EPFL

Thèse n° 8805

Augmenting the performance of impoverished sensor networks using machine learning and time reversal: Application to lightning nowcasting and location

Présentée le 27 octobre 2021

Faculté des sciences et techniques de l'ingénieur
Groupe SCI STI FR
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Amirhossein MOSTAJABI

Acceptée sur proposition du jury

Prof. A. Skrivervik Favre, présidente du jury
Prof. F. Rachidi-Haeri, Prof. M. Rubinstein, directeurs de thèse
Prof. E. Williams, rapporteur
Prof. V. Cooray, rapporteur
Prof. M. Jaggi, rapporteur

École polytechnique fédérale de Lausanne

2021

Acknowledgement

I am profoundly grateful to my unique supervisors, Prof. Farhad Rachidi and Prof. Marcos Rubinstein for their consistent support and guidance during my doctoral study. They continuously provided encouragement and were always willing and enthusiastic to assist in any way they could throughout these years. My sincere and heartfelt gratitude and appreciation for everything you have taught me both inside and beyond the scope of graduate studies!

I also appreciate the insightful comments of my thesis committee members, Professors Earle Williams, Vernon Cooray, Martin Jaggi, and Anja Skrivervik for their constructive comments and suggestions.

During my doctoral study, I have had the incredible serendipity in my career to work on several amazing projects with world-class researchers at Roche, Apple, and CSEM. I would like to thank colleagues at Roche in Basel, Switzerland: Dr. Yohan Farouz, Mrs. Susanne Weissenborn, Dr. Jitao David Zhang, and Dr. Shanon Seger, who gave me the wonderful internship opportunity to gain industrial experience in medical data science. I am truly indebted to the people at Apple: Dr. Jouya Jadidian, Dr. Cheung-Wei Lam, Dr. Misha Tsiklauri, Dr. Sujeet Patole, and Dr. Jessica Desai for a terrific six months research collaboration. I wouldn't have grown as much as I have these six months without your detailed feedback and unwavering support. Particularly, I would like to express my deepest appreciation to Dr. Jouya Jadidian who has been a tremendous source of valuable advice and encouragement throughout my academic life. I'm so grateful to Dr. Andrea Dunbar and Dr.

Pedram Pad from Sector Edge AI and Vision at CSEM in Neuchâtel, Switzerland for their technical advice, helpful contributions, and insightful discussions.

I am thankful to my friends and colleagues at the Electromagnetic Compatibility Laboratory for their collaborative efforts and to Ms. Eulalia Durussel and all other EPFL administrative staff, who always had time for helping me out, no matter how busy they were.

I would like to acknowledge the European Commission H2020 and Laser Lightning Rod (LLR) Consortium for the financial and technical support of this research.

Last but not least, I would like to express my sincere gratitude to my beloved family. Words cannot express how grateful I am to my family members for their unconditional love and unwavering faith throughout these years.

Amirhossein Mostajabi

Lausanne, 04.10.2021

Abstract

Lightning can have a considerable influence on the environment and the economy since it causes energy supply outages, forest fires, damages to infrastructure, injury, and deaths of humans and livestock worldwide. These severe and costly outcomes can be averted or mitigated by predicting the lightning occurrence in advance and taking preventive actions accordingly. Therefore, an accurate and fast lightning prediction method is of considerable value. Lightning is formed in the atmosphere through the combination of intricate dynamic and microphysical processes. Mainly due to this complexity, attempts to solve the important problem of lightning prediction have generally failed to yield accurate results. Furthermore, current prediction systems are slow and very complex, and they require expensive external data acquired by radar or satellites.

In this thesis, we propose a machine learning approach to provide early lightning warnings using data that can be obtained from any weather station. We focused on a small area (usually around a critical infrastructure) targeting a higher spatiotemporal resolution and accuracy. The model is customized for each area of interest to account for the variation of the lightning activity pattern, driving mechanisms, and local conditions from one site to another, hence providing tailor-made, site-specific lightning warnings. The algorithm uses four local atmospheric parameters that can be acquired with readily available sensors to produce excellent prediction of the occurrence of lightning during three subsequent ten-minute intervals and within a radius of 30 km. It allowed to successfully hindcast lightning hazards using single-site ground-based observations from one of 12 stations selected in Switzerland. Being independent of external data sources such as radar, satellite, and weather model outputs, the algorithm can be used with commodity weather stations to scale

up their functionality as an early lightning warning system. That means we can cover remote regions that are out of radar and satellite range and where communication networks are unavailable.

In addition to predicting “when” lightning will occur, determining “where” it struck (localization) is also important in a wide range of research and application domains, including geophysical research, lightning warning, aviation/air traffic, weather services, insurance claims and power transmission and distribution, etc. For example, some lightning warning systems rely on such data to indicate approaching thunderstorms and thus to prevent catastrophic effects of lightning strikes to critical infrastructure, sensitive equipment or systems, and outdoor facilities. Localization is key not only to lightning safety but also to data collection for high-resolution nowcasting with Machine Learning. Over the years, researchers have explored different approaches to lightning localization. However, all current approaches require the presence of multiple RF sensors and, hence, are not applicable in many practical scenarios where the size and power consumption of the device matter. We broke this barrier by developing two machine learning-based lightning location systems. The first one uses the lightning-induced voltage measurements from two preinstalled sensors on transmission lines to localize nearby lightning. The second works on a data from a single electric field sensor to estimate the 2D geolocation of the lightning strike point. The key to reducing the sensor array size to one is hybridizing a powerful RF imaging technique called Electromagnetic Time Reversal (EMTR) with advanced techniques in computer vision and machine learning. Both localization models have been trained and optimized using fully synthetic data.

Where possible, experimental confirmation has been done to determine how the methods generalize to real world scenarios. For example, we evaluated the performance of the second approach on 6 return strokes associated with two upward lightning flashes striking the Säntis tower in Switzerland. The comparison to the results from current lightning location systems reveals that the proposed approach can yield similar location accuracy with a significantly smaller number of sensors. Finally, we also demonstrated the skill of the hybrid EMTR/ML methodology in localizing electromagnetic sources other than lightning, namely RF sources with a bandwidth of 0.1–10 MHz.

Keywords

machine learning, time reversal, convolutional neural networks, deep neural network, lightning nowcasting, source localization, lightning localization, time-to-event analysis, imbalanced classification, transfer learning, digital signal processing.

Résumé

La foudre peut avoir un impact considérable sur l'environnement et l'économie car elle est la cause de pannes d'approvisionnement en énergie, d'incendies de forêt, de dommages aux infrastructures, de blessures et de la mort d'êtres humains et de bétail dans le monde entier. Ces conséquences graves et coûteuses peuvent être évitées ou atténuées en prédisant à l'avance l'occurrence de la foudre et en prenant des mesures préventives en conséquence. Par conséquent, une méthode de prédiction de la foudre précise et rapide est d'une valeur considérable. La foudre se forme dans l'atmosphère par la combinaison de processus dynamiques et microphysiques complexes. Principalement en raison de cette complexité, les tentatives pour résoudre le problème important de la prédiction de la foudre n'ont généralement pas donné de résultats d'une précision suffisante. De plus, les systèmes de prédiction actuels sont lents et très complexes, et ils nécessitent des données externes coûteuses car acquises par radar ou satellites.

Dans cette thèse, nous proposons une approche d'apprentissage automatique pour fournir des alertes de foudre en utilisant des données pouvant être obtenues à partir de n'importe quelle station météo. Nous nous sommes concentrés sur une petite zone (généralement autour d'une infrastructure critique) ciblant une résolution et une précision spatio-temporelles plus élevées. Le modèle est particularisé à chaque zone d'intérêt pour tenir compte de la variation du modèle d'activité de la foudre, des mécanismes sous-jacents et des conditions locales d'un site à l'autre, fournissant ainsi des avertissements de foudre sur mesure et spécifiques au site. L'algorithme utilise quatre paramètres atmosphériques locaux qui peuvent être acquis avec des capteurs facilement disponibles pour produire une excellente prédiction de l'apparition de la foudre pendant trois intervalles consécutifs de dix minutes et dans un rayon de 30 km. L'algorithme permet de simuler

rétrospectivement avec succès les risques de foudre à l'aide d'observations au sol sur un seul site à partir de chacune des 12 stations sélectionnées en Suisse. Étant indépendant des sources de données externes telles que le radar ou les satellite ainsi que de modèles météorologiques, l'algorithme peut être utilisé avec des stations météorologiques de base pour étendre leur fonctionnalité en tant que système d'alerte d'orages. Cela signifie que nous pouvons couvrir des régions éloignées qui sont hors de portée des radars et des satellites et où les réseaux de communication ne sont pas disponibles.

En plus de prédire « quand » la foudre se produira, déterminer « où » elle a frappé (localisation) est également importante dans un large éventail de domaines de recherche et d'application, y compris la recherche géophysique, les systèmes d'alerte, l'aviation/le trafic aérien, les services météorologiques, les réclamations d'assurance et la transmission et la distribution d'électricité, etc. Par exemple, certains systèmes d'alerte à la foudre s'appuient sur ces données pour indiquer l'approche d'orages et ainsi prévenir les effets catastrophiques de la foudre sur les infrastructures critiques, les équipements ou systèmes sensibles et les installations extérieures. La localisation est la clé non seulement de la sécurité contre la foudre, mais également de la collecte de données pour la prévision immédiate haute résolution avec les techniques d'apprentissage automatique (Machine Learning). Au fil des ans, les chercheurs et chercheuses ont exploré différentes approches pour la localisation de la foudre. Cependant, toutes les approches actuelles nécessitent la présence de plusieurs capteurs RF et, par conséquent, ne sont pas applicables dans de nombreux scénarios pratiques où la taille et la consommation d'énergie de l'appareil sont importantes. Nous avons franchi cette barrière en développant deux systèmes de localisation de foudre basés sur l'apprentissage automatique. Le premier utilise les mesures de tension induite par la foudre à partir de deux capteurs préinstallés sur les lignes de transmission pour localiser la foudre à proximité. Le second se base sur des données obtenues à partir d'un seul capteur de champ électrique pour estimer la géolocalisation 2D du point d'impact de la foudre. La clé pour réduire la taille du réseau de capteurs à un est l'hybridation d'une puissante technique d'imagerie RF appelée retournement temporel électromagnétique (EMTR) avec des techniques avancées de vision par ordinateur et d'apprentissage automatique. Les deux modèles de localisation ont été entraînés et optimisés à l'aide de données entièrement synthétiques.

Dans la mesure du possible, une confirmation expérimentale a été effectuée pour déterminer comment les méthodes se généralisent aux scénarios du monde réel. Par exemple, nous avons

évalué les performances de la deuxième approche sur 6 arcs en retour associés à deux éclairs ascendants frappant la tour Säntis en Suisse. La comparaison avec les résultats des systèmes actuels de localisation de la foudre révèle que l'approche proposée peut donner une précision de localisation similaire avec un nombre de capteurs significativement plus petit. Enfin, nous avons également démontré l'habileté de la méthodologie hybride EMTR/ML à localiser les sources électromagnétiques autres que la foudre, à savoir les sources RF avec une bande passante de 0.1 à 10 MHz.

Mots-clés

apprentissage automatique, retournement temporel, réseaux de neurones convolutifs, réseau de neurones profonds, prévision immédiate de la foudre, localisation de la source, localisation de la foudre, analyse du temps jusqu'à l'événement, classification déséquilibrée, apprentissage par transfert, traitement du signal numérique.

