *Machine Learning Research*, **12**, 2825–2830.
- 1455 Peterson, M., S. Rudlosky, and D. Zhang, 2020: Changes to the appearance of opti-  
1456 cal lightning flashes observed from space according to thunderstorm organization and  
1457 structure. *Journal of Geophysical Research: Atmospheres*, **125 (4)**, e2019JD031087, doi:  
1458 [10.1029/2019JD031087](https://doi.org/10.1029/2019JD031087).

- 1459 Rison, W., R. J. Thomas, P. R. Krehbiel, T. Hamlin, and J. Harlin, 1999: A GPS-based  
1460 three-dimensional lightning mapping system: Initial observations in central New Mexico.  
1461 *Geophysical Research Letters*, **26** (23), 3573–3576, doi:10.1029/1999GL010856.
- 1462 Rutledge, S. A., K. A. Hilburn, A. Clayton, B. Fuchs, and S. D. Miller, 2020: Evaluating geo-  
1463 stationary lightning mapper flash rates within intense convective storms. *Journal of Geo-*  
1464 *physical Research: Atmospheres*, **125** (14), e2020JD032827, doi:10.1029/2020JD032827,  
1465 URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD032827>.
- 1466 Schultz, E. V., C. J. Schultz, L. D. Carey, D. J. Cecil, and M. Bateman, 2016: Automated  
1467 storm tracking and the lightning jump algorithm using GOES-R geostationary lightning  
1468 mapper (GLM) proxy data. *Journal of Operational Meteorology*, **4**(7), 92–107, doi:10.  
1469 15191/nwajom.2016.0407.
- 1470 Schulz, W., G. Diendorfer, S. Pedeboy, and D. R. Poelman, 2016: The European lightning  
1471 location system EUCLID – Part 1: Performance analysis and validation. *Nat. Hazards*  
1472 *Earth Syst. Sci.*, **16**, 595–605, doi:10.5194/nhess-16-595-2016.
- 1473 Smola, A. J., and B. Schölkopf, 2004: A tutorial on support vector regression. *Statistics and*  
1474 *Computing*, **14**, 199–222, doi:10.1023/B:STCO.0000035301.49549.88.
- 1475 Spyromitros-Xioufis, E., G. Tsoumakas, W. Groves, and I. Vlahavas, 2016: Multi-target re-  
1476 gression via input space expansion: treating targets as inputs. *Machine Learning*, **104** (1),  
1477 55–98, doi:10.1007/s10994-016-5546-z, URL <https://doi.org/10.1007/s10994-016-5546-z>.
- 1478 Stano, G. T., 2013: Fusing total lightning data with aviation weather center and storm  
1479 prediction center operations during the GOES-R visiting scientist program. *AMS Sixth*  
1480 *Conference on the Meteorological Applications of Lightning Data*, Austin, TX, USA, 724.

- 1481 Takahashi, T., S. Sugimoto, T. Kawano, and K. Suzuki, 2017: Riming Electrification in  
1482 Hokuriku Winter Clouds and Comparison with Laboratory Observations. *Journal of the*  
1483 *Atmospheric Sciences*, **74** (2), 431–447, doi:10.1175/JAS-D-16-0154.1, URL [https://doi.](https://doi.org/10.1175/JAS-D-16-0154.1)  
1484 [org/10.1175/JAS-D-16-0154.1](https://doi.org/10.1175/JAS-D-16-0154.1).
- 1485 Taszarek, M., J. T. Allen, P. Groenemeijer, R. Edwards, H. E. Brooks, V. Chmielewski, and  
1486 S.-E. Enno, 2020: Severe convective storms across Europe and the United States. Part  
1487 I: Climatology of lightning, large hail, severe wind, and tornadoes. *Journal of Climate*,  
1488 **33** (23), 10 239 – 10 261, doi:10.1175/JCLI-D-20-0345.1, URL [https://journals.ametsoc.](https://journals.ametsoc.org/view/journals/clim/33/23/jcliD200345.xml)  
1489 [org/view/journals/clim/33/23/jcliD200345.xml](https://journals.ametsoc.org/view/journals/clim/33/23/jcliD200345.xml).
- 1490 Thomas, R. J., P. R. Krehbiel, W. Rison, T. Hamlin, D. J. Boccippio, S. J. Goodman,  
1491 and H. J. Christian, 2000: Comparison of ground-based 3-dimensional lightning mapping  
1492 observations with satellite-based LIS observations in Oklahoma. *Geophysical Research*  
1493 *Letters*, **27** (12), 1703–1706, doi:10.1029/1999GL010845.
- 1494 Thomas, R. J., P. R. Krehbiel, W. Rison, S. J. Hunyady, W. P. Winn, T. Hamlin, and  
1495 J. Harlin, 2004: Accuracy of the lightning mapping array. *Journal of Geophysical Research*,  
1496 **109**, D14 207, doi:10.1029/2004JD004549.
- 1497 Vaisala, 2013: Vaisala thunderstorm advanced total lightning sen-  
1498 sor LS7002. URL [https://www.vaisala.com/sites/default/files/documents/](https://www.vaisala.com/sites/default/files/documents/WEA-LS7002-Datasheet-B211284EN-A-LOW.pdf)  
1499 [WEA-LS7002-Datasheet-B211284EN-A-LOW.pdf](https://www.vaisala.com/sites/default/files/documents/WEA-LS7002-Datasheet-B211284EN-A-LOW.pdf), accessed on 2018-06-28.
- 1500 Yang, J., Z. Zhang, C. Wei, F. Lu, and Q. Guo, 2017: Introducing the new generation of chi-  
1501 nese geostationary weather satellites, Fengyun-4. *Bulletin of the American Meteorological*  
1502 *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), e2019JD032024, 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://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JD025574>.

1512 **LIST OF TABLES**

1513 **Table 1.** Study dates (year 2018) with the amounts of GLM and NLDN data.  
1514 The three rightmost columns indicate whether the data are used for  
1515 ML-based generator building (GB) or the generator test (GT) part,  
1516 the time of most lightning activity in the region (D: local daytime,  
1517 N: local nighttime), and the primary forcing (Trigger) for storm  
1518 development and lightning. . . . . 71

1519 **Table 2.** Comparison of distribution statistics for observed GLM data and  
1520 the best generator for each target during the full test period. The  
1521 recommended *linSVR*-based generator is shown in bold. Details  
1522 about the target generator names are provided in the Appendix B.b. . . . 72

1523 **Table 3.** Comparison of  $D_{real}$  and  $D_{abs}$  in percent of observed value for the  
1524 FED sum during the full test period. The best 20 and the worst  
1525 5 of the 196 generators (ranked by  $D_{abs}$ ) are included. The recom-  
1526 mended *linSVR*-based generator is shown in bold. Details about  
1527 the target generator names are provided in the Appendix B.b. . . . . 73

1528 **Table 4.** Comparison of  $D_{real}$  and  $D_{abs}$  in percent of observed value for the  
1529 electrified area during the full test period. The best 20 and the worst  
1530 5 of the 196 generators (ranked by  $D_{abs}$ ) are included. In addition,  
1531 the recommended *linSVR*-based generator is shown in bold. Details  
1532 about the target generator names are provided in the Appendix B.b. . . . 74

