

Horizontal Grid Spacing Comparison Among Random Forest Algorithms to Nowcast Cloud-to-Ground Lightning Occurrence

ALICE LA FATA (✉ alice.lafata@edu.unige.it)

University of Genoa Polytechnic School: Universita degli Studi di Genova Scuola Politecnica

<https://orcid.org/0000-0003-3730-2566>

Federico Amato

UNIL: Universite de Lausanne

Marina Bernardi

CESI SpA

Mirko D'Andrea

CIMA Foundation: Fondazione CIMA

Renato Procopio

University of Genoa: Universita degli Studi di Genova

Elisabetta Fiori

CIMA Foundation: Fondazione CIMA

Research Article

Keywords: Lightning, Machine Learning, Nowcasting, Random Forest, Disaster Risk Management

Posted Date: December 28th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1138938/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Horizontal grid spacing comparison among Random Forest algorithms to nowcast Cloud-to-Ground lightning occurrence

Alice La Fata
DITEN

*University of Genoa
Genoa, Italy*
alice.lafata@edu.unige.it

Federico Amato
IDYST

*University of Lausanne
Lausanne, Switzerland*
federico.amato@unil.ch

Marina Bernardi
CESI – s.p.a.

Milan, Italy

marina.bernardi@cesi.it

Mirko D'Andrea
*CIMA Foundation
Savona, Italy*

mirko.dandrea@cimafoundation.org

Renato Procopio
DITEN

*University of Genoa,
Genoa, Italy*
renato.procopio@unige.it

Elisabetta Fiori
*CIMA Foundation
Savona, Italy*

elisabetta.fiori@cimafoundation.org

Abstract — This paper discusses the use of Random Forest (RF), a popular Machine Learning (ML) algorithm, to perform spatially explicit nowcasting of cloud-to-ground lightning occurrence. An application to the Italian territory and the surrounding seas is presented. Specifically, 1-hour ahead lightning occurrences over the months of August, September and October from 2017 to 2019 have been modelled using a dataset including geo-environmental features. Results obtained with three different spatial resolutions have been compared, for nowcasting both positive and negative strokes. The features' importance resulting from the best RF models showed how data-driven models are able to identify the relationships between meteorological variables, in agreement with previous physically based knowledge of the phenomenon. The encouraging results obtained in terms of forecasting accuracy support the idea to use ML-based algorithms in early warning procedures for disaster risk management.

Keywords: Lightning, Machine Learning, Nowcasting, Random Forest, Disaster Risk Management

I. INTRODUCTION

Lightning is an electric discharge characterizing both mesoscale and microscale events that exhibit sudden evolution and complex interaction with the surrounding atmosphere. The strong correlation between lightning and extreme events, such as hail, tornadoes, and heavy rainfalls has been widely documented (Tapia et al., 1998) (Adamo et al., 2009) (Schultz et al., 2015) (Schultz et al., 2017) (Lagasio et al., 2017) (Lynn, 2017). Moreover, Cloud-to-Ground (CG) lightning is a source of damages to infrastructures (such as wind turbines and transmission/distribution lines and buried cables), loss of life and forest fires (Cooper & Holle, 2019). Because of these concerns, having reliable forecasts of the lightning expected activity is of interest in many fields. In particular, a timely forecast of the lightning activity may support early warning systems giving decision makers updated information to take the necessary safety measures. However, nowcasting activities related to the lightning phenomenon is still a great challenge and, considering the complicated interaction between in-cloud and many atmospheric processes, a wide range of approaches is available for such purposes. Many studies have implemented lightning forecasting models based either on the electrification in-cloud processes causing lightning or on

atmospheric variables linked with the lightning phenomenon, such as Convective Available Potential Energy (CAPE) and precipitation rate (McCaull Jr et al., 2009) (Lynn et al., 2012) (Fierro et al., 2014) (Tippett & Koshak, 2018) (McCaull Jr et al., 2020). Early approaches to lightning forecasting were based on analytical studies relating storm lightning rates to convective cloud top height (Price & Rind, 1992). Later, other researchers based forecast methods on the statistical use of lightning climatologies (Bothwell, 2005), i.e., an approach incorporating lightning climatology (e.g., thunderstorm activity, peaks in diurnal CG lightning activity, etc.) and predictor fields from Numerical Weather Prediction (NWP) models. Other researchers used methods based on measures of predicted buoyant instability aloft, as derived from numerical simulations (Bright et al., 2005), i.e., methods helping in delineating potential thunderstorm areas by determining if instability and appropriate thermodynamics for charge separation are coincident in observations and model forecasts. Moreover, in recent decades, elaborated full electrification schemes were developed for inclusion in explicit-convection numerical forecast models, e.g., (Mansell et al., 2002), (Fierro et al., 2007), (Lynn, 2017). These latter methods allow a detailed insight into storm electrical behavior and provide forecasts of the Flash Rate Density (FRD) and flash locations, considering the whole CG and intra-cloud (IC) activity. However, even after simplification of the complex lightning discharge process, most full electrification schemes remain computationally intensive and are still subject to errors in their quantitative forecasts of lightning event flash rates, owing to the intrinsic low predictability of deep convection in the parent explicit-convection model (McCaull Jr et al., 2020). Consequently, the accuracy of FRD results obtained up to now varies on a day-to-day basis, owing to the limitations of model physical parameterizations, input data and procedures, and model numerics. All of these factors produce inherent uncertainty in model forecasts (McCaull Jr et al., 2020). The increasing computational capability in the last ten years have supported researchers to investigate the sensitivity of NWP models to horizontal grid spacing variability, evaluating how to properly calibrate and interpret ensemble output and to optimize trade-offs between model resolution and other computationally constrained parameters (Kain et al., 2008)

(Fiori et al., 2010) (Bryan & Morrison, 2012) (VandenBerg et al., 2014) (Potvin & Flora, 2015).