Contents

Abstract ..... **i**

Résumé ..... **v**

Chapter 1 Introduction **15**

1.1 Thesis Outline 19

Chapter 2 Nowcasting lightning occurrence from commonly-available meteorological parameters using machine learning techniques **21**

2.1 Introduction 22

2.2 Results 29

2.2.1 Machine Learning Model performance for long-range lightning activity at 12 stations in Switzerland 29

2.2.2 Distribution patterns of data among different stations 37

2.2.3 Change of distribution patterns over time 40

2.3 Discusion 41

2.4 Methods 45

2.4.1 Data gathering 45

2.4.2 Stage #1: Model selection, generation, and tuning 50

2.4.3 Stage #2: Training and testing procedure 54

2.4.4 Model evaluation metrics 56

Chapter 3 Machine Learning Based Lightning Localization Algorithm Using Lightning-Induced Voltages on Transmission Lines **59**

3.1 Introduction 60

3.2 Methodology 62

3.2.1 Numerical Simulation and Data Acquisition 62

3.2.2 Selection of the Machine Learning Model 65

3.2.3 Model generation 65

3.3 EVALUATION OF THE MACHINE LEARNING MODEL 67

3.3.1 The Impact of the Sensors' Positions 70

3.3.2 Increasing the Number of Sensors 71

3.3.3 Impact of the Grid Size 72

3.3.4 Sensitivity to the Noise Level and the Risetime of the Lightning Current 73

3.4 Conclusions 75

Chapter 4 Single-Sensor Source Localization Using Electromagnetic Time Reversal and Deep Transfer Learning: Application to Lightning 77

4.1 Introduction 78

4.2 Methods and Results 80

4.2.1 Case Study 1 80

4.2.2 Case study 2: Application to Locate Lightning Flashes 94

4.3 Discussion 99

Chapter 5 Conclusion 103

5.1 Future directions 104

5.1.1 Lightning nowcasting map 104

5.1.2 RF Machine Perception 105

Appendix A: Supplementary Information for Chapter 2 107

Appendix B: Supplementary Information for Chapter 4 116

Bibliography 123

List of Figures

Figure 2-1 Probability density estimates of the surface pressure and surface temperature for lighting-inactive and lighting-active samples in Subsets 1 and 2 using a kernel smoothing function. The Subsets 1 and 2 include observations of four meteorological parameters respectively at Säntis and Monte San Salvatore stations from 2006 up to 2017 with the granularity of ten minutes. The corresponding lead time range is 0-10 minutes. Thus, a recorded observation at the start of a ten-minute interval is labeled according to lightning activity in that interval as either a 'lighting-inactive' sample (without any long-range lightning activity) or a 'lighting-active' sample (with at least one long-range lightning activity recorded). 31

Figure 2-2 Evaluation of the skill of warnings of long-range lightning activity for three ranges of lead times at a, b, c Säntis (Subset 1) and a', b', c' Monte San Salvatore (Subset 2) station. The results from ML model are shown as solid black columns, results from PERSISTENT model are shown as columns filled with dot patterns, results from CAPE model are shown as columns filled with vertical lines, and results from E-FIELD model are shown as solid grey columns (POD: Probability of Detection, FAR: False Alarm Ratio, CSI: Critical Success Index, HSS: Heidke Skill Score). 33

Figure 2-3 Predictors importance estimates for long-range lightning activities. The result includes both studies at individual stations and over all stations. The data from individual stations are standardized according to their local mean and deviation before they are included in the overall study. The presented results correspond to the lead time of 0-10 minutes. 37

Figure 2-4 Comparison of probability density functions (PDFs) of each parameter in lighting-active class samples of Group A and Group B. The corresponding lead time range is 0-10 minutes. For the sake of comparison, the parameters values at each station are standardized based on the local mean and standard deviation values. The standardizing function puts mean of each parameter at zero and scales the parameters by their standard deviations. 39

Figure 2-5 The contribution of individual variables to the first and second principle components for all 12 stations during 2006 to 2017. In each subplot, data for two stations with the largest distance from the cluster center are annotated in red and the rest are presented in black. The corresponding lead time range is 0-10 minutes. 40

Figure 2-6 The contribution of individual variables to the first and second principle components from 2006 to 2017 for a Zurich station (as an example of Group A) and b Säntis station (as an example of Group B). In each subplot, the horizontal axis is the coefficients for PC1 and vertical axis is coefficients for PC2. The corresponding lead time range is 0-10 minutes. 41

Figure 2-7 Parallel coordinates plot from data subset 10. The mean of each predictor is set to zero and the predictors are scaled by their standard deviations. Each line represents a recorded observation at the start of a ten-minute interval and is labeled according to lightning activity in that interval as either blue (without any long-range lightning activity) or orange (with at least one long-range lightning activity recorded). 51

Figure 2-8 Sample of a decision tree grown by the ML model in the ensemble classifier. In this example, the maximum depth of the tree is set to 3 and subset 1 is used as the training set. The prediction score at each leaf would be assigned to its associated observations. The model then combines the

prediction scores for each sample to predict its class as whether lightning active or lightning-inactive. 53

Figure 2-9 Summary of the model selection, generation, tuning, and evaluation. 55

Figure 3-1 Sketch diagram of the detection region. The infinite transmission line is located 50 m beside the detection region. The two voltage sensors are located on the line and 2 km away from each other (Δ=2 km). The coordinates of the sensor positions are annotaed in red. 62

Figure 3-2 A zoomed view of a sample of the induced voltages measured by the two sensors. The 2D coordinates of the lightning strike point is [6996.9 m, 9459.3 m] and the excitation source is a step current ascending along the lightning strike channel. The voltages are calculated using the Rusck's formula and they are shifted so that the first signal starts from time equals to zero. 64

Figure 3-3 Model evaluation results for (a) the x-coordiante and (b) the y-coordiante. (c) is the histogram of the location error considering 2000 randomly selected lightning strike points inside. 68

Figure 3-4 Scatter plot of the target (blue dots) versus estimated (yellow dots) 2D geolocations for the N = 2000 guest lightning strike points using the proposed ML based approach. 69

Figure 3-5 Avarage location error presented as heatmap chart inside the detection region. The colormap represent the location errors in m. The (0,0) point corresponds to the coordinate center (O) shown in Figure 3-1. The x and y labels for each of the cells are the coordinates of the left-bottom corner of the grid cell. 70

Figure 3-6 (a) Cumulative probability and (b) probability density estimate of the location errors for four different sensors' positions. Δ is the distance between the two deployed voltage sensors on th transmission line (see Figure 3-1). 71

Figure 3-7 Sketch diagram of the detection region for the case of three voltage sensors. The coordinates of the sensor positions are written in red. 72

Figure 3-8 Same as in Figure 3-5 when the detection region is changed to the one shown in Figure 3-7 with three voltage sensors in use. 72

Figure 3-9 (a) Cumulative probability and (b) probability density estimate of the location errors for three different grid sizes and three voltage sensors. 73

Figure 3-10 As in Figure 3-8 but for excitation sources with random values for the risetime. The sensor signals are noisy, and the number of data points is increased to 10000. 74

Figure 4-1 Geometry of the problem in the first case study. The blue/filled circles and the red/unfilled circle show the scatterers and the location of the electric field sensor, respectively. The solution space spans 13.44 x 13.44 km2. 81

Figure 4-2 Example of the simulation results for the source and the field at a sensor. (a) The linear current density of the excitation source (Gaussian RF with 10 MHz bandwidth) and (b) the time-reversed version of the signal measured by the electric field sensor normalized to its maximum value. 82

Figure 4-3 (a) Maximum amplitude of vertical electric field at all time steps normalized to its maximum value and (b) 2D profile of the vertical electric field at the last time step. The red cross is the ground truth of the source position, S1 and S2 are the scattering objects, and the white circle is the electric field sensor. The nominated point/points by EMTR are also annotated for on each panel. 84

Figure 4-4 Model performance results for the case of numerical simulations using machine learning. (a) Estimation of the x coordinate and (b) the y coordinate of the randomly selected source locations. (c) Histogram of the location error. 86

Figure 4-5 A sample of the generated images out of the 2D surface vertical electric field values in the detection region. The corresponding source position is shown as a red cross on the image. The images are produced using the jet colormap array with 216 elements. 88

Figure 4-6 Model performance results for the case of numerical simulations using combinational EMTR/ML approach. The detection region is shown in Figure 4-1 including two scattering objects. (a) Estimation of the x coordinate and (b) the y coordinate of the randomly selected source locations. (c) Histogram of the location error. 92

Figure 4-7 Average location error presented as a heatmap chart inside the detection region for the case of numerical simulations using the EMTR/ML combinational approach. The size of the scatterers, their number and locations, and the frequency of the excitation source remained fixed for simulations at each of the source locations. The colormap represents the location errors in meters. The (0,0) point corresponds to the coordinate center (O) shown in Figure 4-1. The x and y labels for each of the cells are the coordinates of the bottom-left corner of the grid cell. 93

Figure 4-8 Average location error presented as a heatmap chart inside the detection region for the case of numerical simulations using the combinational EMTR/ML approach. A third scatterer, was added randomly to some of the cases with its center at [8.05, 8.05]. The size of the scatterers as well as the frequency of the excitation source were selected randomly for simulations at each of the source locations. The colormap represents the location errors in meters. The (0,0) point corresponds to the coordinate center (O) shown in Figure 4-1. The x and y labels for each of the cells are the coordinates of the bottom-left corner of the grid cell. 94

Figure 4-9 Geometry of the problem in the second case study, the blue/filled circles represent four actual tall mountains around the Säntis, modeled as cylinders. The red/unfilled circle shows the electric field sensor at Herisau. The red cross is the position of the Säntis Tower. The solution space spans 15.78 x 15.78 km2. 96

Figure 4-10 Simultaneous recordings of the vertical electric field at Herisau and the channel-base current associated with a return stroke of an upward lightning flash that occurred on on July 18th, 2017 at 18:31:35 (local time) at the Säntis Tower. The vertical electric field data were used for the experimental validation of the proposed approach. 97

Figure 4-11 EUCLID pulse location errors for upward negative flashes striking the Säntis Tower recorded in the period of June 2010 until December 2013. The ground truth target for all pulses is the Säntis Tower and the estimated locations by EUCLID are presented as blue dots. The size of the dots is proportional to the pulse peak current amplitude. The length and width of the shown area are, respectively, 3.34 and 1.06 km. The figure is adopted from Azadifar et al. [152]. 98

Figure 4-12 Experimental validation result for the proposed combinational approach: six return strokes (RS1-6) associated with two upward lightning flash that occurred at the Säntis Tower were the excitation source and the single-sensor recording of the associated electric fields 14.7 km away were the input data for the model. The estimated lightning strike point for each of the RSs is also shown as a red cross around the target (i.e., the Säntis Tower). The map is generated using 3D Map Generator–Atlas plugin (https://graphicriver.net/item/3d-map-generator-atlas-from-heightmap-to-real-3d-map/22277498) for Adobe Photoshop CC 2017.1.1 release. 101

Figure A-1 Distribution of the lightning-active class samples for 12 stations during 2006 to 2017. In each subplot, the horizontal axis is the air pressure at the station level (QFE) and the vertical axis is the air temperature. For the sake of comparison, the parameter values at each station are standardized based on the local mean and standard deviation values. The standardizing function sets the mean of each parameter to zero and scales the parameters by their standard deviations. In this analysis, samples are classified as lightning-active or lightning-inactive based on the imminent long-range lightning activity (i.e. lead time = 0-10 minutes). 110