1533 **Table B1.** ML model types with abbreviation. . . . . 75

1534 **Table B2.** Naming conventions of used target generator configurations. The  
1535 name extensions as of column 1 are used following the ML model  
1536 type. The three columns indicate the utilized features during the  
1537 ML training for each of the three targets GLM flash duration (Flash  
1538 Duration), number of events per flash (Event Number), and GLM  
1539 flash extent (Flash Extent). NLDN indicates that NLDN flash du-  
1540 ration, the number of pulses/strokes per flash, NLDN flash extent,  
1541 and the maximum LF amplitude are used as features. The GLM  
1542 pseudo-features Flash Duration, Flash Extent and/or Event Num-  
1543 ber can complement the NLDN features for some configurations.  
1544 Feature set selections define how one target (header) is generated,  
1545 i.e., ST or multi-step approach. The attributes can or cannot be  
1546 applied and may replace *default* in the generator name. Combina-  
1547 tions of a feature set selection with 0, 1, or 2 attributes are possible.  
1548 76

1549 **Table D1.** Comparison of distribution statistics for observed GLM data and  
1550 the best generator for each target on 07 April 2018. The recom-  
1551 mended *linSVR*-based generator is shown in bold. Details about  
1552 the target generator names are provided in the Appendix B.b. . . . . 77

1553 **Table D2.** Comparison of distribution statistics for observed GLM data and  
1554 the best generator for each target on 26 May 2018. The recom-  
1555 mended *linSVR*-based generator is shown in bold. Details about  
1556 the target generator names are provided in the Appendix B.b. . . . . 78

1557 **Table D3.** Comparison of distribution statistics for observed GLM data and  
1558 the best generator for each target on 31 July 2018. The recom-  
1559 mended *linSVR*-based generator is shown in bold. Details about  
1560 the target generator names are provided in the Appendix B.b. . . . . 79

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
<i>linSVR num ext raw</i>	0.46	0.30	0.00	9.41	0.77	0.15	656.5
<b><i>linSVR num ext(a) plus</i></b>	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
<i>linSVR num ext(a) raw plus</i>	35.3	30.0	2	411	79.8	0.38	6687.6
<b><i>linSVR num ext(a) plus</i></b>	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
<i>linSVR num ext(a2)</i>	30.1	26.3	0.0	157.3	30.9	0.24	3479.2
<b><i>linSVR num ext(a) plus</i></b>	29.0	26.2	8.5	305.5	31.5	0.24	2738.4

1568 TABLE 3. Comparison of  $D_{real}$  and  $D_{abs}$  in percent of observed value for the FED sum during the  
1569 full test period. The best 20 and the worst 5 of the 196 generators (ranked by  $D_{abs}$ ) are included.  
1570 The recommended *linSVR*-based generator is shown in bold. Details about the target generator  
1571 names are provided in the Appendix B.b.

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

1572 TABLE 4. Comparison of  $D_{real}$  and  $D_{abs}$  in percent of observed value for the electrified area  
1573 during the full test period. The best 20 and the worst 5 of the 196 generators (ranked by  $D_{abs}$ ) are  
1574 included. In addition, the recommended *linSVR*-based generator is shown in bold. Details about  
1575 the target generator names are provided in the Appendix B.b.

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

Table B1. ML model types with abbreviation.

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

---

<sup>1</sup>This study uses ETR as a special form of RF.

1576 Table B2. Naming conventions of used target generator configurations. The name extensions  
1577 as of column 1 are used following the ML model type. The three columns indicate the utilized  
1578 features during the ML training for each of the three targets GLM flash duration (Flash Duration),  
1579 number of events per flash (Event Number), and GLM flash extent (Flash Extent). NLDN indicates  
1580 that NLDN flash duration, the number of pulses/strokes per flash, NLDN flash extent, and the  
1581 maximum LF amplitude are used as features. The GLM pseudo-features Flash Duration, Flash  
1582 Extent and/or Event Number can complement the NLDN features for some configurations. Feature  
1583 set selections define how one target (header) is generated, i.e., ST or multi-step approach. The  
1584 attributes can or cannot be applied and may replace *default* in the generator name. Combinations  
1585 of a feature set selection with 0, 1, or 2 attributes are possible.

Name extension	Flash Duration	Event Number	Flash Extent
Feature set selections			
<i>default</i>	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)		
<i>num</i>	NLDN	NLDN + Flash Duration + Flash Extent	NLDN
<i>num(a)</i>	NLDN	NLDN + Flash Duration	NLDN
<i>num ext</i>	NLDN	NLDN + Flash Duration + Flash Extent	NLDN + Flash Duration + Event Number
<i>num ext(a)</i>	NLDN	NLDN + Flash Duration	NLDN + Flash Duration + Event Number
<i>num ext(a2)</i>	NLDN	NLDN + Flash Duration	NLDN + Flash Duration
Attributes			
<i>plus</i>	NLDN with mean LF amplitude and CG fraction	NLDN with mean LF amplitude and CG fraction	NLDN with mean LF amplitude and CG fraction
<i>raw</i>	features and target not scaled	features and target not scaled	features and target not scaled

1587 Table D1. Comparison of distribution statistics for observed GLM data and the best generator  
1588 for each target on 07 April 2018. The recommended *linSVR*-based generator is shown in bold.  
1589 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
<i>BAGR KNN dist num ext default</i>	0.57	0.46	0.01	10.56	0.92	0.25	949.2
<b><i>linSVR num ext(a) plus</i></b>	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
<i>BAGR KNN dist num raw</i>	57.8	48.0	3	467	99.7	0.24	996.6
<b><i>linSVR num ext(a) plus</i></b>	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
<i>BAGR KNN dist num ext(a2) raw plus</i>	41.24	35.1	0.0	166.0	33.7	0.24	582.7
<b><i>linSVR num ext(a) plus</i></b>	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
<i>linSVR raw</i>	0.41	0.25	0.01	4.73	0.56	0.21	599.9
<b><i>linSVR num ext(a) plus</i></b>	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
<i>linSVR num ext(a) raw plus</i>	32.2	29.0	3	341	54.0	0.45	3593.1
<b><i>linSVR num ext(a) plus</i></b>	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
<i>linSVR num ext raw</i>	27.4	26.1	0.0	154.8	27.8	0.28	1029.2
<b><i>linSVR num ext(a) plus</i></b>	26.7	26.1	8.5	157.6	27.2	0.29	1130.4