In recent decades, Machine Learning (ML) helped improving the prediction skills of multi-data weather related phenomena thanks to the integration of its strengths with atmospheric science. The ability of ML algorithms of modelling highly nonlinear functions is fundamental for its application in the analysis of the spatial and temporal variability of geo-environmental data (Kanevski et al., 2009), sometimes allowing uncertainty estimation (James et al., 2013) (Guignard et al., 2021). Moreover, ML tools can take advantage of large datasets and of the use of numerous input variables, or features (James et al., 2013). Therefore, they can solve regression and classification problems in high-dimensional (geo)input spaces, generally constituted by the geographical space and a set of spatially, temporally, or spatio-temporally referenced features. Different ML algorithms have been successfully applied to model phenomena such as, among the others, water pollution (Leuenberger & Kanevski, 2015), landslides (Taalab et al., 2018) and forest fires (Tonini et al., 2020), susceptibility, air temperature (Amato et al., 2020), wind speed (Veronesi et al., 2016). In recent years, some ML algorithms have been developed also in the domain of lightning occurrence forecasting. Mostajabi et. al. (Mostajabi et al., 2019) used XGBoost to perform 30 minutes ahead forecasting of lightning occurrence based on a set of single-site observations of meteorological parameters. (Blouin et al., 2016) used a tree-based classification algorithm to predict 6-hours and 24-hours CG lightning. A 1-hour nowcasting model is proposed in (Mecikalski et al., 2015), showing that lightning forecast are made 45 minutes before rainfall occurs. Differently, (Zhou et al., 2020) developed a 0–1-hour nowcasting model based on data from geostationary meteorological satellites as input for a Deep Learning (DL) algorithm. Since satellite data have the advantage of monitoring and detecting the initial stages of convective clouds, they can be used for convective initiation (CI) monitoring and early warnings, when the detection over one spot is available.

In this paper, the use of ML to perform spatially explicit nowcasting of CG lightning occurrence is presented. Precisely, the goal of the work is to create a model, able to perform pattern recognition for nowcasting the CG occurrence, one hour in advance over the Area Of Interest (AOI), here covering the Italian territory and the surrounding seas. Specifically, a 1-hour ahead forecasting is performed by classifying each pixel of the study area with the presence or absence of CG strokes. The classification is performed with Random Forest (RF). Since the data used as input for the algorithm have different spatial resolution, an appropriate grid of interpolation may be sought. A comparison among results obtained using 3 different horizontal grids for data interpolation is presented to test the sensitivity of the algorithm in detecting the processes driving to the development of CG lightning. Summarizing, the primary goal of the nowcasting model proposed is to test the ability of ML to nowcast the possibility of CG strokes' occurrence and, secondly, to compare outcomes' accuracies depending on the horizontal grid of the input data. The comparison is performed over 3 months of 3 years (i.e., August, September and October from 2017 to 2019).

The paper is organized as follows: Section II presents the methodology used, discussing input data, the horizontal space

resolutions compared, and the classification algorithm utilized, Section III shows results and proposes a related discussion, Section IV provides the conclusion of the paper.

II. METHODOLOGY

A. Input Data

Aiming at enhancing results obtained in (La Fata et al., 2021), the models proposed in this study are trained with meteorological data covering the period from August to October of 2017, 2018 and tested with data related to the same months in 2019. This choice is due to the particularly high lightning activity over the Italian peninsula and the surrounding seas during August-October 2018, as shown in (Nicora et al., 2021) (Paliaga et al., 2019). At the intense lightning activity occurred in this period corresponded also a significant amount of precipitation, as shown in Nicora et al. (2021) and the fact that August 2018 was the fourth warmest after 1880, 2016, 2017, 2015 at European scale, (Paliaga et al., 2019). The AOI has been chosen to have a reasonably comparable number of pixels over land and sea, as shown in Figure 1. Pixels over which the AOI has been analyzed include a sea area of $A_{\text{sea}}=448.734 \text{ km}^2$ and a land area of $A_{\text{land}} = 301.338 \text{ km}^2$, thus A_{sea} is ~1.5 times A_{land} . Data have been interpolated on 3 different spatial grids covering the AOI. All features of each pixel are organized in a Data Frame (DF) and only pixels containing all data are considered to train the models, i.e., if one feature is missing in a pixel, the pixel is excluded from the DF. Considering that the final goal of the work is to create an operative tool working in near real time, in this study the selection of the input variables is performed considering their availability, the spatio-temporal resolution and the easiness in the retrieve and process phases, i.e., the timings needed to receive, download, and process the data. Hence, all data needed to nowcast the lightning activity in the models proposed (from either observations/direct measurements or radar/satellites measurements) must be available sufficiently in advance, giving analysts enough time to pre-process them and run the forecasting algorithm.

The features chosen to train the model range from spatio-temporal information to observations or forecasted data produced by NWP, i.e.,

- Latitude and Longitude: coordinates belong to the spatial matrix on which all the features are gridded. Outcomes of the 10 years analyses in (Nicora et al., 2021) show a significant difference in the lightning activity along the Italian peninsula. Moreover, the correlation between lightning activity and the latitude is also exposed in (Underwood, 2006) (Rakov, 2013) (Enno et al., 2020).
- Digital Elevation Model (DEM [m ASL]): analyses performed in (Paliaga et al., 2019) suggest that the orographic effect may be considered as possible influencing factor enhancing the seasonal difference between the lightning activity over sea and land. Moreover, in (Underwood, 2006) (Mazarakis et al., 2008) (Vogt & Hodanish, 2014) the relative flash density was found to be correlated with terrain elevation and also (Poelman, 2014) associates the CG lightning peak currents with the terrain elevation. Nevertheless (Kotroni & Lagouvardos, 2008) states that a positive correlation between lightning activity and terrain elevation is evident

during spring and summer but not in autumn and winter. Consequently, aiming at creating a generalized ML algorithm, the terrain elevation [m ASL] is added as input feature.

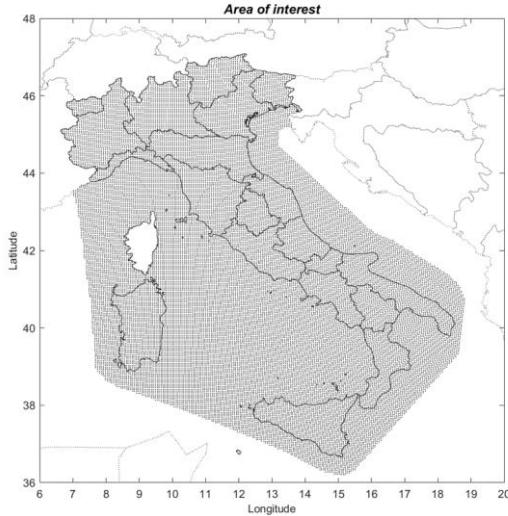


Figure 1: AOI for algorithm application.