Figure A-2 Distribution of lightning-active class samples for 12 stations during 2006 to 2017. In each subplot, the horizontal axis is the relative humidity and the vertical axis is the wind speed. For the sake of comparison, the parameter values at each station are standardized based on the local mean and standard deviation values. The standardizing function sets the mean of each parameter to zero and scales the parameters by their standard deviations. In this analysis, samples are classified as lightning-active or lightning-inactive based on the imminent long-range lightning activity (i.e. lead time = 0-10 minutes). 111

Figure A-3 Distribution of lightning-active class samples for 12 stations during 2006 to 2017. In each subplot, the horizontal axis is the air temperature and the vertical axis is the relative humidity. For the sake of comparison, the parameters values at each station are standardized based on the local mean and standard deviation values. The standardizing function sets the mean of each parameter to zero and scales the parameters by their standard deviations. In this analysis, samples are classified as lightning-active or lightning-inactive based on the imminent long-range lightning activity (i.e. lead time = 0-10 minutes). 112

Figure A-4 Distribution of lightning-active class samples for 12 stations during 2006 to 2017. In each subplot, the horizontal axis is the wind speed and the vertical axis is the air pressure at the station level (QFE). For the sake of comparison, the parameters values at each station are standardized based on the local mean and standard deviation values. The standardizing function sets the mean of each parameter to zero and scales the parameters by their standard deviations. In this analysis, samples are classified as lightning-active or lightning-inactive based on the imminent long-range lightning activity (i.e. lead time = 0-10 minutes). 113

Figure A-5 Percentage of total variance explained by each of the first and second principle components at each station. The PCA analysis is done using data subsets 1 to 12 considering lead time range of 0-10 minutes. 114

Figure A-6 The contribution of individual variables to the first and second principle components from 2006 to 2017 for **a** Zurich station (lead time: 10-20 minutes), **b** Zurich station (lead time: 20-30 minutes), **c** Säntis station (lead time: 10-20 minutes), **d** Säntis station (lead time: 20-30 minutes). In each subplot, the horizontal axis is the coefficients for PC1 and vertical axis is coefficients for PC2. 115

Figure B-1 Visualization of the first 64 feature maps of the 2nd layer in the VGG-19 model. The input image is the one presented in Figure 4-5. 117

Figure B-2 Visualization of the first 64 feature maps of the 5th layer in the VGG-19 model. The input image is the one presented in Figure 4-5. 118

Figure B-3 Visualization of the first 64 feature maps of the 10th layer in the VGG-19 model. The input image is the one presented in Figure 4-5. 119

Figure B-4 Visualization of the first 64 feature maps of the 15th layer in the VGG-19 model. The input image is the one presented in Figure 4-5. 120

Figure B-5 Visualization of the first 64 feature maps of the 20th layer in the VGG-19 model. The input image is the one presented in Figure 4-5. 121

Figure B-6 Ambiguity of the solution in case of one symmetrical scatterer. A is an arbitrary point inside the medium and A' is the mirror of A with respect to the symmetry line. 122

List of Tables

Table 2-1 Parameters (with acronyms and definitions) used in performance evaluations of models. 29

Table 3-1 Performance results (RMSE) for state-of-the-art algorithms in the case of two sensors as shown in Figure 3-1. The dataset included 2000 samples and the evaluation method was 5-fold cross validation. 66

Table 4-1 Coordinates of the elements shown in Figure 4-9. 96

Table 4-2 Characteristics of the current waveforms associated with the considered 6 return strokes (RS1 to RS6) and their corresponding location error using the EMTR/ML approach 100

Table A-1 Geographical information of studied stations 108

Table A-2 A sample of contingency table for a two-class prediction scheme 108

Table A-3 Specifications of the engaged data subsets 109

Table A-4 List of measuring instruments for meteorological observations from SwissMetNet [95] 109

Chapter 1

Introduction

Lightning is responsible for considerable financial losses, injury and death. It can strike humans, animals, vehicles, forests, and human-made structures. Lightning can start a fire in properties and forests, it can lead to livestock death, and it can cause injuries and fatalities in humans directly and indirectly [1]–[3]. According to Curran et al. [4] and Ashley and Gilson [5], nearly 1000 annual casualties as well as damages for more than one billion U.S. dollars worldwide can be attributed to lightning [4]–[6]. In Switzerland, lightning was the second most frequent cause of human loss among the natural hazards from 1964 to 2015 [7]. Lightning may also interfere with electronic devices and circuits, buildings, and human-made structures such as transmission lines, wind turbines, and photovoltaics. Lightning is reported to have affected the aviation industry in many ways. It can cause an electrical surge striking an airplane, which in turn can lead to equipment failure and crashing. It also has an adverse impact on outdoor ramp operations such as aircraft fueling, baggage handling, and tug operations. In space centers and lunch sites, it can disturb and endanger ground and rocket lunch operations [8]–[10]. Lightning is also a major cause of damage to the growing industry of wind turbines. It can cause severe damage to wind turbine blades and induce transient surges and overvoltages in the power grid, resulting in damaging control systems and wind turbine components [11], [12].

Although the consequences of these events are severe and costly for the society and the economy, they can be mitigated or averted by predicting lightning occurrences in advance and making preparations accordingly. Therefore, lightning prediction is of great importance, and it has gained attention from many researchers and industries [13]. The process of charge separation and lightning occurrence involves complex interactions between atmospheric processes that are still not well understood; hence, the study of this phenomenon is still an ongoing field of research. Therefore, the lightning prediction generated many different approaches among researchers from various disciplines, e.g., physicists, meteorologists, and even data scientists. One direction is to implement numerical models describing the electrification process and charge separation in clouds [14]–[17]. Fierro et al. [18] proposed and implemented an explicit electrification model capable of lightning prediction within the Weather Research and Forecasting (WRF) Model. It includes in-cloud, inductive, and non-inductive collisional charging, an elliptic solution of the 3D component of the ambient electric field, and discharge parameterization [6], [18]. Another approach is to benefit from less complicated and more practical parameterization, relying only on cloud-resolving models without utilizing electrification subroutines [19]–[24]. For example, it is possible to predict both intracloud (IC) and cloud-to-ground (CG) lightning flashes dynamically by extracting the electrical potential energy from WRF's cloud-resolving model [25]. A more recent study by Tippett et al. [26] uses the product of precipitation rate and convective available potential energy (CAPE) as a proxy to CG flash prediction. This model is capable of producing significant results for lead times up to 15 days. It can predict the number of flashes, their spatial extent, and lightning/no-lightning maps.

The mentioned numerical weather models rely on lightning parameterization for prediction. However, they suffer from a significant drawback; although the recorded historical data has been taken into account for the improvement of parameterization and understanding the physics behind the phenomenon, these data are hardly utilized in the process of prediction [27]. Therefore, another approach emerges by combining historical data with numerical models, which results in better performance. Geng et al. [27] used the recorded recent data to calibrate the existing numerical prediction models.

Although these models can outperform purely numerical models, they suffer from latency in the process of simulation and prediction. They also need to process the state of the systems sequentially, which can limit the models for nowcasting scenarios.

To overcome these issues, it is possible to go in the direction of machine learning approaches utilizing emerging achievements in the domain of artificial intelligence. The advancement of computation power in the past two decades in general and also the advent of general-purpose graphics processing units (GPGPUs) enabled the processing of data beyond what has ever been possible. This breakthrough resulted in the implementation of methods to learn patterns from data in an efficient way by using machine learning. Therefore, many recent approaches prefer to benefit from machine/deep learning toolboxes in order to implement lightning prediction modules, which can learn extremely complex patterns using training datasets and, after the completion of the training process, can predict the lightning flashes in fractions of a second.

In the first part of the thesis, we focused on leveraging advances in machine learning and pattern recognition algorithms to develop a bespoke lightning nowcasting scheme. Apart from lightning diagnostic schemes producing weather maps over large areas, we aimed at increasing the temporal resolution and the accuracy by focusing on a small area (usually around a critical infrastructure). The model is customized for each area of interest to account for the variation of the lightning activity pattern, driving mechanisms, and local conditions from one site to another, hence providing a tailor-made, site-specific lightning nowcasting. Furthermore, to increase the practicality of the approach, we limited the input data to a set of commonly-available surface station parameters. In this way, raw data from a simple personal weather station are enough for the model to operate - independent of the availability of other resources, such as sophisticated upper-atmosphere measurements or numerical weather model outputs. The proposed technique consists of a machine learning model that is trained to nowcast lightning incidence during any one of three consecutive 10-minute time intervals and within a circular area of 30 km radius around a meteorological station. The algorithm

uses four local atmospheric parameters that can be acquired with readily available sensors and produce an excellent prediction of lightning occurrence up to 30 minutes in advance and within a radius of 30 km. The four parameters are (i) the surface air pressure, (ii) the air temperature, (iii) the relative air humidity, and (iv) the wind speed. The detailed results are reported in Chapter 2.

In addition to predicting lightning occurrence, knowing the exact geolocation of a lightning strike is important in a wide range of applications. Lightning Location Systems (LLS) provide data on lightning intensity, its location, and thunderstorm movements. These data can be then utilized by, e.g., public utilities and insurance companies for protection and damage estimation. Some lightning warning systems rely on such data to indicate the approaching thunderstorms [28], [29] and thus to prevent catastrophic effects of lightning strikes to critical infrastructure, sensitive equipment or systems, and outdoor facilities. Even in high-energy atmospheric physics the lightning location data are needed, for instance, to seek for the lightning flashes associated with the Terrestrial Gamma-ray Flashes (TGFs) seen from space [30].

The most widely used lightning location techniques are the time-of-arrival method (ToA) and the magnetic direction finding method [31]–[33]. More recently, the Electromagnetic Time Reversal (EMTR) has also been applied to locate lightning discharges [34]–[37]. Despite the existence of different approaches applied to the important problem of lightning localization, there are still two main limitations: (i) They require the installation of dedicated sensors such as, for instance, ELF, VLF, or VHF networks, and (ii) the existing approaches rely on the data from multiple sensors (typically three or more).

To tackle these issues, we devised a new machine-learning-based 2D lightning localization algorithm that, first, takes advantage of the preinstalled voltage measurement systems on power transmission lines to get the data and, second, needs only two sensors to operate. We then continued to reduce the required sensors even further by combining machine learning with electromagnetic time reversal. The proposed method was applied to the localization of lightning discharges in the Säntis

region in Northeastern Switzerland. The model was trained based on FDTD simulation results and tested using experimental observations of lightning flashes in the Säntis region. The experimental validation results show a high accuracy for the proposed hybrid approach in finding the 2D geolocation of the lightning strike point at the Säntis Tower using only one sensor.

1.1 Thesis Outline

The thesis is organized into two main parts: I) Lightning nowcasting and II) Lightning localization. Each part consists of one or two published papers.

We start, in Part I, with our paper on nowcasting lightning occurrence from ground-based meteorological measurements using machine learning. The paper, presented as Chapter 2, demonstrates that the values of four readily measurable local atmospheric quantities can be used to produce excellent predictors for the occurrence of lightning during three subsequent ten-minute intervals and within a radius of 30 km. The data were obtained from the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) online database for the period between 2006 to 2017 with a granularity of 10 minutes. We also took advantage of the lightning data available from MeteoSwiss to label each of these meteorological observations based to their imminent lightning activity condition. Once the database was formed, an Extreme Gradient Boosting (XGBoost) algorithm was trained with some portion of the data (training set) to learn how to predict the lightning activity given the meteorological observations. Once trained, an unseen portion of the data (testing set) was used to test the accuracy of warnings. Based on the results, the ML model shows recalls exceeding 70% and precisions above 90% for all three considered lead time ranges. The results confirm that the model has statistically-considerable predictive skill for lead times up to 30 minutes.