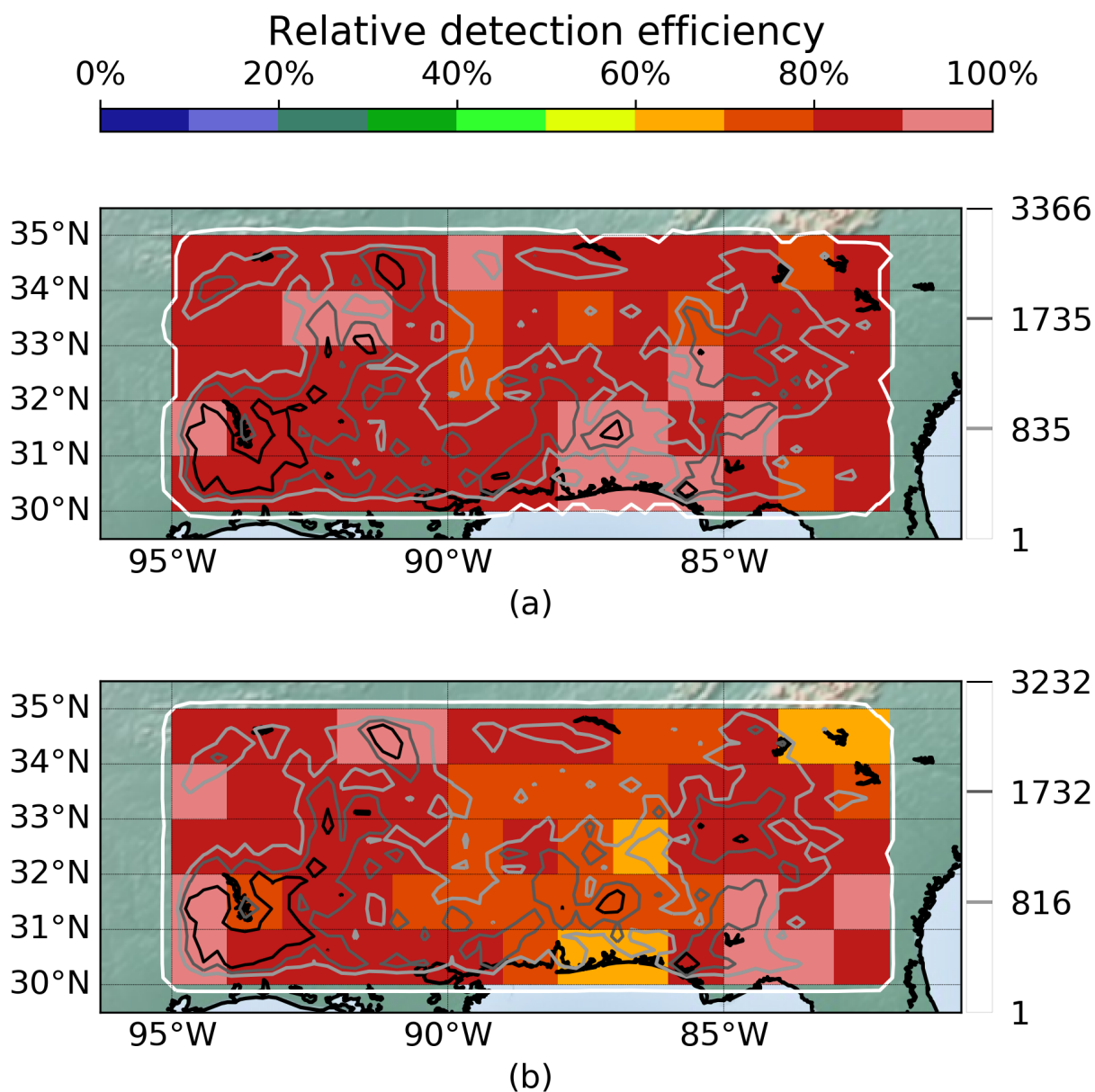
1595 Table D3. Comparison of distribution statistics for observed GLM data and the best generator  
1596 for each target on 31 July 2018. The recommended *linSVR*-based generator is shown in bold.  
1597 Details about the target generator names are provided in the Appendix B.b.

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
<b><i>linSVR num</i></b> <b><i>ext(a) plus</i></b>	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
<i>linSVR num</i> <i>ext(a) raw plus</i>	33.9	30.0	2	316	76.4	0.40	2441.8
<b><i>linSVR num</i></b> <b><i>ext(a) plus</i></b>	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
<i>linSVR raw</i>	29.6	26.2	0.0	180.6	33.5	0.22	745.2
<b><i>linSVR num</i></b> <b><i>ext(a) plus</i></b>	29.5	26.2	8.5	305.5	32.6	0.24	871.4