- Temperature: a positive correlation between lightning and Sea Surface Temperature (SST) in autumn is shown in (Kotroni & Lagouvardos, 2016), where it is suggested that the reason may lie in the fact that higher SST destabilizes the lower tropospheric layers, thus enhancing convection and therefore lightning. Analyses in (Kotroni & Lagouvardos, 2016) also suggest that this finding could be used to forecast the intensity of lightning activity. Similar considerations can be found in (Nicora et al., 2021), in which, during autumn, a strong lightning and precipitation activity is detected and linked with the strong interaction between a warmer SST and a significant amount of unstable moist air. Moreover, the lightning activity has been linked with the solar heating cycle in (Enno et al., 2020); indeed, a summer peak in the lightning activity was observed at mid-latitudes whereas the Mediterranean experienced an autumn maximum. Most of the lightning occurred over land from March to August, whereas from September to February it was concentrated over the Mediterranean. Consequently, the Infrared brightness temperature [K] measured by Meteosat Second Generation in the infrared channel (10.8MHz) is used as input parameter for the creation of the ML algorithm. The refresh time is 15 minutes. Since all other features in the input space have a time resolution of 1 hour, the hourly temperature mean value and hourly standard deviation are computed.
- Zonal u (or x-coordinate) and meridional v (or y-coordinate) components of horizontal wind vector [ms^{-1}]: The microphysical and kinematic characteristics of the relationship between lightning and convective processes have been extensively studied. Particularly, because lightning needs an electric field region to be initiated, flash initiations tend to cluster in the vicinity of updraft cores, where either sedimentation alone or sedimentation combined

with wind shear or turbulence produces gradients in the charged particles, creating the electric fields needed to initiate lightning (Calhoun et al., 2013). Consequently, both zonal and meridional wind components are used in the DF at 4 pressure levels (1000 hPa, 850 hPa, 700 hPa, 500 hPa) to consider the wind shear effect. The components are provided by the daily forecast of the COSMO-I5 model over the period under investigation. COSMO-I5 (Steppeler et al., 2003) is the limited-area, non-hydrostatic model used by the COSMO consortium with 2 daily run (00UTC-12UTC) and 72 hours of forecast over the Mediterranean basin.

- Vorticity: many studies demonstrate that deep atmospheric convective processes are characterized by intense vertical velocities, able to reach the zone of the atmosphere where lightning phenomena occurs (Petersen et al., 1999) (Deierling & Petersen, 2008) (Wang et al., 2015) (Huang, 2021). Findings in (Mazarakis et al., 2008) confirm such idea. The data availability for this analysis included only the zonal and meridional components of horizontal wind vector, thus, the vertical component of vorticity, ω_z , has been calculated at 1000 hPa to support the ML algorithm in identifying the most intense convection zones at lower level of the atmosphere, i.e.,

$$\vec{\omega}_z = \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) \quad (1)$$

where the spatial derivative has been approximated with first order Finite Difference.

- Precipitation: Many studies (Adamo et al., 2009)(Tapia et al., 1998) (Soula & Chauzy, 2001) (Adamo et al., 2009) (Lagasio et al., 2017) (Soula & Chauzy, 2001) have revealed the strong intercorrelation between the lightning phenomena and severe rainfall process evolution in thunderstorms, confirming the hypothesis that lightning activity may be useful to track the convective cores' motion associated with severe rainfall processes. Consequently, measured precipitation data are included as input feature in the DF to train the models. From the observational point of view, the 1-hour precipitation accumulation from the Integrated Multi-satellitE Retrievals for GPM (IMERG) is used. Moreover, to validate data retrieved from COSMO-I5 model, the precipitation forecasted by COSMO-I5 is compared with precipitation values detected by the Italian radar network, obtained with the Modified Conditional merging (MCM) technique. Precipitation data computed by COSMO-I5 model are considered to be validated if:

$$|rr_{\text{measured}} - rr_{\text{COSMO}}| \leq T \quad (2)$$

where rr_{measured} is the precipitation from radar, rr_{COSMO} represents the precipitation modelled by COSMO-I5 and T is a threshold, defined for 4 different hourly cumulated rainfall levels, as shown in Table 1.

TABLE 1. THRESHOLDS USED TO VALIDATE THE COSMO I5 MODEL.

Measured precipitation [mm]	Threshold T [mm]
$0 \leq rr_{measured} \leq 15$	10
$15 < rr_{measured} \leq 35$	20
$35 < rr_{measured} \leq 55$	25
$rr_{measured} > 55$	30

- Distance Sea/land: analyses in (Nicora et al., 2021) (Paliaga et al., 2019) (Enno et al., 2020) found that the lightning activity may differ over sea and land areas. Moreover results in (La Fata et al., 2021) suggest that longitude may have a relevant importance when trying to perform a classification of pixels with or without CG strokes. Consequently, the distance from each land pixel from the sea has been added as feature in the input space.

All the above-mentioned input data have different spatial and temporal resolution; thus, interpolation is needed to create a DF to train and test the ML algorithm. Ideally, to introduce the fewest approximation, the influence of interpolation errors needs to be as reduced as possible. Considering the data availability, their different temporal and horizontal spatial resolution and since it is unknown a priori what's the best resolution on which interpolate all the data, the 3 spatial grids available are compared:

- Horizontal grid of Temperature data (HT): $0.05^\circ \times 0.0611^\circ$,
- Horizontal grid of Wind components data (HW): average pixel dimension of $0.045^\circ \times 0.0681^\circ$,
- Horizontal grid of Precipitation data (HP): $0.1^\circ \times 0.1^\circ$.

The temporal resolution is set to 1 hour for all input data.

The output of each model created is a Boolean variable indicating the presence or absence of strokes in a specific spatial location in the following hour. For all the 3 spatial resolutions over which input data have been gridded, 2 models have been created, one aiming at forecasting pixels in which positive strokes are detected, one aiming at forecasting pixels in which negative strokes are detected. The reason of the creation of differentiated models for positive and negative strokes lays in the different numerosity and creation mechanism of them. (Diendorfer et al., 2009) and (Cooray & Arevalo, 2017) say that negative lightning flashes account for about 90% or more of global CG lightning, and that 10% or less of CG discharges transport positive charge to Earth. Moreover (Rakov, 2013) and (Diendorfer et al., 2009) state that positive lightning can be the dominant type of CG lightning during the cold season, during the dissipating stage of a thunderstorm, and in some other situations including severe storms and thunderclouds formed over forest fires or contaminated by smoke. Differences in the scenarios leading to positive and negative lightning can be found in (Nag & Rakov, 2012) and (Cooray & Arevalo, 2017). (Nag & Rakov, 2012) also states that several properties of positive lightning (e.g., number of strokes per flash, occurrence of continuing current, leader propagation mode, and branching) appear to be distinctly different from those of negative lightning.