In Part II, we introduce two papers on localizing lightning strike points. In the first paper, presented as Chapter 3, we investigate the question of how preinstalled voltage sensors on transmission lines can be used to localize lightning flashes. In particular, we focused on devising a machine learning model that can estimate the 2D geolocation of lightning

strike points from real-time values of lightning-induced voltages measured by two sensors on the transmission line.

The second paper, presented as Chapter 4, discusses how hybridizing machine learning with advanced electromagnetic localization techniques can help lower the number of required sensors for lightning localization to only one. To do this, the data on the incident electric field are recorded by a single remote electric field sensor. Electromagnetic time reversal is used to convert the time-domain electric field signal to a sequence of 2D electric field patterns in the region of interest. Then, pretrained image recognition Convolutional Neural Networks (CNNs) are used to extract a feature vector from EMTR outputs. Finally, a feed forward neural network is used to map the feature vectors to the X-Y position of the incident lightning flash. The paper also explains how the proposed algorithm can be effective in localizing RF sources other than lightning impulses.

Chapter 5 presents a summary and conclusions of the thesis. Potential development threads are also discussed for the future studies.

Chapter 2

Paper title: Nowcasting lightning occurrence from commonly-available meteorological parameters using machine learning techniques[1]

List of authors: Amirhossein Mostajabi, Declan L. Finney, Marcos Rubinstein, and Farhad Rachidi

Author contributions: A.M. conceived the study, developed the model and ran the classifications. D.L.F. advised on interpreting the results and model evaluation strategies. M.R. and F.R. supervised the study and contributed to the results analysis. A.M. led the manuscript preparation with contribution from D.L.F.; all co-authors reviewed the manuscript.

Abstract: Lightning discharges in the atmosphere owe their existence to the combination of complex dynamic and microphysical processes. Knowledge discovery and data mining methods can be used for seeking characteristics of data and their teleconnections in complex data clusters. We have used machine learning techniques to successfully hindcast nearby and distant lightning hazards by looking at single-site observations of meteorological parameters. We developed a four-parameter model based on four commonly-available surface weather

[1] Postprint version of the article published in Nature Partner Journal Climate and Atmospheric Science (DOI: https://doi.org/10.1038/s41612-019-0098-0)

variables (air pressure at station level (QFE), air temperature, relative humidity and wind speed). The produced warnings are validated using the data from lightning location systems. Evaluation results show that the model has statistically-considerable predictive skill for lead times up to 30 minutes. Furthermore, the importance of the input parameters fits with the broad physical understanding of surface processes driving thunderstorms (e.g., the surface temperature and the relative humidity will be important factors for the instability and moisture availability of the thunderstorm environment). The model also improves upon three competitive baselines for generating lightning warnings: (i) a simple but objective baseline forecast, based on the persistence method, (ii) the widely-used method based on a threshold of the vertical electrostatic field magnitude at ground level, and, finally (iii) a scheme based on CAPE threshold. Apart from discussing the prediction skill of the model, data mining techniques are also used to compare the patterns of data distribution, both spatially and temporally among the stations. The results encourage further analysis on how mining techniques could contribute to further our understanding of lightning dependencies on atmospheric parameters.

2.1 Introduction

Lightning is responsible for human injuries and fatalities, the death of livestock, and house and forest fires [1]–[3]. It is also a major source of electromagnetic interference and damage to electronic circuits, buildings, and other exposed man-made structures such as transmission lines, wind turbines and photovoltaics. Based on the reports for 1023 fatalities associated with natural hazard processes in Switzerland during the period from 1946 to 2015, more than 16% of the cases were caused by lightning, making it the second most frequent cause of loss of life among the natural hazards in Switzerland [7]. Furthermore, lightning is reported to have an adverse impact on the aviation industry due to hazard to outdoor ramp operations, such as aircraft fueling, baggage handling, food service, and tug operations. In space centers, lightning is also a danger to fuel crews, ground operations and rocket launch operations [9], [10]. Lightning is also a major cause of damage to wind turbines, one of the fastest growing sectors of renewable energy production, causing transient surges and overvoltages in the power grid, inducing interference in control systems and, most importantly, causing

significant damage to the blades and other wind turbine components [11], [12]. The consequences of these events can be very costly due to energy production losses, extra maintenance costs or even loss of operating equipment [38].

Given its noteworthy socioeconomic impact, appreciable attention has been given to accurate lightning prediction.

The widely accepted mechanism for charging in the thunderstorms is the non-inductive mechanism [39], [40]. In this mechanism, charge separation occurs when ice crystals and graupel particles collide in the presence of supercooled liquid water. Charge is transferred between the different types of particles and then the particles separate by weight under the influence of gravity and convective motions. Several past studies have reported on observational findings regarding how charge structures relate to different lightning types over a wide range of convective regimes. For example, analyzing nine distinct mesoscale regions of severe storms, Carey and Buffalo [41] found significant and systematic differences in the mesoscale environments of positive and negative storms. They hypothesized that the mesoscale environment indirectly influences CG lightning polarity by directly controlling the storm structure, dynamics, and microphysics, which in turn control storm electrification and ground flash polarity. Since lightning involves complex interactions between many atmospheric and in-cloud processes, it is unsurprising that research into that phenomenon continues to generate a wide range of approaches for its prediction. Many studies have implemented sophisticated electrification physics within cloud-resolving numerical models [6], [14], [16], [17], [42]. For example, Fierro et al. [43] implemented an explicit electrification and lightning forecast module within the Weather Research and Forecasting (WRF) Model which includes in-cloud, non-inductive and inductive collisional charging, an explicit elliptic solution of the 3D component of the ambient electric field, and two discharge parameterizations [18] On the other hand, some studies employ a simpler and practical approach of parameterization to allow for useful lightning forecasts without the need for adding electrification subroutines to cloud-resolving models [19], [20], [22], [24], [44], [45]. For example, Lynn et al. [25] implemented a dynamic forecast scheme for both cloud-to-ground (CG) and intracloud (IC) lightning flashes based on the electrical potential energy parameter derived from the WRF cloud-resolving model. More recently, Tippett et al. [26]

used the product of convective available potential energy (CAPE) and precipitation rate as a proxy to predict cloud-to-ground (CG) lightning over the United States. The prediction scheme showed significant skill for lead times up to 15 days for predicting the number of flashes and their spatial extent as well as the lightning/no lightning maps.

Apart from lightning diagnostic schemes producing bulk flash metrics such as average lightning activity with useful skill up to several days forecast, some studies have focused on assessment of the lightning threat in the very near future and providing early warning by nowcasting individual flashes and/or the onset of lightning within the storm. Lightning warning systems in parks, sport complexes, schools, local government buildings, airports, space centers, etc. benefit from the output of such lightning nowcasting schemes to give decision-makers enough time to make take the necessary safety precautions for staff and visitors, stop lightning-sensitive operations, and protect equipment. For example, Mecikalski et al. [46] introduced an integrated 0-1 hour first-flash lightning nowcasting scheme. By merging the satellite and radar systems with numerical models, the lightning forecast is made 30-45 minutes before rainfall occurs. Chambra et al. [47] implemented a cloud-to-ground (CG) lightning probability and binary occurrence or non-occurrence forecast in 20-km grid boxes for 2-hour periods over the continental United States. This lightning guidance product is provided in the framework of the Localized Aviation MOS Program (LAMP) and it is issued at hourly intervals. Recently, Meng et al. [48] developed an early lightning warning system by integrating observational data from radar, satellites, lightning detection systems, ground electric instruments and sounding instruments with synoptic pattern forecasting products, and numerical simulation of the electrification and discharge model. The system is able to provide lightning activity potential and warning products for the upcoming 0-1 h. Seroka et al. [49] used radar-derived parameters, namely the isothermal reflectivity and Vertically Integrated Ice (VII), as the proxy to nowcast lighting over the Kennedy Space Center. Despite several different approaches applied to the important problem of lightning nowcasting and early warning generation, the complex processes and large number of parameters involved in the problem lends themselves to the potential application of a machine learning approach.

Knowledge Discovery in Databases (KDD) is an interdisciplinary area focusing on the process of discovering meaningful correlations, patterns, trends, and on extracting useful knowledge

by mining large amounts of data [50]. KDD has become a powerful tool for turning data into useful, task-oriented knowledge in a wide variety of fields such as business intelligence, marketing or genetics and it has contributed to several of the most recent breakthroughs [50]–[55]. In atmospheric science, enormous proliferation of databases from remote sensing platforms and global-scale earth system models provides a large flow of data [56]. The availability of very large volumes of such data has provided great opportunities for the big data-spun revolution to happen in atmospheric science [56]. Machine learning algorithms such as KDD techniques could give computers the ability to learn a skill (such as making predictions) from sets of archived data and to apply the skill on new data. While conventional algorithms depend on developers entering reams of regulations and principles [57], forecasters and researchers have mixed machine learning with atmospheric science aiming towards improving the communities' prediction skills for multiple weather-related phenomena at different scales [58]. For example, Manzato et al. [59] presented a neural network ensemble forecast for hail in Northeastern Italy. Gagne et al. [60] used machine learning models to predict the probability of a storm producing hail and the radar-estimated hail size distribution parameters for each forecast storm. Herman et al. [61] explored the internals of some regression and tree-based models and what physical and statistical insights they reveal about forecasting extreme precipitation from a global, convection-parameterized model. Lagerquist et al. [60] described a machine learning system that forecasts the probability of damaging straight-line wind for each storm cell in the continental United States. Karstens et al. [62] developed a human–machine mix for forecasting severe convective events. As an application in lightning nowcasting and early warning systems, this paper examines how the mining of basic atmospheric datasets can be used to explore correlation patterns between lightning incidence and atmospheric data and, thus, for nowcasting of lightning activity. To achieve this, a machine-learning-based model is trained to nowcast whether or not there would be any lightning incidence inside a specific region up to 30 minutes in advance, given the real-time measured values of four meteorological parameters which are relevant to the mechanisms of electric charge generation in thunderstorms [63]–[65], namely the air pressure at station level (QFE), the air temperature 2 m above ground, the relative humidity, and the wind speed.

Although the selected meteorological parameters do not necessarily represent upper level meteorology within the thunderstorm charging zone, they are indicators of low-level factors involved in thunderstorms. In addition, they can also be more reliably and continuously measured than many upper-atmosphere parameters that could be more closely linked to lightning generation.

Some lightning predictive schemes use operational radars and satellites to detect storm initiation and development, perhaps also aided by Convective-Allowing Models, and can provide calibrated thunder guidance up to at least a day in advance. For example, the High-Resolution Ensemble Forecast (HREF) Calibrated Thunder guidance produces probabilities that represent the likelihood of at least one cloud-to-ground (CG) lightning strike within 12 miles (20 km) of a point location over a 4-hour forecast period [66]. This guidance generates forecasts over the Continental United States using a rolling 4-hour window out to 48 hours. Using commonly-available surface data makes the warning system in this study independent of external sources of data such as numerical model outputs, satellite and radar. In this regard, the proposed approach could benefit the current lightning predictive schemes. Two potential contributions are: (i) While satellites can provide broad nowcasting information for people, the ML approach provides an opportunity for much more localized forecasting and alerts, and this could be facilitated for users through a web interface where they can upload their own data. Furthermore, the method can provide information in areas where radars are not present, where weather forecaster resources are limited, or where nowcasting is not in operation, for instance in isolated areas in low-income countries in Asia, South America, and Africa [67]. In fact, the method can be applied to any weather station (with extended data records to ensure appropriate training samples) to give localized forecasting, independent of the availability of other resources. (ii) The input data are not subjected to the typical scan cycles, limited forecast steps, or processing and post-processing delays. Indeed, the predictors used are commonly available in real time and they have high temporal resolutions. Given this, the proposed ML model is able to provide early lightning warnings with short lead time ranges (0-30 minutes), as opposed to methods that have forecast periods measured in hours. Such warnings could contribute to the reduction of air traffic congestion at airports and to the decrease in disruptions to energy generation from wind turbines farms.