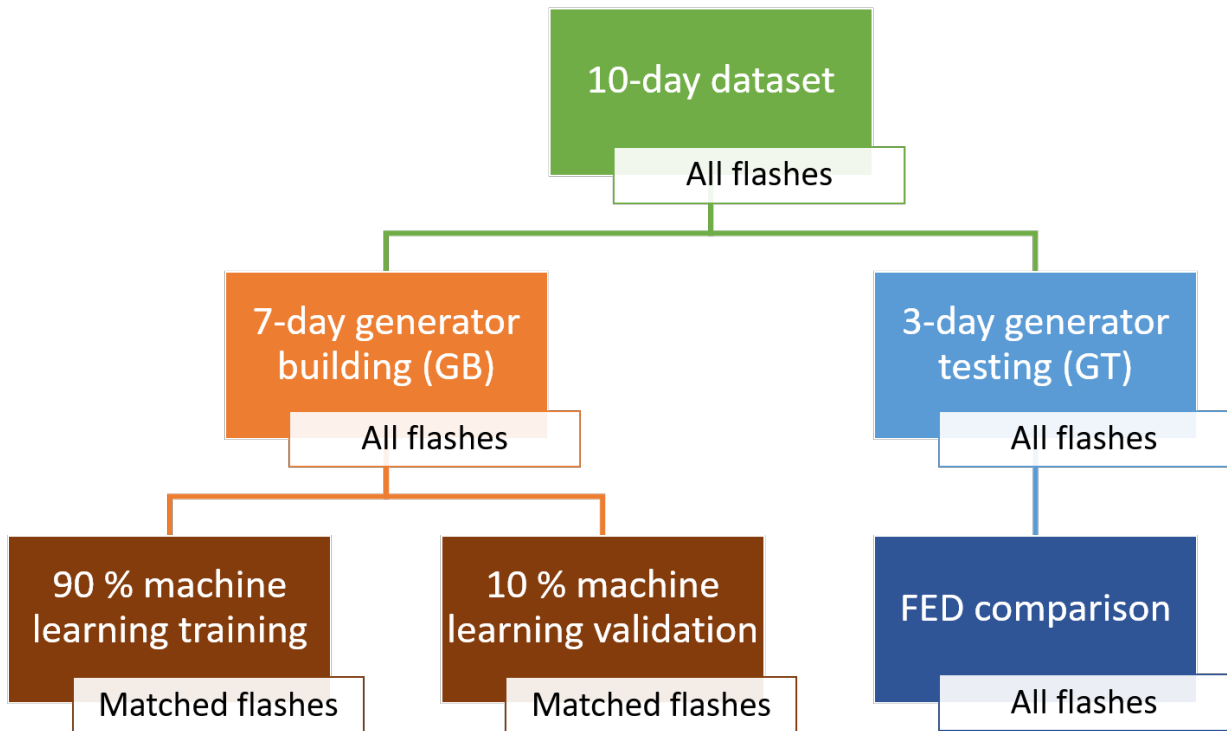


1599	<b>LIST OF FIGURES</b>	
1600	<b>Fig. 1.</b>	Relative flash detection efficiency per $1^\circ \times 1^\circ$ pixel (color) for the full 10-day dataset. In (a) for GLM and in (b) for NLDN. Grey-scale lines contour the flash number at the 0th (1 flash), 50th, 80th, and 95th percentile of the flash number distribution per $0.25^\circ \times 0.25^\circ$ pixel (only for pixels with flash activity). The corresponding percentile values are stated on the right-side color bar. . . . . 82
1601		
1602		
1603		
1604		
1605		
1606	<b>Fig. 2.</b>	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. . . . . 83
1607		
1608		
1609		
1610	<b>Fig. 3.</b>	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 <i>num ext(a)</i> configuration (Appendix B.b). . . . . 84
1611		
1612		
1613		
1614	<b>Fig. 4.</b>	An example of one simulated flash with corresponding GLM and NLDN observations on 26 May 2018. The final GEO lightning pseudo-observation generator is used including a linear SVR model, i.e., <i>linSVR num ext(a) plus</i> . Time series of latitudes (a), longitudes (b), and a map (c). The map (c) includes characteristics of the observed and simulated GLM flash. The time interval shown matches the simulated flash duration of 640 ms. . . . . 85
1615		
1616		
1617		
1618		
1619		
1620		
1621	<b>Fig. 5.</b>	Normalized absolute difference of statistics and scores (titles) between distributions of observed and simulated GLM flash extent (0 means equal to observation, 1 represents the worst simulation). The boxplots represent the distributions of 28 target generator results per ML type (x-axis) including the Inter-Quartile-Range (IQR, blue box), 1.5 times the IQR (whiskers), and outliers (black cross). The horizontal green line give the median. Results for the full test dataset. ML type abbreviations provided in Table B1. . . . . 86
1622		
1623		
1624		
1625		
1626		
1627		
1628		
1629	<b>Fig. 6.</b>	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. . . . . 87
1630		
1631		
1632		
1633		
1634		
1635		
1636	<b>Fig. 7.</b>	Observed (a) and simulated hourly FED using <i>linSVR num ext(a) plus</i> (b), <i>MLP alpha8 num raw</i> (c), and <i>BAGR KNN dist num plus</i> (d) generator on 26 May 2018, 20:00 to 21:00 UTC. The FED grid uses pixels of $5 \text{ km} \times 5 \text{ km}$ . ML type abbreviations provided in Table B1. . . . . 88
1637		
1638		
1639		
1640		
1641	<b>Fig. 8.</b>	Pixel-to-pixel ( $5 \text{ km} \times 5 \text{ km}$ ) simulated versus observed hourly FED for the 3-day test period using the same <i>linSVR</i> (a), <i>MLP</i> (b), and <i>BAGR KNN dist</i> (c) based generators as in Figure 7. The gray box
1642		
1643		

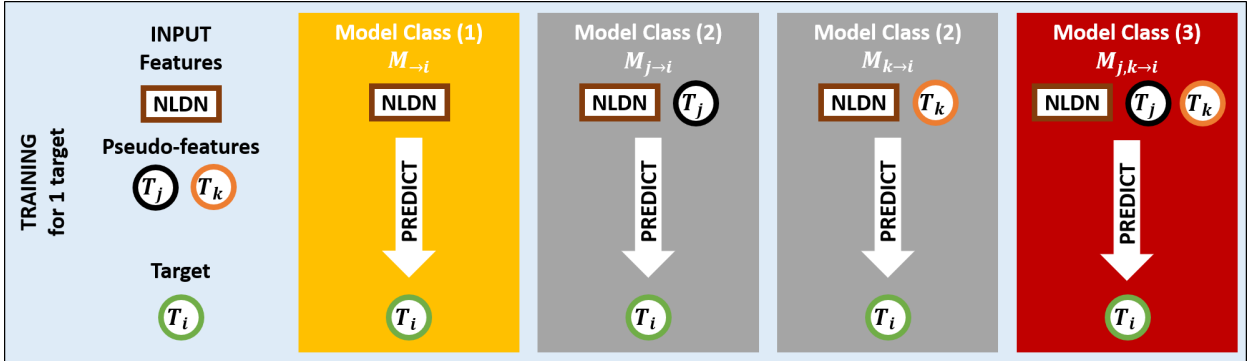
1644		and white margins indicate the upper limits of distributions on each	
1645		axis. ML type abbreviations provided in Table B1. . . . .	89
1646	<b>Fig. 9.</b>	Hourly sum of FED (a) and hourly electrified area (b) within the	
1647		region of interest. Top (1): Absolute values and number of simu-	
1648		lated flashes per hour. Bottom (2): Difference of simulation minus	
1649		observation. The observation is plotted in blue, the remaining col-	
1650		ors represent the 10 best generators for FED sum (a) and electrified	
1651		area (b), respectively. The best generator of FED sum (a) is also	
1652		included in (b) [orange], and the best generator from (b) is included	
1653		in (a) [lime]. Results for 31 July 2018. Details about the generator	
1654		names are provided in Appendix B.b. . . . .	90
1655	<b>Fig. 10.</b>	As Figure 9 with 10 repetitions of the recommended <i>linSVR num</i>	
1656		<i>ext(a) plus</i> generator. Median (line) and range (shaded) of 10 gen-	
1657		erator repetitions for 31 July 2018. . . . .	91
1658	<b>Fig. D1.</b>	Normalized absolute difference of statistics and scores (titles) be-	
1659		tween distributions of observed and simulated GLM flash duration	
1660		(0 means equal to observation, 1 represents the worst simulation).	
1661		The boxplots represent the distributions of 28 target generator re-	
1662		sults per ML type (x-axis) including the Inter-Quartile-Range (IQR,	
1663		blue box), 1.5 times the IQR (whiskers), and outliers (black cross).	
1664		The horizontal green line give the median. Results for the full test	
1665		dataset. ML type abbreviations provided in Table B1. . . . .	92
1666	<b>Fig. D2.</b>	As Figure D1 but for the normalized absolute difference of statistics	
1667		and scores (titles) between distributions of observed and simulated	
1668		GLM event number per flash (0 means equal to observation, 1 rep-	
1669		resents the worst simulation). . . . .	93