Ground Stroke Density (GSD) data are available thanks to CESI-SIRF Lightning Location System (LLS), now property of Meteorage SAS. Sensors of CESI-SIRF LLS were installed in 1994 and consist today of VAISALA LS7002 sensors over the Italian territory. SIRF is a founding member

of the EUCLID (EUropean Cooperation for Lightning Detection) network, a pan European union aiming at sharing LLSs' data. All sensors of EUCLID operate in the same frequency range. The sensors' redundancy of the EUCLID network allows the Italian territory to be covered by theoretically 150 sensors; actually at least by the 15 sensors closest to the national borders. Strokes data have been processed defining a Boolean matrix whose entries are 1 in correspondence of pixels in which at least one stroke is detected (class 1) and 0 elsewhere (class 0). The DF over the AOI, for each hour, is composed of pixels including the 16 abovementioned features and the 17th dimension representing the target, i.e., the presence/absence of strokes.

B. Classification algorithm

The problem of determining the presence or absence of strokes in a specific spatial location during the next hour has been formulated in terms of classification of each pixel of the AOI. The classification has been performed using RF. The choice is due to the demonstrated potential and benefits in the prediction ability of the RF technique for nowcasting problems strictly correlated with the lightning phenomenon, such as the prediction of small-scale storm initiation, diagnosis of turbulence, mesoscale convective system initiation and lightning activity, respectively shown in (Breiman, 2001) (Williams, 2014) (Blouin et al., 2016) (Ahijevych et al., 2016). RF is an ensemble method based on Classification and Regression Decision Trees, originally introduced in (Breiman, 2001). A classification tree is an algorithm in which the input space is divided into L non-overlapping regions R_1, R_2, \dots, R_L . A recursive binary splitting starting from the top of the tree is used to formulate the R_L regions: on the top of the tree all the observations are still included in one single region and, successively, each region of the predictor space is split into two new ones. Specifically, calling s the cut point and for any $l \in L$, the pair of half-plains produced at each step is defined as:

$$R_1(l, s) = \{X | X_l < s\} \quad \text{and} \quad R_2(l, s) = \{X | X_l \geq s\} \quad (3)$$

where $\{X | X_l\}$ is the conditional probability related to X given X_l , i.e., a measure of the probability of occurrence of the event X given that another event (X_l) has already occurred. Once the R_L regions are defined, predictions are simply computed by assigning to each sample the label of the most commonly occurring class of training observation in its same region. At each step of the binary splitting, the values of s and l are obtained by minimizing the Gini index G , computed as:

$$G = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk}) \quad (4)$$

where \hat{p}_{mk} is the proportion of the training observations in the m^{th} region belonging to the k^{th} class. RF is built with the ensemble of several trees generated from bootstrapped training samples. However, to ensure a proper decorrelation among the trees, only a subset m of the p features is chosen as split candidate for each single tree. An advantage of this approach is the possibility to study the importance of each feature, hence improving the understanding of the phenomenon analysed. Specifically, it is possible to average

the reduction of the Gini index due to splits over a given predictor and averaged for all the trees. Therefore, high values of this metric will correspond to more important predictors.

To apply RF over the AOI for the creation of the ML algorithm to nowcast lightning, data have been split into a train and a test set and organized as follows: pixels containing at least 1 stroke (class1) are counted; pixels containing 0 strokes (class 0) are randomly selected among pixels of the DF where the precipitation forecasted by COSMO-I5 model is validated, as indicated Table 1, so that class 0 and class 1 are balanced (i.e. contain the same number of pixels). The balancing procedure has been performed for a DF containing only locations of positive strokes and for a DF containing only locations of negative strokes. The resulting DF is composed of balanced 0-1 pixels related to August-September and October 2017 and 2018 as train set and pixels related to August-September and October 2019 as test set. Although the approach adopted to split the data considers neither the spatial nor the temporal dependencies among data, these are explicitly taken into account by the addition into the input space of the geographical coordinates (latitude and longitude). Once normalized the input features in the interval [0,1], a 5-fold Cross-Validation (CV) has been performed to train RF and determine optimal values for parameters.

III. RESULTS AND DISCUSSION

A. Results

The Confusion Matrices (CMs) of the best models obtained with the RF technique with data gridded over 3 different horizontal spatial resolutions have been calculated for all data selected for August-October 2019. CMs are composed of 4 elements: true negatives (TN, i.e., pixels belonging to class 0 correctly labelled as class 0), true positives (TP, i.e., pixels belonging to class 1 correctly labelled as class 1), false positives (FP, i.e., pixels belonging to class 0 wrongly labelled as class 1) and false negatives (FN, i.e., pixels belonging to class 1 wrongly labelled as class 0). The predicted probability threshold is set to 0.5, i.e., a pixel is labelled as belonging to class 1 if the probability associated to such pixel calculated by the model is ≥ 0.5 . To compare results obtained using HT, HW and HP, typical parameters related to lightning detection are computed. The Probability of Detection (POD), is calculated as

$$POD = \frac{TP}{TP + FN}, \quad (5)$$

the False Alarm Rate (FAR) is calculated as

$$FAR = \frac{FP}{TP + FP}, \quad (6)$$

and the Accuracy is calculated as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (7)$$

Results are summarized in Table 2 and Table 3 and shown in Figure 2. Considering POD, FAR and Accuracy, the best performing model is the one for which input data have been re-gridded using HW resolution, confirming the results

showing that higher resolutions of data allow to create more reliable and better performing models (Kain et al., 2008) (VandenBerg et al., 2014) (Potvin & Flora, 2015). Consequently, the features' importance resulting from the best RF models obtained with data re-gridded using HW is shown in Figure 3. For data gridded with HW resolution, for both Positive and Negative strokes, precipitation resulted to be the highest impacting feature, highlighting the ability of the model to link the lightning phenomenon and the meteorological variables driving to the creation of cumulonimbus and precipitation, in agreement also with outcomes of (Paliaga et al., 2019) (Enno et al., 2020) (Nicora et al., 2021). Longitude resulted to be the second most important feature. Nevertheless, the variable representing the distance from land to sea surface is the lowest impacting for Positive strokes and the second lowest impacting for Negative strokes. The importance related to wind components varies with pressure levels: for both Positive and Negative strokes the importance of wind components decreases as pressure level increases, i.e., zonal and meridional wind components at 500 hPa is the highest impacting feature among such data. Importance of temperature data is located between wind components at lower pressure levels. This may be due to the fact that temperature values are available thanks to infrared satellite measurements. Since the source of lightning is usually a cumulonimbus (thundercloud) (Rakov, 2013), it is reasonable to think that temperature values over a cloud area are linked with wind components at lower pressure levels (i.e., 500-700 hPa). The low importance related to Latitude and DEM show how the impact of meteorological features describing the atmosphere at higher altitudes is more relevant than the impact related to topology data.