In supervised learning, algorithms are designed to learn from a given dataset with an already known output (training set). After the learning phase, predictions are made on new datasets (testing sets). In this study, we implemented the proposed nowcasting scheme using data from 12 meteorological stations in Switzerland between 2006-2017 (see Table A-1 for the list of the selected stations). The stations were selected based on two criteria, namely (i) the availability of both meteorological and lightning activity data during the study period, and (ii) the fact that they are well distributed among different ranges of altitude and terrain topographies. Among the stations, 6 are located in an urban area inside cities with altitudes ranging between 273 and 776 m above sea level and one is the weather station at the Geneva airport. Five out of the 12 stations are located in mountainous regions with 3 of them having an altitude of more than 1000 m above sea level. A common feature of the stations in Switzerland is the presence of nearby topography, which increases the probability of storms being initiated near them. While this makes the results presented here directly relevant to locations that experience topographically induced thunderstorms, further work is needed to evaluate the skill of the approach in environments with different triggering mechanisms.

At each of these single-site meteorological stations, we first formed a tabular database with each row containing the observations during a specific time window with a granularity of ten-minutes. In each row, the corresponding meteorological measurements are used as the predictors (also called features) and the recorded lightning activity is used as the response. Once the database was formed, pattern recognition and data mining algorithms were employed to identify regularities between predictors and responses using a portion of the data which, as mentioned above, is called the training set. The model could then use the explored correlations for nowcasting the long-range lightning threat (within a circular area of 30 km radius around the meteorological station) for the unseen cases (testing set). The model predictions and observations are then compared to evaluate the model's prediction skill. The evaluation results are presented by means of four common indices in forecasting rare events described in Table 2-1 namely the Probability of Detection (POD), False Alarm Ratio (FAR), Critical Success Index (CSI), and Heidke Skill Score (HSS).

Detailed information on the data acquisition, database formation, training and testing procedures, performance evaluation process, and the selection and generation of the applied machine learning algorithm in this study are presented in Section 2-4.

Table 2-1 Parameters (with acronyms and definitions) used in performance evaluations of models.

Parameter		Full name	Definition	Equation
Variables in the contingency table	**H**	Hit (or true positive)	Number of observed lightning-active samples correctly identified by the classifier	
	M	Miss (or false negative)	Number of observed lightning-active samples falsely classified as lightning-inactive by the classifier	
	FA	False Alarm (or false positive)	Number of observed lightning-inactive samples falsely classified as lightning-active by the classifier	
	C	Correct rejection (or true negative)	Number of observed lightning-inactive samples correctly identified by the classifier	
Selected evaluation metrics	**POD**	Probability of Detection (or true positive rate)	Proportion of observed lightning-active samples correctly identified by the classifier	$\frac{H}{H+M}$
	FAR	False Alarm Ratio	Proportion of observed lightning-inactive samples falsely classified as lightning-active by the classifier	$\frac{FA}{H+FA}$
	CSI	Critical Success Index (or threat index)	Ratio of successful event forecasts to the total number of event forecasts that were either made (H+FA) or needed (M)	$\frac{H}{\mathrm{H+FA+M}}$
	HSS	Heidke Skill Score	Total number of correct forecasts (H+C) minus the correct random forecasts divided by the total number of forecasts (H+M+FA+C) minus the correct forecasts due to chance	$\frac{2(H.C-FA.M)}{((H+M)(M+C)+(H+FA)(FA+C))}$

2.2 Results

2.2.1 Machine Learning Model performance for long-range lightning activity at 12 stations in Switzerland

The database consists of the observations of 4 meteorological parameters with a granularity of ten minutes recorded at 12 selected meteorological stations in Switzerland over a time period ranging from 2006 to 2017. In order to see how far in advance the lightning alarms could be generated, three ranges for lead time were investigated: (i) 0-10 minutes, which corresponds to imminent lightning activity, (ii) 10-20 minutes, and (iii) 20-30 minutes. At each station, the data are labeled according to the presence or absence of long-range lightning activity (within 30 km distance from the station) and with respect to the three aforementioned lead-time ranges to form Subsets 1-12 (see Table A-3 for the list of studied subsets and Section 2-4 for the description of the data gathering).

The list of stations with their geographical information is presented in Table A-1. Among the selected stations, the Säntis and Monte San Salvatore stations have been of great interest for lightning studies in the literature [68]–[71] The atmospheric data at these two stations were gathered in Subsets 1 and 2. Figure 2-1 shows the visualization of the data at the Säntis and Monte San Salvatore stations (Subsets 1 and 2, respectively) for the 0-10 minute lead-time range where the inter-relation between the considered parameters is illustrated. A recorded observation at the start of a ten-minute interval is labeled according to long-range lightning activity in that interval as either a 'lighting-inactive' sample (without any long-range lightning activity) or a 'lighting-active' sample (with at least one long-range lightning activity recorded). Figure 2-1 shows the probability density estimates of the surface pressure and surface temperature for lighting-inactive and lighting-active samples in Subsets 1 and 2 using a kernel smoothing function. Although the visualized data suggest the existence of different distribution patterns among the two classes, they seem to be mixed together at both stations, which makes it difficult to explicitly extract any classification criteria. Alternatively, the developed machine learning model (the ML model) is used to recognize the patterns. The machine success is evaluated in two ways: (i) By measuring how accurately it can classify the data into two distinct classes (lighting-inactive or lighting-active), and, (ii) by investigating how it can improve upon three competitive baselines, namely the persistence forecasting method, the electrostatic field method, and a scheme based on CAPE threshold.

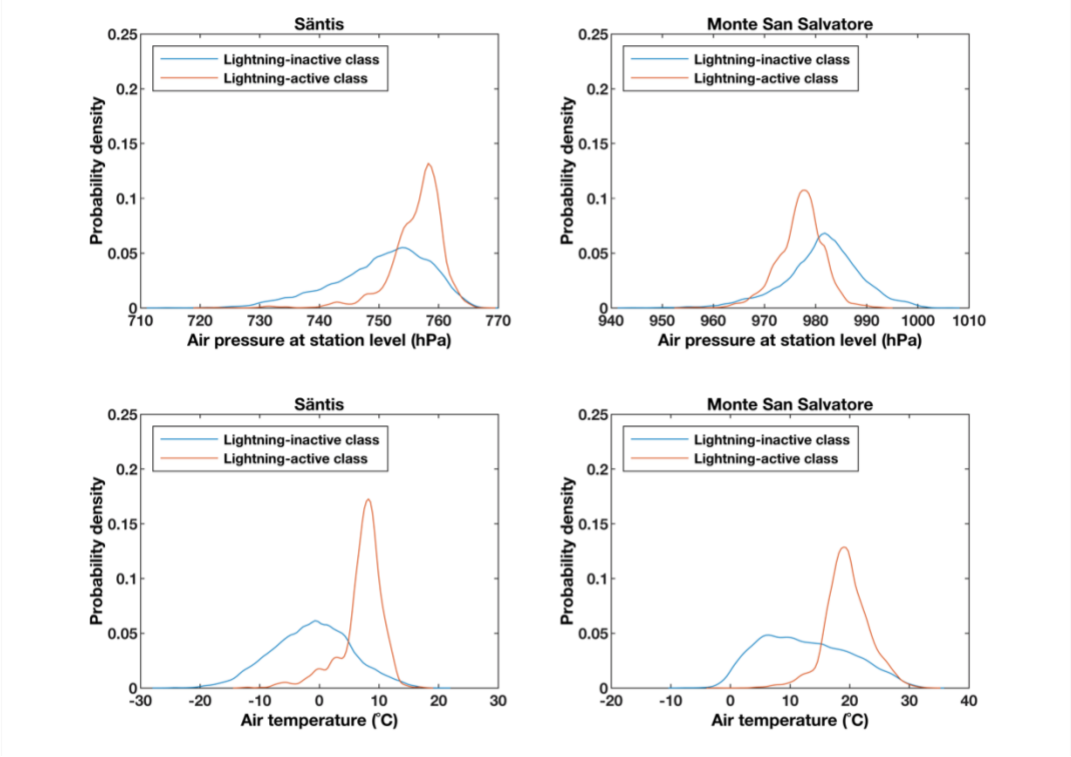


Figure 2-1 Probability density estimates of the surface pressure and surface temperature for lighting-inactive and lighting-active samples in Subsets 1 and 2 using a kernel smoothing function. The Subsets 1 and 2 include observations of four meteorological parameters respectively at Säntis and Monte San Salvatore stations from 2006 up to 2017 with the granularity of ten minutes. The corresponding lead time range is 0-10 minutes. Thus, a recorded observation at the start of a ten-minute interval is labeled according to lightning activity in that interval as either a 'lighting-inactive' sample (without any long-range lightning activity) or a 'lighting-active' sample (with at least one long-range lightning activity recorded).

To generate forecasts based on the data, the PERSISTENCE model assumes that in every ten-minute interval, the forecast is the same as the observation in the previous interval. So, if there was lightning in the previous ten-minute period, then, according to the PERSISTENCE model, there will be lightning in the next period too. The persistence forecast represents a realistic and applicable competitive option against the ML model, since the preceding lightning activity is continuously stored by lightning location systems around the world. If one wanted to predict whether there will be lightning in the next ten-minute interval based on the PERSISTENCE model, it would be possible to check the online lightning and thunderstorm detection networks and see if there was any lightning activity recorded in the previous time interval. Note that the prediction for the second and third lead time ranges was carried out using the observed lightning in the ten-minute interval prior to the forecast time, which is the same interval that was used to forecast for the 0-10 minute lead time.

Electrostatic field readings have shown to be affected both before and during the thunderstorm due to the approaching charge centers, their rearrangement inside the thunderclouds, as well as cloud electrification and rearrangement of space charge in the atmosphere [72]. Some previous studies have used these variations to forecast approaching lightning activity [28], [72], [73]. The electrostatic field method (the E-FIELD model) used in this study is based on detecting when the vertical electrostatic field ($E(z)$) exceeds a specific threshold to issue the warning. The corresponding threshold used by the E-FIELD model for each subset and lead time is defined in a way that the CSI is maximized. The choice of CSI as the optimization criterion is mainly based on its inherent consideration of both, POD and FAR and, hence, it is suitable when a trade-off between these two is desired. Among the selected stations, the Säntis station was the only one equipped for vertical electrostatic field measurements. These data were available from August 2016 to July 2018. Hence, the performance results for the E-FIELD model are presented in Figure 2-2 only for the Säntis station and from August 2016 to December 2017 to also match with the time period of this study.