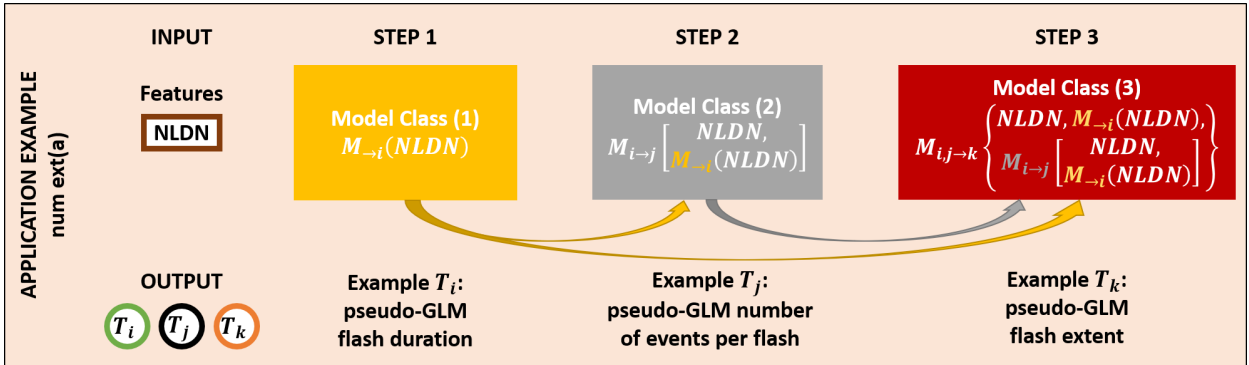
1670 FIG. 1. Relative flash detection efficiency per  $1^\circ \times 1^\circ$  pixel (color) for the full 10-day dataset. In  
 1671 (a) for GLM and in (b) for NLDN. Grey-scale lines contour the flash number at the 0th (1 flash),  
 1672 50th, 80th, and 95th percentile of the flash number distribution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



1675 FIG. 2. Illustration of splitting the 10-day dataset with GLM and NLDN flashes in 7-day gener-  
 1676 ator building (GB) and 3-day generator testing (GT) data. The GB data is further processed for  
 1677 the machine learning (ML) part.

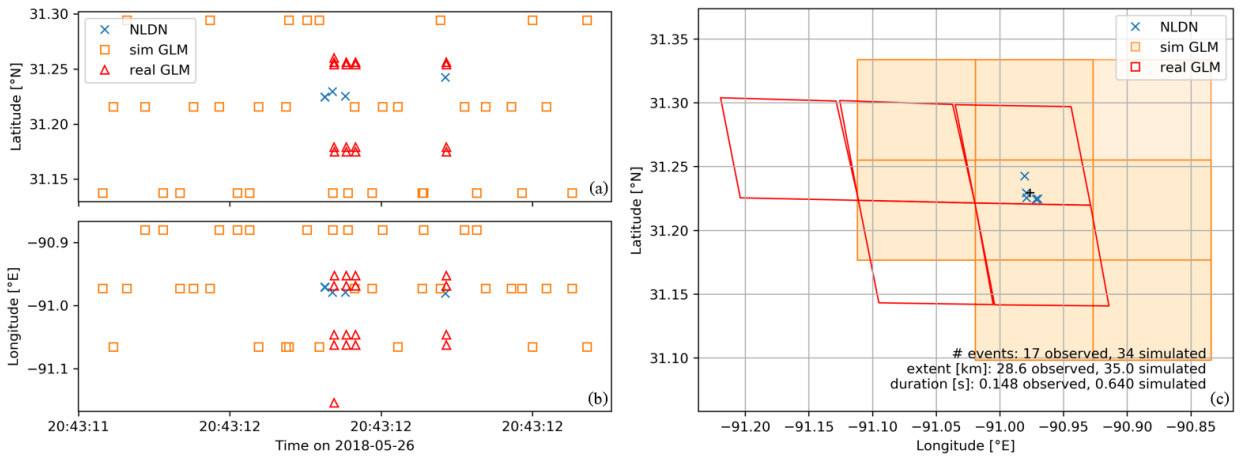


(a)

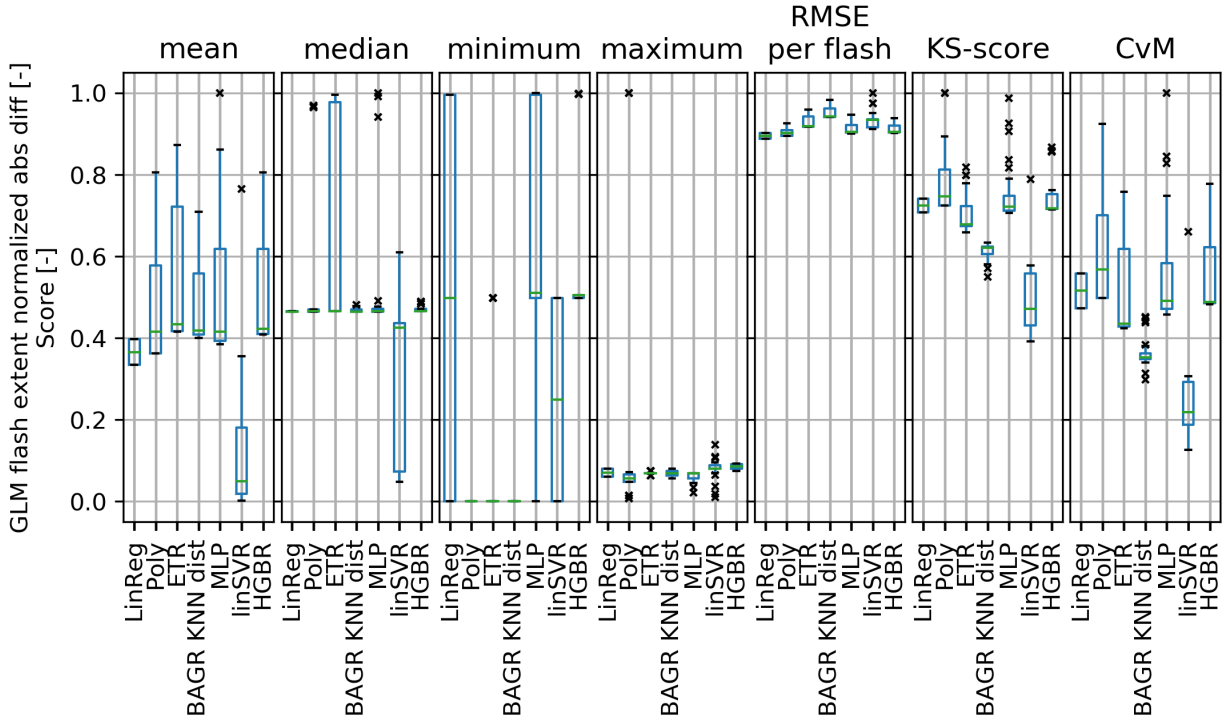


(b)