B. Discussion

Accuracy values, reaching values higher than 57% and the FAR values lower than 37%, for both positive and negative strokes, highlight the ability of the models to deal with forecasts with all the three different horizontal spatial resolutions. Nevertheless, a reliable nowcasting tool should reach higher Accuracies. Thus, in the following paragraph, possible solutions are explored and further data to be added in the input space are discussed.

TABLE 2. PERCENTAGE VALUES OF POD, FAR AND ACCURACY FOR POSITIVE STROKES

RESOLUTION	Positive strokes		
	POD [%]	FAR [%]	Accuracy [%]
HP	44,11	37,03	59,08
HW	54,11	18,20	71,84
HT	36,37	32,77	59,32

TABLE 3. PERCENTAGE VALUES OF POD, FAR AND ACCURACY FOR NEGATIVE STROKES

RESOLUTION	Negative strokes		
	POD [%]	FAR [%]	Accuracy [%]
HP	39,04	37,10	58,00
HW	44,51	19,07	67,01
HT	24,93	29,69	57,20

Results of Accuracy obtained with HP and HT resolutions are approximately similar for both positive and negative strokes, while they are differentiated when calculating POD

and FAR. Despite, for negative strokes, FAR for data gridded on HT resolution is lower than FAR for data gridded on HP resolution, POD for data gridded on HP resolution is higher. At an overall look, for both positive and negative strokes, the best performing algorithm is the one with data gridded using COSMO resolution, i.e., HW. Figure 2 shows how the models with data gridded over HW reaches the highest POD and Accuracy and the lowest FAR. This result may be attributed to the fact that half of the features used to train the model are available on COSMO resolution, i.e., all zonal and meridional wind components. Thus, training the model re-gridding all data using COSMO resolution introduce interpolation errors on less than half among the variables (the vertical vorticity of wind at 1000 hPa is computed starting from wind data). Differences in the performances reached by models trained with data gridded using HW resolution support literature results indicating better results when using high spatial resolution. Distinction in the results obtained for positive and negative strokes may be also attributed to the differences in the creation mechanisms of them (Nag & Rakov, 2012) (Cooray & Arevalo, 2017).

Results reached confirm the yet high POD and low FAR obtained in (Blouin et al., 2016), in which a lower spatial resolution was used, and in (Mostajabi et al., 2019), in which the best results were obtained with an XGBoost algorithm to perform 30 minutes ahead forecasting of lightning occurrence based on a set of single-site observations of meteorological parameters. Differently, results in (Zhou et al., 2020), obtained using data from geostationary meteorological satellites as input to create a DL algorithm, reach higher performances with respect to the presented model when dealing with hours characterized by high intense lightning activity. Consequently, future enhancement of the models here presented may consider the introduction of satellite and radar data as input parameters, with a sufficient resolution (Weisman et al., 1997) (Skamarock, 2004) (Kain et al., 2006) able to detect and capture CI and its evolution.

The results presented in this paper must be intended as part of an ongoing research, and many relevant questions still must be investigated. Future studies will deal with the optimization of the spatio-temporal correlation among data. Indeed, for all the presented models, the spatial structure of the data was considered by explicitly adding the geographical coordinates in the input space. Nevertheless, the effect of neighbouring observation in both space and time was not considered in the proposed models. Embedding strategies may be investigated to better consider the spatio-temporal nature of the investigated phenomena. The use of DL algorithms may also ensure a better modelling of the spatio-temporal dependences typical of environmental data because

of their capability to automatically extract features in both spatial and temporal domains. Finally, this manuscript analysed the forecasting problem in terms of binary classification of the presence/absence of strokes, only considering CG lightning. Further variables allowing to distinguish the creation mechanisms for positive and negative lightning should be added in the input space in future studies. Consequently, future improvements of the work presented in this paper will include investigations related to embedding strategies and the definition of a further classification or regression analysis:

- one option could be the definition of a threshold related to the typical values of the GSD (as reported in (Nicora et al., 2021) and (Paliaga et al., 2019) and references therein: the classification problem gives an indication in case such threshold is overcome.
- the problem could be analysed in terms of regression having as output feature the GSD.

IV. CONCLUSION

In this paper results of an application of RF used to perform spatially explicit 1-hour ahead nowcasting of lightning occurrence have been presented. Precisely, the proposed ML models allowed the nowcasting of CG strokes occurrence one hour in advance over a defined area. An application to the Italian national territory, including its surrounding seas, has been presented. By comparing the results obtained with three different resolutions, it is evident that the first step in this proposed approach is to interpolate all the data selected as input on the finer spatial resolutions available to increase the ability of the model in nowcasting both positive and negative strokes. The features' importance resulting from the best RF models showed how data-driven models are able to identify relationships between variables, in agreement with previous physically based knowledge of the phenomenon. The encouraging results obtained in terms of forecasting Accuracy suggest how, after proper improvements, ML-based algorithms could find their place in wider early-warning systems to support disaster risk management procedures.



Figure 2: Histograms of the percentage values of POD, FAR and Accuracy on the Test set obtained with the resulting best RF models for positive and negative strokes. Orange bars refer to the model created re-gridding all data of the DF over the HP resolution. Yellow bars refer to the model created re-gridding all data of the DF over the HW resolution. Green bars refer to the model created re-gridding all data of the DF over the HT resolution.