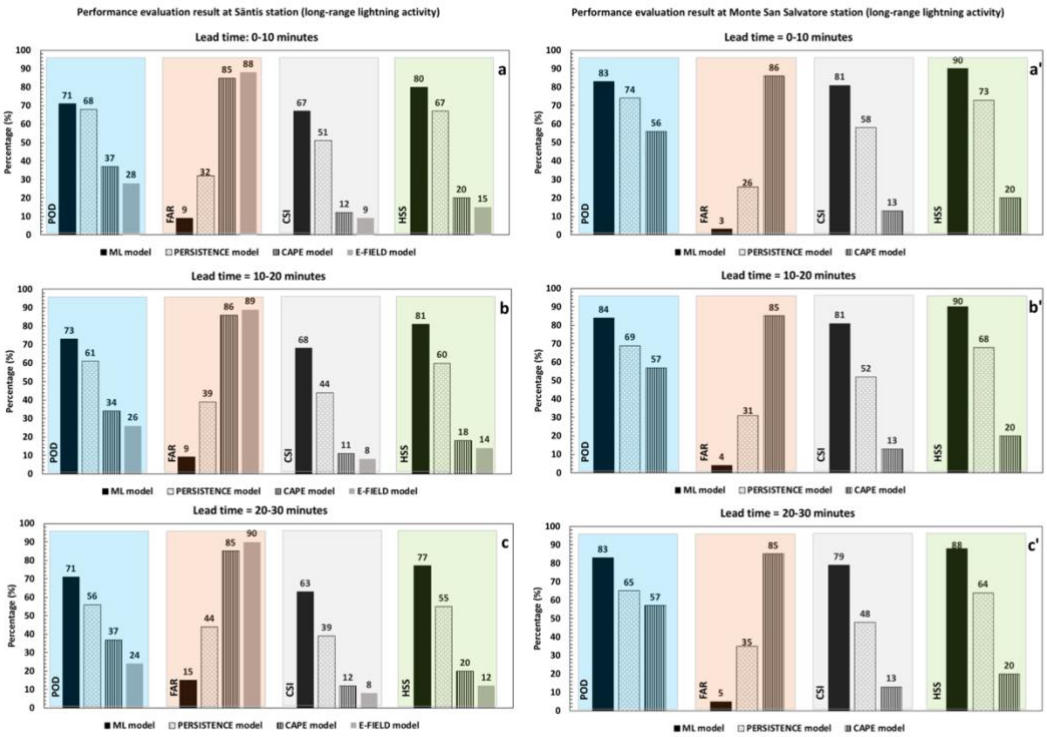


Figure 2-2 Evaluation of the skill of warnings of long-range lightning activity for three ranges of lead times at a, b, c Säntis (Subset 1) and a', b', c' Monte San Salvatore (Subset 2) station. The results from ML model are shown as solid black columns, results from PERSISTENT model are shown as columns filled with dot patterns, results from CAPE model are shown as columns filled with vertical lines, and results from E-FIELD model are shown as solid grey columns (POD: Probability of Detection, FAR: False Alarm Ratio, CSI: Critical Success Index, HSS: Heidke Skill Score).

Not having direct measurements of the electrostatic field at stations other than the Säntis, we considered a CAPE model in addition to the PERSISTENCE model as the competitive schemes for the proposed ML model. The CAPE model uses a threshold of Convective Available Potential Energy (CAPE) to assess the risk of lightning. Assessing the level of CAPE which is well understood to be reflective of environments favorable to the development of lightning [21], [45], [74], the CAPE model would be a more objective comparative scheme. Similar to the E-FIELD model, the corresponding threshold for each subset and lead time is defined in a way that the CSI is maximized. The CAPE model's performance has been found to have low sensitivity to the selected threshold. For example, at the Säntis and Monte San Salvatore stations, the CSI changed respectively less than 7.5% and 5% for a ± 30% change in the threshold value at each of the lead time ranges. As mentioned in the Data Section, in this study, the CAPE data are retrieved from the ERA5 hourly reanalysis data on single levels provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). The temporal resolution of the available CAPE data was originally 1 hour. To make it consistent with the granularity of the meteorological parameters, which is ten-minutes, one possibility was to interpolate the data for missing ten-minute time frames. While this interpolation would increase the number of available samples to be used by the CAPE Model, it might also negatively affect its performance skill due to possible inaccurate interpolation of samples for this highly variable and hard to measure parameter. Consequently, the CAPE model was only tested on the ten-minute intervals for which the CAPE data were originally available without any temporal interpolation.

The performance results for the ML model and the three baselines are presented in Figure 2-2 for the Säntis and Monte San Salvatore stations. The four selected evaluation metrics shown in the figure are introduced in the Methods section and a summary is given in Table 2-1. Based on the results at both stations, the ML model has consistently higher scores than the three competitive models for all three lead time ranges. The results from Figure 2-1 show clear differences in both the absolute recorded values and the distribution patterns of data at the

two stations. However, the performance results for the ML model at the two stations in Figure 2-2 exhibit similar evaluation results. In other words, although the atmospheric data for both, lighting-inactive and lighting-active classes have different ranges of values and are distributed differently at the two stations, the ML model was able to learn from the local patterns and predict with reasonable accuracy at both stations. Indeed, the results for the POD index at both stations reveal that more than 71% of the lighting-active samples are classified accurately.

It is worth noting that the accuracy of classification is not only important as a measure of prediction skill but it also accounts for how successful the ML model is in finding the complex correlations in the data. Low accuracy in the model does not necessarily imply that there is no correlation in the data since the low accuracy might be attributable to deficiencies in the model. High model performance, in contrast, is an indication of a strong correlation pattern between the predictors and the response and also of the capability of the model to recognize such patterns.

Using the feature reduction method, the impact of excluding individual variables from the ML model input was investigated on each metric for the two subsets. The sensitivity is calculated using the following equation,

$$S = \frac{m_{wo} - m_w}{m_w} \times 100 \qquad (2\text{-}1)$$

where S is the sensitivity of each one of the four metrics to a specific feature, m_{wo} and m_w are, respectively, the values of the metric with and without the feature included in the ML model. A positive value for S means that the associated index would increase if that feature is not included in the study. If the majority of the indices show positive sensitivity to a particular feature, one could get better results if the feature is excluded from the predictors list. Looking at the sensitivity results for the long-range activity and for the three investigated ranges of lead time, no such feature could be found. This suggests that the best result is the one with all meteorological variables included.

The results in Figure2-2a,b,c for the CAPE model show a similar performance for all three lead time ranges compared to the ones for the E-FIELD model. Both models show low POD and

very high FAR values at all three lead time ranges. On the other hand, the results shown in Figure2-2a, for example, indicate a good prediction skill for the PERSISTENCE model for the imminent threat warning. This noticeable lack of skill for the E-FIELD model at such lead time ranges has already been reported in the literature. For example, Aranguren et al. [73] analyzed the skill of a lightning warning system based on the measurement of the electrostatic field in Northeastern Spain. Using a threshold-based lightning warning system, the FAR was around 90%, and the POD was between 10% to 70% based on the low and high threshold levels. Analyzing 7 storms, they also reported that the best achieved lead time for the forecast was less than 6.5 minutes.

The results in Figure2-2 show how the model's nowcasting skill changes when a longer forecast time is required. This is important to give sufficient time for the safety actions to be undertaken. Comparing results in Figure2-2a' and Figure2-2c' shows that the PERSISTENCE model is more sensitive to the increase of lead time compared to the CAPE and the ML models. The results from the PERSISTENCE model show 12% drops in each, POD, CSI, and HSS, and a 12% increase in FAR compared to the 0-10 minute lead time, whereas the performance of the CAPE and the ML models is not affected by an increase in the lead time. In other words, the results in Figure2-2 suggest that, although looking at the preceding lightning activity records (i.e. what is done by the PERSISTENCE model) is good enough to warn for the very near future lightning threat (0-10 minute lead time), this would not be reliable in most of the applications where longer forecast times are needed. On the other hand, the ML model ensures that the accuracy of its warnings up to 30 minutes in advance will be maintained.

While nowcasting rare events, special emphasis needs to be given to the no-event cases, which dominate the dataset. Since no-event instances represent the majority of the samples, lacking the skill to correctly classify them would lead to a large number of false alarms. According to the results for FAR in all subsets of Figure2-2, the ML model is seen to perform much better than the other models concerning the correct rejection of no-events.

Although the sensitivity analysis described in Eq. 1 gives insights into each predictor's importance in the ML model performance, its value often varies from one metric to the other, which makes it difficult to rank the predictors in the sense of their overall importance. To bridge the gap, the predictors importance estimates calculated by the ML model could be

used to rank the predictors. As explained in Section 2-4, the learning process was done by growing an ensemble of decision trees. To grow each of these trees, the ML model starts from the root and creates decision nodes and branches by calculating the split gains. The model would then be able to estimate the predictor importance by summing the changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes. The predictors importance associated with each split is computed as the difference between the risk for the parent node and the total risk for the two children [75] To compare the predictor rankings due to their importance at different stations, the importance estimate of each predictor is computed relative to the sum of the estimates for all predictors in that study, calculated as

$$imp_i(\%) = \frac{imp_i}{\sum_{j=1}^{4} imp_j} \times 100 \qquad i = 1, \dots, 4 \qquad (2\text{-}2)$$

where imp_i is the absolute value of the importance estimate for predictor i. These predictors importance estimates are reported in Figure2-3 for studies at each station. All statistics and results presented in this figure are presented for a lead time of 0-10 minutes. One should note here that the predictors importance does not relate to the model accuracy and it just relates to the importance of each predictor in making a prediction, not whether or not the prediction is accurate.

The result in Figure2-3 indicates that at most of the stations, for the prediction of the long-range lightning activity, the variation of the surface pressure, relative humidity, and surface temperature were more important than the wind speed. The dependence of the long-range model performance on these parameters can be explained as follows: The impact of the surface temperature can be attributed to the fact that anomalously high local temperatures are more likely to be associated with instability. This parameter could also be important due to the arrival of the gust front (cooler temperatures), or higher CAPE at the meso-scale. High relative humidity suggests a higher chance of having sufficient moisture supply to generate deep convection, as well as being related to instability of the environment. The surface pressure identifies local troughs propagating through. Alternatively, the increases or decreases of pressure perturbations could be associated with the propagation of a gust front. Results indicate that the wind speed was also found to be useful by the classifier. In the long-

range, it might be important due to the fact that large organized systems may induce strong winds far from the charging regions of the storms. The fact that the predictors importance estimates change from one station to the other might indicate that the models for individual stations may offer new insights into the relevant processes locally.