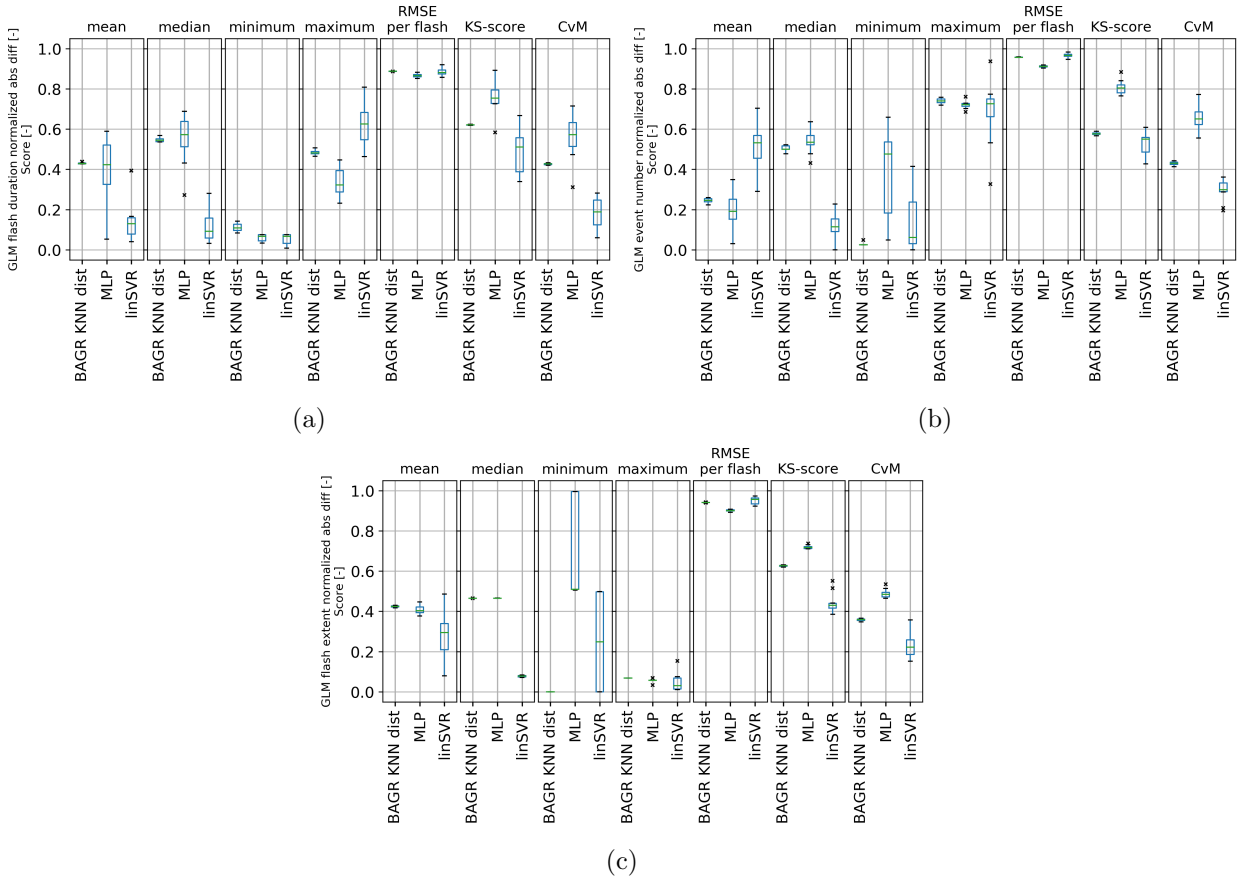
1678 FIG. 3. Flow chart of the multi-step approach illustrating the possible predictions of a given  
 1679 target using different combinations of features and pseudo-features (a, Training). The Application  
 1680 (b) shows the example of the *num ext(a)* configuration (Appendix B.b).



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.

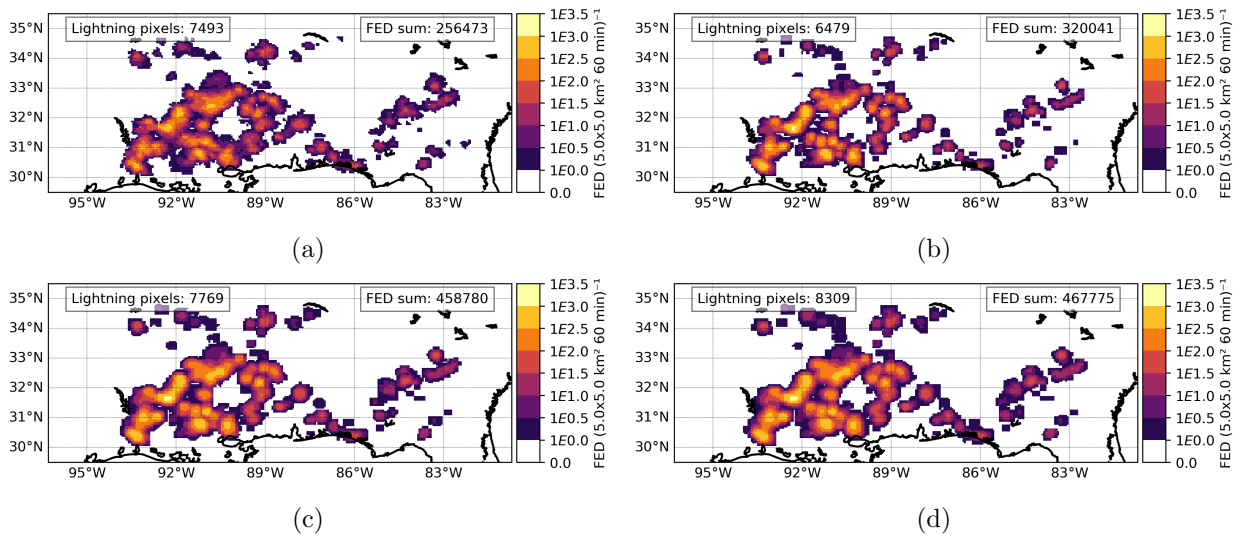


1686 FIG. 5. Normalized absolute difference of statistics and scores (titles) between distributions of  
 1687 observed and simulated GLM flash extent (0 means equal to observation, 1 represents the worst  
 1688 simulation). The boxplots represent the distributions of 28 target generator results per 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.