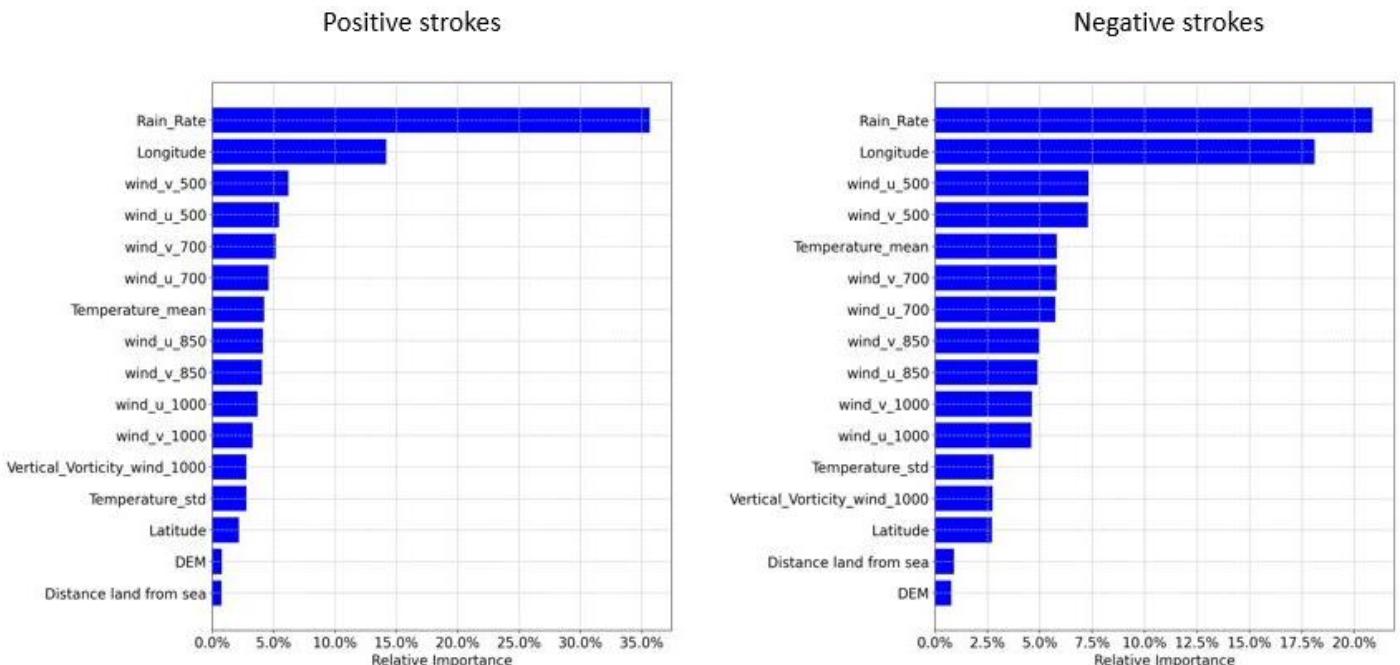


Figure 3: Features' importance of the resulting best RF model obtained for data gridded with HW resolution, for positive and negative strokes. Wind vorticity is labelled as “vertical_vorticity_wind_1000”, zonal (meridional) components of horizontal wind are labelled as wind_u (wind_v) followed by the pressure level (e.g., “wind_u_1000”).

V. STATEMENTS AND DECLARATIONS

Authors disclose financial interests that are directly or indirectly related to the work submitted for publication.

ACKNOWLEDGMENT

The data have been received by CESI S.p.A and Meteorage SAS, previous and actual owner of the SIRF network, and by the Italian Civil Protection Department, that we here thank.

REFERENCES

Adamo, C., Goodman, S., Mugnai, A., & Weinman, J. A. (2009). Lightning measurements from satellites and significance for storms in the Mediterranean. In *Lightning:*

Principles, Instruments and Applications (pp. 309–329). Springer. https://doi.org/10.1007/978-1-4020-9079-0_14

Ahijevych, D., Pinto, J. O., Williams, J. K., & Steiner, M. (2016). Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Weather and Forecasting*, 31(2), 581–599.

Amato, F., Guignard, F., Robert, S., & Kanevski, M. (2020). A novel framework for spatio-temporal prediction of environmental data using deep learning. *Scientific Reports*, 10(1), 1–11. <https://doi.org/10.1038/s41598-020-79148-7>

Blouin, K. D., Flannigan, M. D., Wang, X., & Kochtubajda, B. (2016). Ensemble lightning prediction models for the province of Alberta, Canada. *International Journal of Wildland Fire*, 25(4), 421–432. <https://doi.org/10.1071/WF15111>

Bothwell, P. D. (2005). Development of an operational statistical scheme to predict the location and intensity of