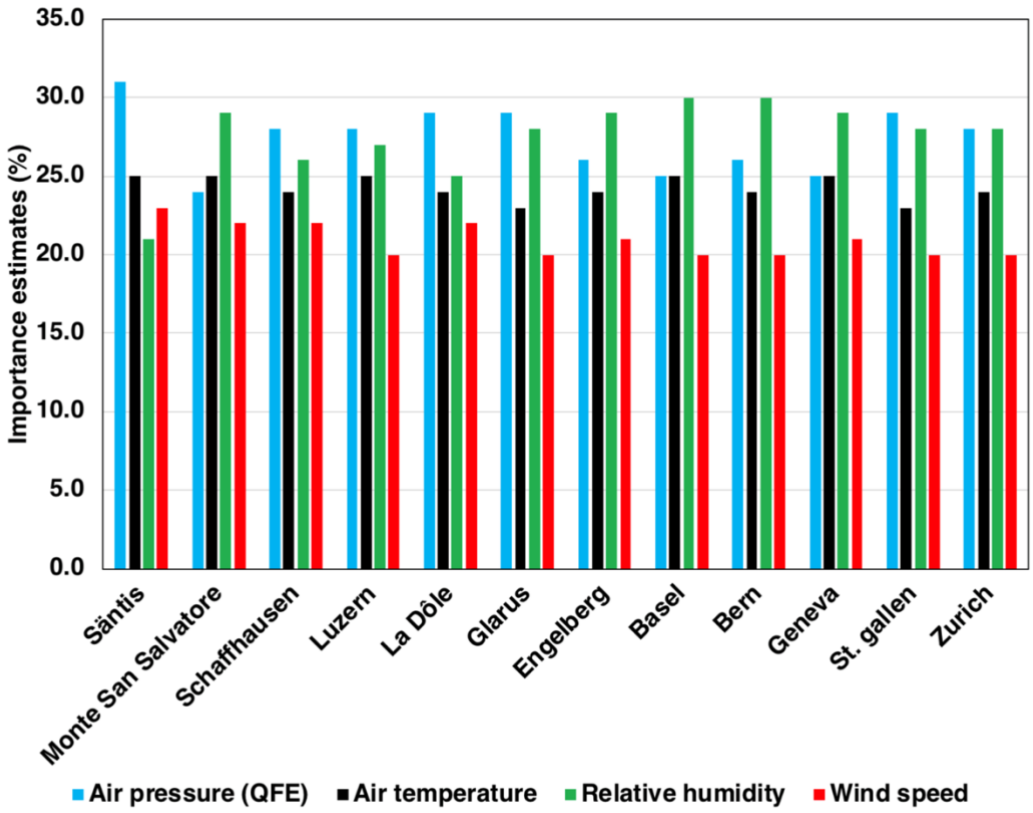


Figure 2-3 Predictors importance estimates for long-range lightning activities. The result includes both studies at individual stations and over all stations. The data from individual stations are standardized according to their local mean and deviation before they are included in the overall study. The presented results correspond to the lead time of 0-10 minutes.

2.2.2 Distribution patterns of data among different stations

As stated in the Introduction, the involved stations are selected from regions with different topographies and altitudes. In order to see how these differences affect the distribution patterns of data among the stations, the probability density functions (PDFs) of each parameter in the lighting-active class for all 12 stations are compared in Figure 2-4. The PDFs are plotted using each of the data Subsets 3 to 14 and with a lead time of 0-10 minutes. The reason for choosing these data subsets is that each one of them includes data from an individual station and it has the same temporal coverage as the others (2006 to 2017). The

choice of a lead time range of 0-10 minutes accounts for the imminent lightning activity at each interval. One should note here that in order for the results to be comparable between stations, the data in the subsets were standardized. The standardizing function shifts the mean of each predictor to zero and scales the predictors by their standard deviations. The distribution patterns of the lighting-active samples for all 12 stations are also visually illustrated in Figures A-1 to A-4. Looking at the PDF plots and the distribution patterns for all stations suggests the existence of two groups: (i) Group A, including 10 stations with their altitude lower than 1050 m above sea level, and (ii) Group B, which includes the remaining two stations with higher altitudes (Säntis and La Dôle). Figure 2-4 shows that the probability density of surface pressure and relative humidity are different in the stations belonging to Group A and Group B. These differences in the densities can be clearly seen in the distribution patterns shown in Figures A-1 to A-4, where the lighting-active class data in Group B are clustered around higher pressure and higher relative humidity when compared to Group A. Looking at Table A-1, one can find the potential reason for this difference in the patterns. Säntis and La Dôle are the sites with the highest elevations in the dataset and are approximately 2000 m and 1100 m higher than the bulk of the other sites. The altitude differences have a considerable effect on pressure, temperature and wind speed and, therefore, we have performed a more detailed investigation on these sites.

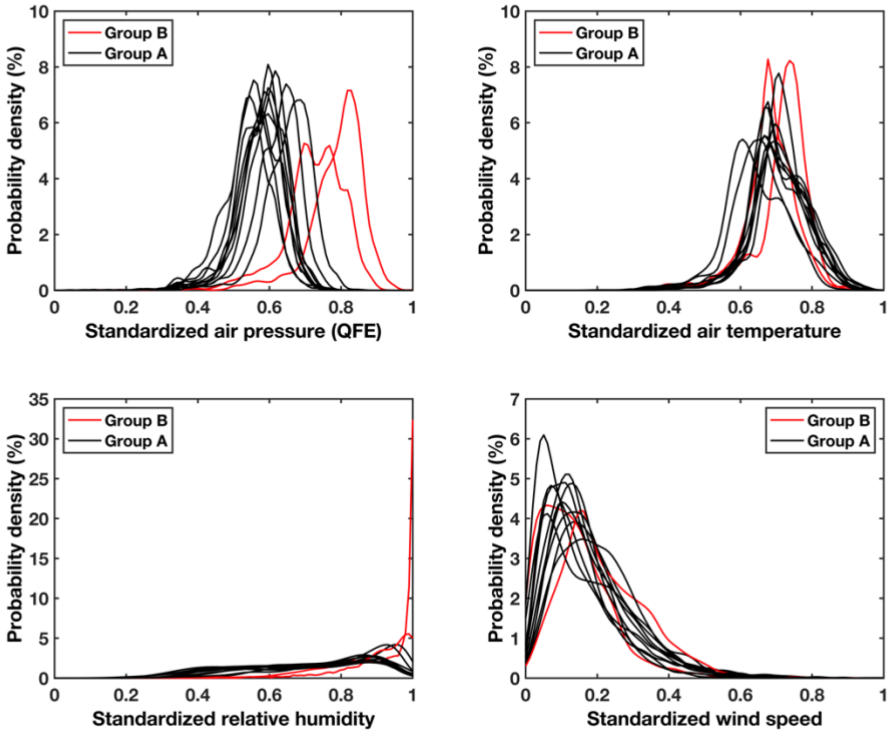


Figure 2-4 Comparison of probability density functions (PDFs) of each parameter in lighting-active class samples of Group A and Group B. The corresponding lead time range is 0-10 minutes. For the sake of comparison, the parameters values at each station are standardized based on the local mean and standard deviation values. The standardizing function puts mean of each parameter at zero and scales the parameters by their standard deviations.

To better understand these differences, the principle component analysis (PCA) is used to explain the distribution patterns of individual features in different subsets. PCA is a data mining tool that transforms a number of interrelated variables into a new set of uncorrelated variables called principle components (PCs), while retaining as much as possible of the variation that exists in the original data [76], [77]. The first principal component retains most of the variations in the data, and each succeeding component accounts for as much of the remaining variability as possible. The Singular Value Decomposition (SVD) algorithm is used to perform the PCA analysis on data from individual stations (Subsets 1 to 12). At each station, the first two Principal Components (PC1 and PC2) are kept. Figure A-5 shows the percentage of total variance explained by each of these two leading PCs. Furthermore, the contribution of each original variable to each principal component is defined by sets of coefficients. Figure 2-5 shows the loadings of each variable on the first two components (PC1 and PC2) for all 12

stations. In each subplot, the two stations with the highest distance from the center of the main cluster are defined using cluster analysis and marked as red dots. Looking at the four subplots, it can be seen that, in all cases, black and red dots correspond exactly to stations of Groups A and B, respectively. In other words, Group B stations show quite different PC1 and PC2 coefficients in all variables compared to Group A. The results suggest different patterns of distribution between the two groups and they confirm what was visually concluded earlier from Figure 2-4 and Figures A-1 to A-4.

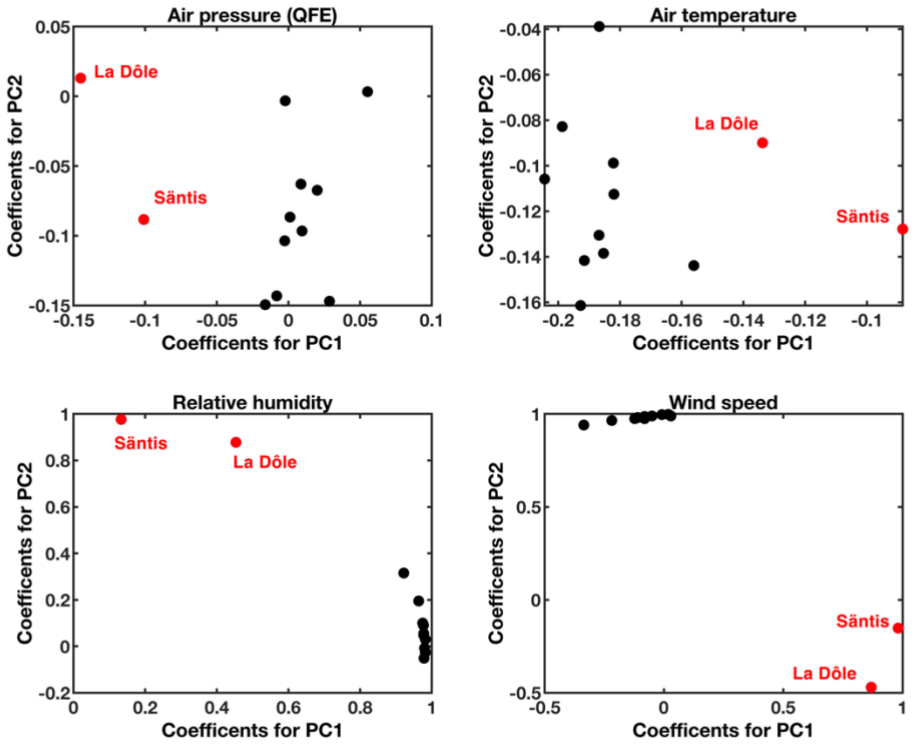


Figure 2-5 The contribution of individual variables to the first and second principle components for all 12 stations during 2006 to 2017. In each subplot, data for two stations with the largest distance from the cluster center are annotated in red and the rest are presented in black. The corresponding lead time range is 0-10 minutes.

2.2.3 Change of distribution patterns over time

Further analysis of the data showed that these pattern variations are not limited to geographical characteristics. Mining the data for all stations on an annual basis using PCA reveals that patterns of distribution for stations belonging to Group B have changed widely from year to year while Group A stations exhibited only a slight change. In this regard, each

of the Subsets 1 to 12 was split into 12 segments, each assembling the data for one year between 2006 to 2017. Then, the PC1 and PC2 coefficients were plotted and the differences in feature contributions were observed from year to year. Figure 2-6 presents the results for one representative station of each group, namely Zurich from Group A and Säntis from Group B for a lead time range of 0-10 minutes. The results for the other two lead time ranges are presented in Figure A-6.

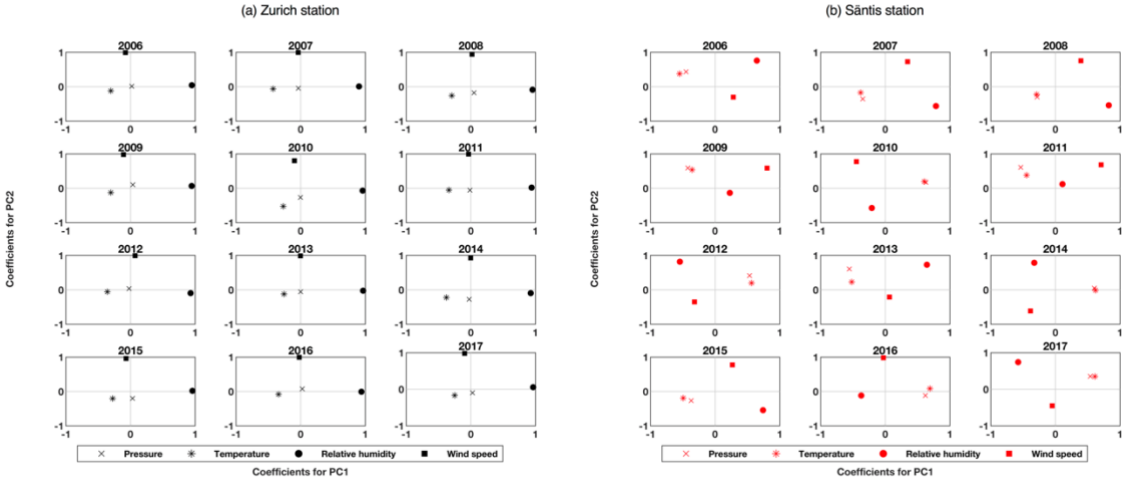


Figure 2-6 The contribution of individual variables to the first and second principle components from 2006 to 2017 for a Zurich station (as an example of Group A) and b Säntis station (as an example of Group B). In each subplot, the horizontal axis is the coefficients for PC1 and vertical axis is coefficients for PC2. The corresponding lead time range is 0-10 minutes.