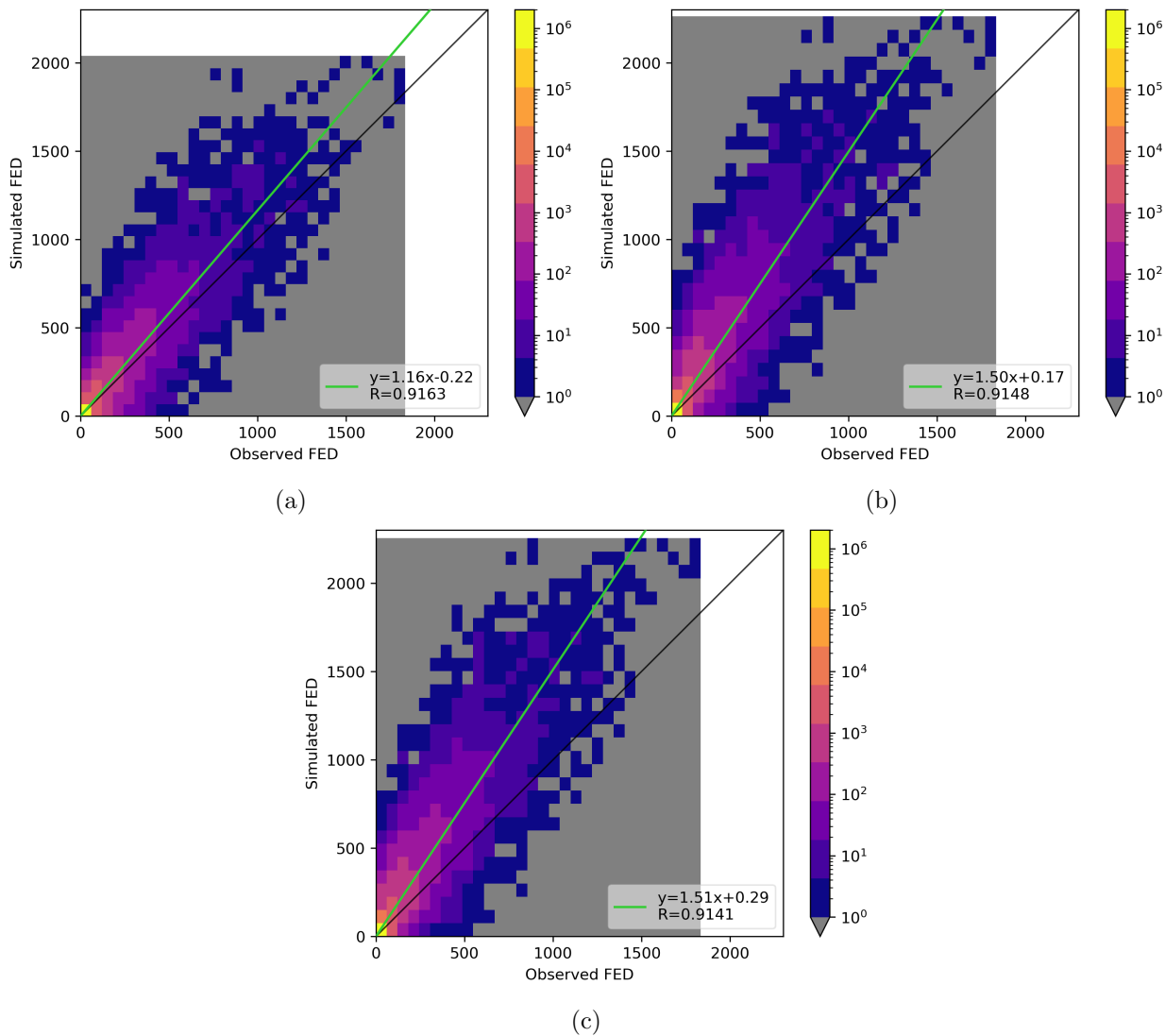


1692 FIG. 6. Normalized absolute difference of statistics and scores (titles) between distributions of  
 1693 observed and simulated GLM flash duration (a), event number per flash (b), and flash extent (c);  
 1694 y-range (0 to 1) as in Figures D1 (for a), D2 (for b), and 5 (for c). Boxplots (as in 5) represent the  
 1695 distribution for training the same model (x-axis) 10 times during the first step of the simulation.  
 1696 ML type abbreviations provided in Table B1.

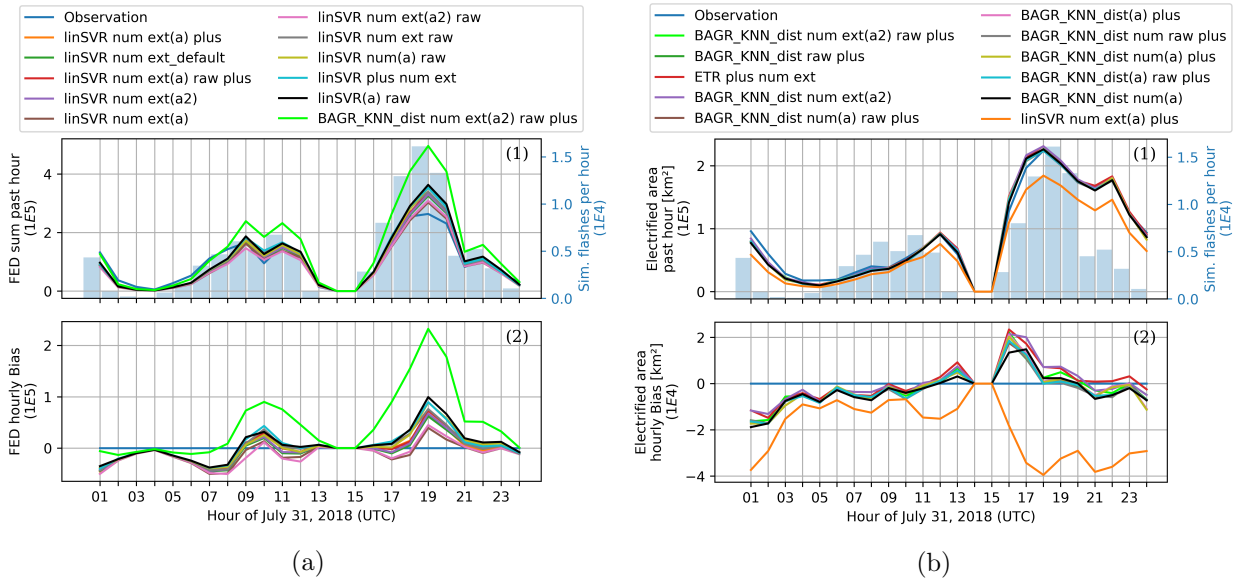




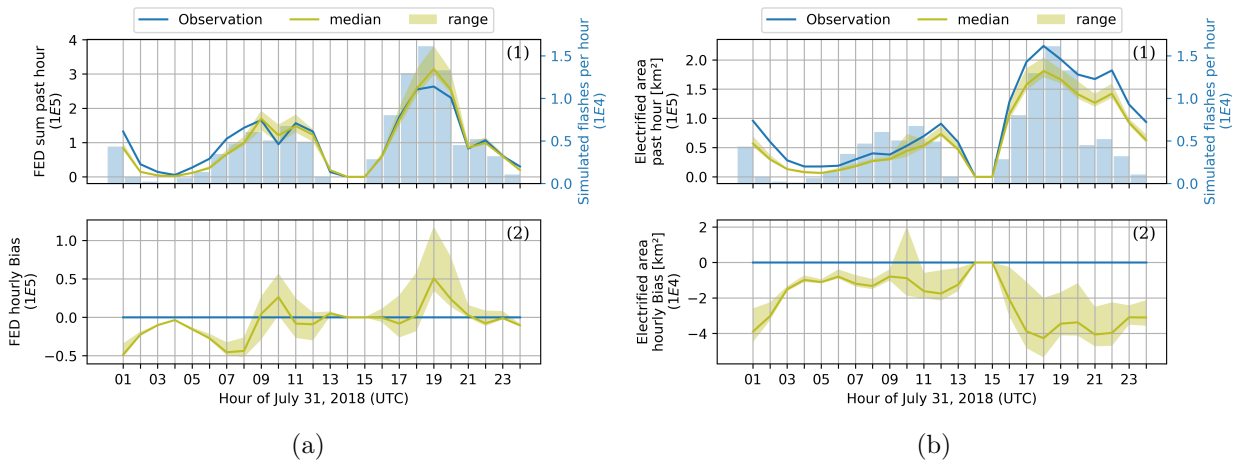
1697 FIG. 7. Observed (a) and simulated hourly FED using *linSVR num ext(a) plus* (b), *MLP alpha8*  
 1698 *num raw* (c), and *BAGR KNN dist num plus* (d) generator on 26 May 2018, 20:00 to 21:00 UTC.  
 1699 The FED grid uses pixels of 5 km × 5 km. ML type abbreviations provided in Table B1.