- lightning. *Conference on Meteorological Applications of Lightning Data.* https://ams.confex.com/ams/Annual2005/techprogram/paper_85013.htm.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Bright, D. R., Wandishin, M. S., Jewell, R. E., & Weiss, S. J. (2005). A physically based parameter for lightning prediction and its calibration in ensemble forecasts. *Preprints, Conf. on Meteor. Appl. of Lightning Data, Amer. Meteor. Soc., San Diego, CA*, 3496, 30. https://ams.confex.com/ams/Annual2005/techprogram/paper_84173.htm.
- Bryan, G. H., & Morrison, H. (2012). Sensitivity of a simulated squall line to horizontal resolution and parameterization of microphysics. *Monthly Weather Review*, 140(1), 202–225. <https://doi.org/10.1175/MWR-D-11-00046.1>
- Calhoun, K. M., MacGorman, D. R., Ziegler, C. L., & Biggerstaff, M. I. (2013). Evolution of lightning activity and storm charge relative to dual-Doppler analysis of a high-precipitation supercell storm. *Monthly Weather Review*, 141(7), 2199–2223. <https://doi.org/10.1175/MWR-D-12-00258.1>
- Cooper, M. A., & Holle, R. L. (2019). Current Global Estimates of Lightning Fatalities and Injuries. In *Reducing Lightning Injuries Worldwide* (pp. 65–73). Springer. https://doi.org/10.1007/978-3-319-77563-0_6
- Cooray, V., & Arevalo, L. (2017). Modeling the stepping process of negative lightning stepped leaders. *Atmosphere*, 8(12), 245. <https://doi.org/10.3390/atmos8120245>
- Deierling, W., & Petersen, W. A. (2008). Total lightning activity as an indicator of updraft characteristics. *Journal of Geophysical Research: Atmospheres*, 113(D16). <https://doi.org/10.1029/2007JD009598>
- Diendorfer, G., Schulz, W., Cummins, C., Rakov, V., Bernardi, M., De La Rosa, F., Hermoso, B., Hussein, A. M., Kawamura, T., & Rachidi, F. (2009). Review of CIGRE Report “Cloud-to-Ground Lightning Parameters Derived from Lightning Location Systems—The Effects of System Performance”. *CIGRE Colloq. Harmonizing Environment, Power Qual. Power Syst.*
- Enno, S.-E., Sugier, J., Alber, R., & Seltzer, M. (2020). Lightning flash density in Europe based on 10 years of ATDnet data. *Atmospheric Research*, 235, 104769. <https://doi.org/10.1016/j.atmosres.2019.104769>
- Fierro, A. O., Leslie, L., Mansell, E., Straka, J., MacGorman, D., & Ziegler, C. (2007). A high-resolution simulation of microphysics and electrification in an idealized hurricane-like vortex. *Meteorology and Atmospheric Physics*, 98(1), 13–33. <https://doi.org/10.1175/2010JAS3659.1>
- Fierro, A. O., Mansell, E. R., Ziegler, C. L., & MacGorman, D. R. (2014). Explicit electrification and lightning forecast implemented within the WRF-ARW model. *XV International Conference on Atmospheric Electricity*.
- Fiori, E., Parodi, A., & Siccardi, F. (2010). Turbulence closure parameterization and grid spacing effects in simulated supercell storms. *Journal of the Atmospheric Sciences*, 67(12), 3870–3890. <https://doi.org/10.1175/2010JAS3359.1>
- Guignard, F., Amato, F., & Kanevski, M. (2021). Uncertainty quantification in extreme learning machine: Analytical developments, variance estimates and confidence intervals. *Neurocomputing*, 456, 436–449. <https://doi.org/10.1016/j.neucom.2021.04.027>
- Huang, X.-G. (2021). Vorticity and Spin Polarization—A Theoretical Perspective. *Nuclear Physics A*, 1005, 121752. <https://doi.org/10.1016/j.nuclphysa.2020.121752>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer. <https://doi.org/10.1007/978-1-4614-7138-7>
- Kain, J. S., Weiss, S. J., Bright, D. R., Baldwin, M. E., Levit, J. J., Carbin, G. W., Schwartz, C. S., Weisman, M. L., Drogemeier, K. K., & Weber, D. B. (2008). Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Weather and Forecasting*, 23(5), 931–952. <https://doi.org/10.1175/WAF2007106.1>
- Kain, J. S., Weiss, S. J., Levit, J. J., Baldwin, M. E., & Bright, D. R. (2006). Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Weather and Forecasting*, 21(2), 167–181. <https://doi.org/10.1175/WAF906.1>
- Kanevski, M., Pozdnoukhov, A., Pozdnukhov, A., & Timonin, V. (2009). *Machine learning for spatial environmental data: Theory, applications, and software*. EPFL press. <https://doi.org/10.1201/9781439808085>
- Kotroni, V., & Lagouvardos, K. (2008). Lightning occurrence in relation with elevation, terrain slope, and vegetation cover in the Mediterranean. *Journal of Geophysical Research: Atmospheres*, 113(D21). <https://doi.org/10.1029/2008JD010605>
- Kotroni, V., & Lagouvardos, K. (2016). Lightning in the Mediterranean and its relation with sea-surface temperature. *Environmental Research Letters*, 11(3), 034006. <https://doi.org/10.1088/1748-9326/11/3/034006>
- La Fata, A., Amato, F., Bernardi, M., D’Andrea, M., Procopio, R., & Fiori, E. (2021). Cloud-to-Ground lightning nowcasting using Machine Learning. *Cloud-to-Ground Lightning Nowcasting Using Machine Learning*.
- Lagasio, M., Parodi, A., Procopio, R., Rachidi, F., & Fiori, E. (2017). Lightning Potential Index performances in multimicrophysical cloud-resolving simulations of a back-building mesoscale convective system: The Genoa 2014 event. *Journal of Geophysical Research: Atmospheres*, 122(8), 4238–4257. <https://doi.org/10.1002/2016JD026115>
- Leuenberger, M., & Kanevski, M. (2015). Extreme Learning Machines for spatial environmental data. *Computers & Geosciences*, 85, 64–73. <https://doi.org/10.1016/j.cageo.2015.06.020>
- Lynn, B. H. (2017). The usefulness and economic value of total lightning forecasts made with a dynamic lightning scheme coupled with lightning data assimilation. *Weather and Forecasting*, 32(2), 645–663. <https://doi.org/10.1175/WAF-D-16-0031.1>
- Lynn, B. H., Yair, Y., Price, C., Kelman, G., & Clark, A. J. (2012). Predicting cloud-to-ground and intracloud lightning