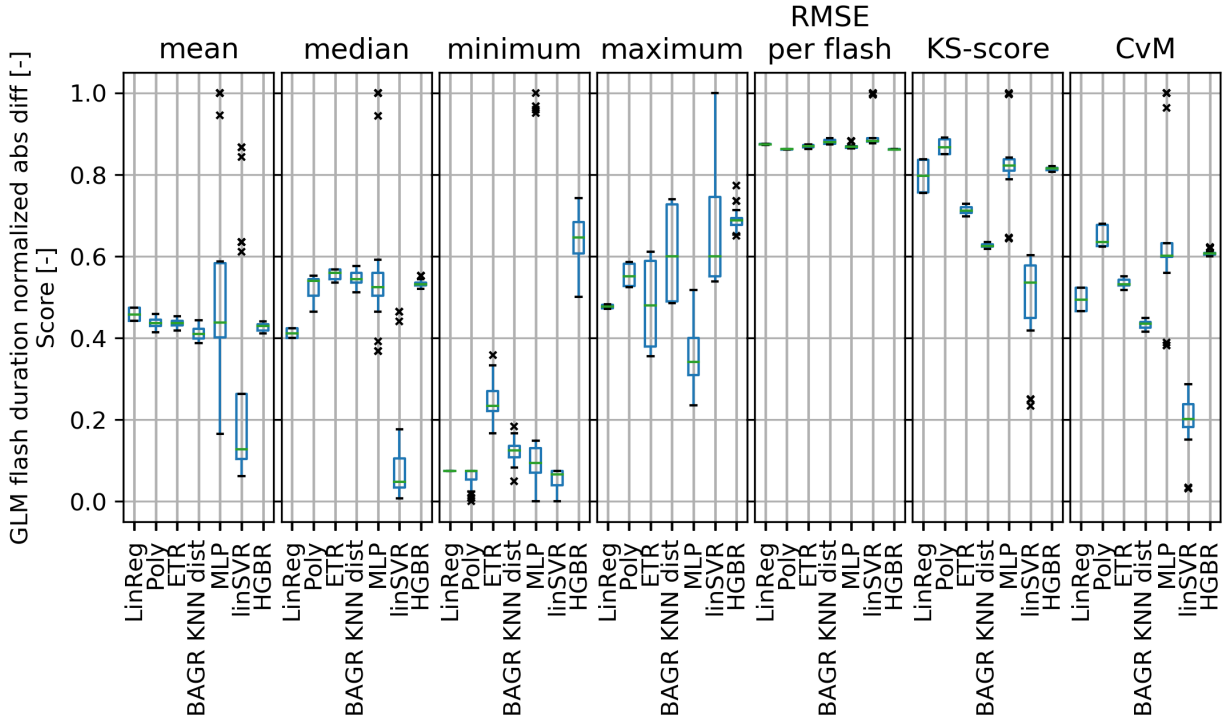
1700 FIG. 8. Pixel-to-pixel ( $5\text{ km} \times 5\text{ km}$ ) simulated versus observed hourly FED for the 3-day test  
 1701 period using the same *linSVR* (a), *MLP* (b), and *BAGR KNN dist* (c) based generators as in  
 1702 Figure 7. The gray box and white margins indicate the upper limits of distributions on each axis.  
 1703 ML type abbreviations provided in Table B1.



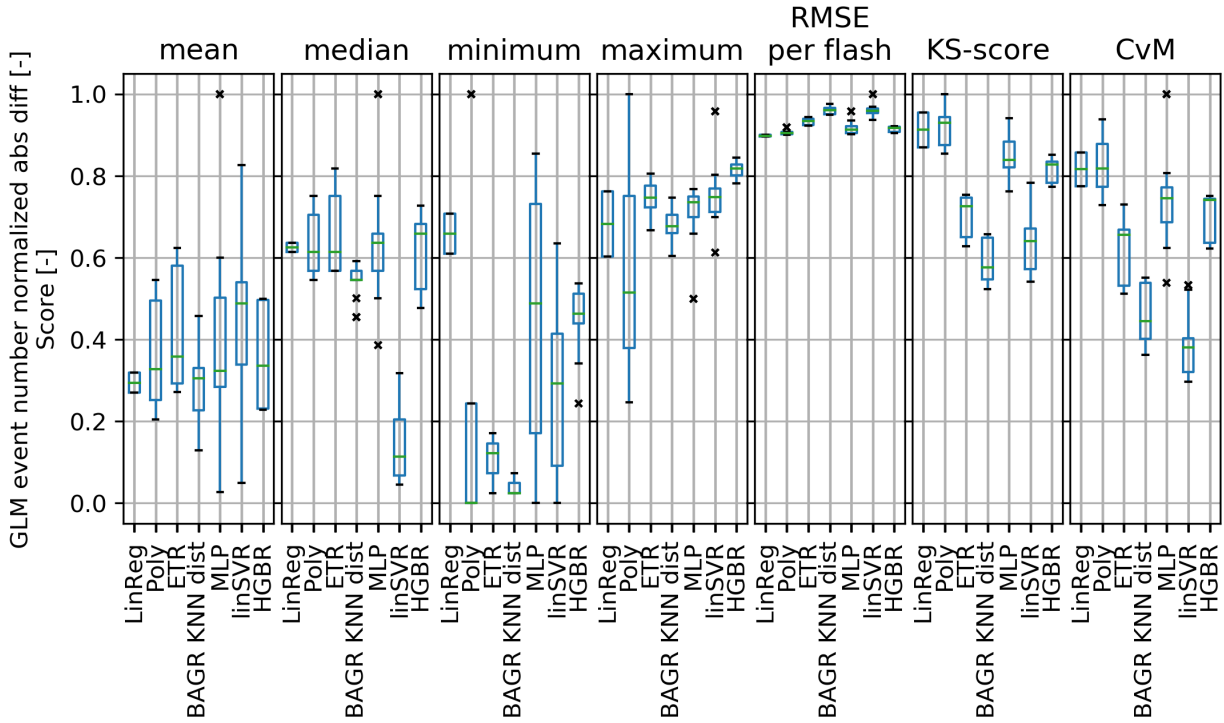
1704 FIG. 9. Hourly sum of FED (a) and hourly electrified area (b) within the region of interest.  
 1705 Top (1): Absolute values and number of simulated flashes per hour. Bottom (2): Difference of  
 1706 simulation minus observation. The observation is plotted in blue, the remaining colors represent  
 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.



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.



1718 FIG. D2. As Figure D1 but for the normalized absolute difference of statistics and scores (titles)  
 1719 between distributions of observed and simulated GLM event number per flash (0 means equal to  
 1720 observation, 1 represents the worst simulation).