- in weather forecast models. *Weather and Forecasting*, 27(6), 1470–1488. <https://doi.org/10.1175/WAF-D-11-00144.1>
- Mansell, E. R., MacGorman, D. R., Ziegler, C. L., & Straka, J. M. (2002). Simulated three-dimensional branched lightning in a numerical thunderstorm model. *Journal of Geophysical Research: Atmospheres*, 107(D9), ACL 2-1–ACL 2-12. <https://doi.org/10.1029/2000JD000244>
- Mazarakis, N., Kotroni, V., Lagouvardos, K., & Argiriou, A. A. (2008). Storms and lightning activity in Greece during the warm periods of 2003–06. *Journal of Applied Meteorology and Climatology*, 47(12), 3089–3098. <https://doi.org/10.1175/2008JAMC1798.1>
- McCaull Jr, E. W., Goodman, S. J., LaCasse, K. M., & Cecil, D. J. (2009). Forecasting lightning threat using cloud-resolving model simulations. *Weather and Forecasting*, 24(3), 709–729. <https://doi.org/10.1175/2008WAF2222152.1>
- McCaull Jr, E. W., Priftis, G., Case, J. L., Chronis, T., Gatlin, P. N., Goodman, S. J., & Kong, F. (2020). Sensitivities of the WRF Lightning Forecasting Algorithm to Parameterized Microphysics and Boundary Layer Schemes. *Weather and Forecasting*, 35(4), 1545–1560. <https://doi.org/10.1175/WAF-D-19-0101.1>
- Mecikalski, J. R., Williams, J. K., Jewett, C. P., Ahijevych, D., LeRoy, A., & Walker, J. R. (2015). Probabilistic 0–1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *Journal of Applied Meteorology and Climatology*, 54(5), 1039–1059. <https://doi.org/10.1175/JAMC-D-14-0129.1>
- Mostajabi, A., Finney, D. L., Rubinstein, M., & Rachidi, F. (2019). Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. *Npj Climate and Atmospheric Science*, 2(1), 1–15. <https://doi.org/10.1038/s41612-019-0098-0>
- Nag, A., & Rakov, V. A. (2012). Positive lightning: An overview, new observations, and inferences. *Journal of Geophysical Research: Atmospheres*, 117(D8). <https://doi.org/10.1029/2012JD017545>
- Nicora, M., Mestriner, D., Brignone, M., Bernardi, M., Procopio, R., & Fiori, E. (2021). A 10-year study on the lightning activity in Italy using data from the SIRF network. *Atmospheric Research*, 256, 105552. <https://doi.org/10.1016/j.atmosres.2021.105552>
- Paliaga, G., Donadio, C., Bernardi, M., & Faccini, F. (2019). High-resolution lightning detection and possible relationship with rainfall events over the Central Mediterranean Area. *Remote Sensing*, 11(13), 1601. <https://doi.org/10.3390/rs11131601>
- Petersen, W. A., Cifelli, R. C., Rutledge, S. A., Ferrier, B. S., & Smull, B. F. (1999). Shipborne dual-Doppler operations during TOGA COARE: Integrated observations of storm kinematics and electrification. *Bulletin of the American Meteorological Society*, 80(1), 81–98. [https://doi.org/10.1175/1520-0477\(1999\)080<0081:SDDDT>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<0081:SDDDT>2.0.CO;2)
- Poelman, D. R. (2014). A 10-year study on the characteristics of thunderstorms in Belgium based on cloud-to-ground lightning data. *Monthly Weather Review*, 142(12), 4839–4849. <https://doi.org/10.1175/MWR-D-14-00202.1>
- Potvin, C. K., & Flora, M. L. (2015). Sensitivity of idealized supercell simulations to horizontal grid spacing: Implications for Warn-on-Forecast. *Monthly Weather Review*, 143(8), 2998–3024. <https://doi.org/10.1175/MWR-D-14-00416.1>
- Price, C., & Rind, D. (1992). A simple lightning parameterization for calculating global lightning distributions. *Journal of Geophysical Research: Atmospheres*, 97(D9), 9919–9933. <https://doi.org/10.1029/92JD00719>
- Rakov, V. A. (2013). The physics of lightning. *Surveys in Geophysics*, 34(6), 701–729. <https://doi.org/10.1007/S10712-013-9230-6>
- Schultz, C. J., Carey, L. D., Schultz, E. V., & Blakeslee, R. J. (2015). Insight into the kinematic and microphysical processes that control lightning jumps. *Weather and Forecasting*, 30(6), 1591–1621. <https://doi.org/10.1175/WAF-D-14-00147.1>
- Schultz, C. J., Carey, L. D., Schultz, E. V., & Blakeslee, R. J. (2017). Kinematic and microphysical significance of lightning jumps versus nonjump increases in total flash rate. *Weather and Forecasting*, 32(1), 275–288. <https://doi.org/10.1175/WAF-D-15-0175.1>
- Skamarock, W. C. (2004). Evaluating mesoscale NWP models using kinetic energy spectra. *Monthly Weather Review*, 132(12), 3019–3032. <https://doi.org/10.1175/MWR2830.1>
- Soula, S., & Chauzy, S. (2001). Some aspects of the correlation between lightning and rain activities in thunderstorms. *Atmospheric Research*, 56(1–4), 355–373. [https://doi.org/10.1016/S0169-8095\(00\)00086-7](https://doi.org/10.1016/S0169-8095(00)00086-7)
- Steppeler, J., Doms, G., Schättler, U., Bitzer, H. W., Gassmann, A., Damrath, U., & Gregoric, G. (2003). Mesogamma scale forecasts using the nonhydrostatic model LM. *Meteorology and Atmospheric Physics*, 82(1), 75–96. <https://doi.org/10.1007/s00703-001-0592-9>
- Taalab, K., Cheng, T., & Zhang, Y. (2018). Mapping landslide susceptibility and types using Random Forest. *Big Earth Data*, 2(2), 159–178. <https://doi.org/10.1080/20964471.2018.1472392>
- Tapia, A., Smith, J. A., & Dixon, M. (1998). Estimation of convective rainfall from lightning observations. *Journal of Applied Meteorology*, 37(11), 1497–1509. <https://doi.org/10.1002/asl2.453>
- Tippett, M. K., & Koshak, W. J. (2018). A baseline for the predictability of US cloud-to-ground lightning. *Geophysical Research Letters*, 45(19), 10,719–10,728. <https://doi.org/10.1029/2018GL079750>
- Tonini, M., D'Andrea, M., Biondi, G., Degli Esposti, S., Trucchia, A., & Fiorucci, P. (2020). A Machine Learning-Based Approach for Wildfire Susceptibility Mapping. The Case Study of the Liguria Region in Italy. *Geosciences*, 10(3), 105. <https://doi.org/10.3390/geosciences10030105>
- Underwood, S. J. (2006). Cloud-to-ground lightning flash parameters associated with heavy rainfall alarms in the Denver, Colorado, Urban Drainage and Flood Control

- District ALERT Network. *Monthly Weather Review*, 134(9), 2566–2580. <https://doi.org/10.1175/MWR3201.1>
- VandenBerg, M. A., Coniglio, M. C., & Clark, A. J. (2014). Comparison of next-day convection-allowing forecasts of storm motion on 1-and 4-km grids. *Weather and Forecasting*, 29(4), 878–893. <https://doi.org/10.1175/WAF-D-14-00011.1>
- Veronesi, F., Grassi, S., & Raubal, M. (2016). Statistical learning approach for wind resource assessment. *Renewable and Sustainable Energy Reviews*, 56, 836–850. <https://doi.org/10.1016/j.rser.2015.11.099>
- Vogt, B. J., & Hodanish, S. J. (2014). A high-resolution lightning map of the state of Colorado. *Monthly Weather Review*, 142(7), 2353–2360. <https://doi.org/10.1175/MWR-D-13-00334.1>
- Wang, F., Zhang, Y., Zheng, D., & Xu, L. (2015). Impact of the vertical velocity field on charging processes and charge separation in a simulated thunderstorm. *Journal of Meteorological Research*, 29(2), 328–343. <https://doi.org/10.1007/s13351-015-4023-0>
- Weisman, M. L., Skamarock, W. C., & Klemp, J. B. (1997). The resolution dependence of explicitly modeled convective systems. *Monthly Weather Review*, 125(4), 527–548. [https://doi.org/10.1175/1520-0493\(1997\)125<0527:TRDOEM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2)
- Williams, J. K. (2014). Using random forests to diagnose aviation turbulence. *Machine Learning*, 95(1), 51–70.
- Zhou, K., Zheng, Y., Dong, W., & Wang, T. (2020). A Deep Learning Network for Cloud-to-Ground Lightning Nowcasting with Multisource Data. *Journal of Atmospheric and Oceanic Technology*, 37(5), 927–942. <https://doi.org/10.1175/JTECH-D-19-0146.1>