2.3 Discusion

Early warning systems are useful to help prevent effects of lightning strikes to critical infrastructure, sensitive equipment or systems, and outdoor facilities. Taking advantage of both, the large amount of available data for meteorological parameters and advances in data mining and knowledge discovery, we used KDD techniques to investigate the correlation between lightning and selected meteorological parameters and thus warn against the risk of long-range lightning activity. To do that, the machine was programmed to automatically extract and learn hidden regulations inside previously labeled data in order to predict the labels for unseen data. These regulations could be any information in the data that the machine could use during the training process to learn a target function that best maps the input variables to the output. The evaluation results for 12 locations in Switzerland revealed

that, by looking at values for four principle meteorological parameters, the ML model was able to warn with a reasonable accuracy of future lightning activity up to 30 minutes in advance and in an area of 30 km around the observation point.

For each station, the ML model has been set up to predict lightning activity in three future forecast times: (i) the interval from the present to 10 minutes into the future which corresponds to imminent lightning activity, (ii) the interval between 10-20 minutes into the future, and (iii) the interval between 20-30 minutes into the future. These lead time ranges are likely shorter than the lifetime of many thunderstorms, organized systems or isolated storms. This implies that the ML model is looking at the changes in the local surface conditions leading to the occurrence of lightning within the storm life cycle, and this applies to any kind of storm. Given the fact that isolated storms (short-lived, topographically and diurnally forced) and propagating organized storms (longer-lived) can both lead to lightning, a strength of the ML model is that it is capable of accounting for different situations leading to lightning generation, with the caveat that the different situations need to be discernible in the input data.

As discussed in the Methods section, the main challenge in developing the appropriate predictive scheme was the high imbalance seen between lighting-inactive and lighting-active classes. The situation gets even worse when the lead time is increased. This high level of imbalance not only highlights the need for special techniques in developing the machine learning algorithm, but also requires specific considerations for model performance evaluation. For the considered lightning prediction application, the cost of misclassifying lightning-inactive samples is much less than that for lighting-active ones. The average of PODs for the studied local subsets showed that more than 76% of the long-range lightning threats were correctly predicted by the ML model up to 30 minutes in advance.

Unlike some lightning warning systems that are based on data from lightning detection networks [28], [29], the ML model provides a tool for building lightning nowcasting schemes without using nearby or preceding lightning data as the precursor of the imminent threat. In other words, it does not rely on the first detection of the lightning for generating the warnings. Instead, it uses the lightning location systems' data for labeling the archived data

and thus training the model with historical data. Once trained, the model does not need such data to predict lightning occurrence risk in future time windows.

Compared to other conventional warning techniques, the machine learning approach also offers automatic extraction of regulations and it does not require advanced background knowledge about the predicted task. Although the ML model eliminates the need for manually adjusting the thresholds or prediction criteria due to automatic detection of regulations, it is susceptible to changes that might occur in the environment and that may affect the extracted rules by the model during the training process. For example, an increase in the number of tall buildings nearby and weather and climate changes might alter the dependencies found by the model between predictors and response. On the other hand, the ML model could be readily updated to the new situation with the required periodicity using new achieved data.

Although some vulnerable sites such as airports and space centers need to warn for total lightning activity (both CG and IC flashes), being able to split the warnings for each type of flash would, in general, be a desirable feature in an early lightning warning system. The ML model presented in this study does not differentiate between the various types of lightning flashes. The reason is that the available data from lightning location systems used to train the ML model have no distinction between cloud-to-ground (CG) and intra-cloud (IC) flashes. However, one could evaluate the skill of warnings for different types of lightning activities by training the model separately using data from each flash type. The application of the method to the task of forecasting different types of lightning is beyond the scope of this work and will be dealt with in future research.

Although no data were available for direct comparison with other state of the art lightning warning approaches based on radar or satellite, their performance results are reported in the literature. For example, Seroka et al. [78] investigated the use of optimized radar-derived predictors along with IC flashes to predict CG lightning over the Kennedy Space Center. The values obtained for POD, FAR, CSI and the average lead time for the two leading predictors of CG flashes were, respectively, 78%, 35%, 55%, and 2.4 minutes for IC as the predictor, and 78%, 46%, 47%, and 6.4 minutes for a radar reflectivity value of 25 dBZ at -20 ˚C as the predictor. Taking advantage of the wide range of input data including radar and satellite

observations, ground measurements of the electric field, data from lightning location systems, sounding instruments with synoptic pattern forecasting products and a two-dimensional charge-discharge model, Meng et al. [79] made an integrated system to predict lightning within the upcoming 0-60 minutes. The nowcasting system gives the probability of lightning occurrence in grids of 1 km x 1 km and in 15-minute lead time steps. The probabilities above 25% lead to early warnings. The verification results for the issued warnings during 16 thunderstorms in Beijing and the surrounding area show that the mean values for the POD score drop from 49% to 37%, the FAR score increases from 67% to 77% and the CSI score decreases from 24% to 16% while coming from the minimum (0-15 minutes) to the maximum lead time range (45-60 minutes).

The primary goal of this study was to examine the effectiveness of using single-site meteorological observations to train machine learning algorithms for nowcasting lightning and warning against the lightning threat. Secondarily, data mining techniques were used to further investigation of possible geographical and temporal dependencies in the data.

Explaining the reasoning behind the machine decisions to humans can be difficult because they often do not make use of the same intermediate abstractions that humans use [80]. This kind of issues, in turn, could be addressed by the interpretable machine learning. Interpreting machine learning aims to increase the model transparency and thus make it more useful and trustable by giving explanations for model predictions. For example, applying interpretation techniques could reveal an explanation of the hidden trends found by the ML model once it is implemented at stations in different climate zones. Comparing the basis for model prediction in different stations would shed light on whether there are different regulations that correlate meteorological parameters to lightning activity or not. Such kind of findings would be obviously more valuable when more relevant atmospheric parameters to lightning initiation available in satellite and radar data are used as the predictors. This would enable the machine learning approach to better contribute to a further understanding of lightning and atmospheric interactions.

The use of surface measurements as input data in this study does not put any limitation on other relevant parameters to be used by the ML model. Large amounts of atmospheric data are available from numerical model outputs, atmospheric soundings, satellite and radar

observations. Given this and also the fact that lightning activity is now readily detected with high spatiotemporal resolution by means of space-borne instruments and ground-based lightning location systems, an extensive amount of work is in progress by the authors to apply the machine learning approach to provide lightning predictive schemes (i) capable of estimating the flash rates as well as the lightning threat itself, (ii) with large spatial coverage, and (iii) with good skill for lead times up to 24 hours. However, for the first-stage research presented in this study, we restricted our efforts to the selection of surface data since we wanted the scheme to be easy to implement and widely applicable to a variety of vulnerable sites. In fact, the idea behind the choice of input variables for this early warning scheme in this study was to use types of predictors that are commonly-available, that have a high temporal resolution, and that are easy and fast to retrieve in real time.

Rapid increases in total lightning activity have been demonstrated to be a precursor for the occurrence of severe weather at the ground [81]. As a potential application, the proposed ML model could be trained to provide an early indication of severe weather events other than lightning at short time scales. Such a model could be evaluated alongside the lightning jump algorithm [82].

Even though we have not used real time data in this study, the selected meteorological parameters are available from Personal Weather Stations (PWS) with refresh rates of less than 2 s. Being small, precise, and easy to install and operate, individuals often own these devices and upload the data to an online platform aiming to improve weather forecasting [83], [84]. Given the fact that these sensors measure all four predictors needed by the ML model, it could be easily integrated into these devices. In this regard, PWSs would be converted to an accurate early lightning warning system at any arbitrary point of interest while keeping their main functionality as weather stations.

2.4 Methods

2.4.1 Data gathering

The dataset used in the ML model consists of data used as predictors, namely available meteorological data (air pressure, air temperature, relative humidity, and wind speed) and

lightning activity data as the response. Vertical electrostatic field data measured by the E-field mill device at one of the stations and Convective Available Potential Energy (CAPE) are also gathered and used as predictors and competitive baselines in this study.

The spatial and temporal coverage of the study was set according to the availability of both atmospheric and lightning data. Furthermore, and considering the effects of the terrain topography and topological effects on the lightning incidence [85], the stations were selected to be properly distributed among different ranges of altitude and terrain topographies. The time interval of the study was set to 2006 to 2017 (12 years) and 12 stations in Switzerland were selected. Note that the vertical electrostatic field data were only available at the Säntis station from August 2016 to July 2018. In order to use these data, the time coverage at that station was therefore extended up to July 2018. More information about the selected stations is presented in Table A-1. The used data are described in what follows.

In this study, data on surface air pressure at station level (QFE), air temperature 2 m above the ground, relative air humidity, and wind speed were obtained from the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) online database and they were measured by the automatic monitoring network of MeteoSwiss (SwissMetNet). SwissMetNet now comprises 160 measurement sites equipped with high-precision measurement instruments and state-of-the-art communication technology. The measurement instruments and respective sources of errors are listed in Table A-4. All devices in SwissMetNet comply with the standards of the Word Meteorological Organization (WMO) regarding location selection, measurement height, and the degree of measurement precision [86]. The measurement data from each station are automatically transmitted to the MeteoSwiss central database, where various quality assurance checks are performed. In the MeteoSwiss data warehouse, measurement data are processed and systematically reviewed on a continuous basis. Measurement gaps are filled, additional parameters are calculated and corrections are made [87]. The pre-processing stages applied to the measurement values from the station to the end user are illustrated in SwissMetNet user guides [87], [88].

The minimum available granularity level of meteorological data in SwissMetNet is 10 minutes. As a result, the time period of the study is quantized into ten-minute intervals. For each

interval, the observation records at the starting point are assigned to the predictor fields in the database.

Raw surface data contain numerous unrelated trends that are likely to reduce the ability of the model to find useful regulations. Hence, removing seasonal and diurnal dependencies would help more useful meteorological signals to stand out. Note that two different de-trending algorithms were tested. However, they were not used since they were shown not to provide any gain in the prediction performance.

The data from lightning location systems are used to first train the ML model and then to validate the accuracy of lightning warnings that it generates as well as the competitive base lines. MeteoSwiss receives lightning localization data from the Météorage company to detect and locate lightning discharges [89]. Météorage is a part of the European Cooperation for Lightning Detection (EUCLID), which is a network of Lightning Location Systems (LLS) operating in western Europe. Detailed information on the EUCLID network can be found in Azadifar et al. and Schulz et al. [90], [91] The average flash detection efficiency for the used datasets in this study is reported to be 95% for cloud-to-ground (CG) flashes and 45% for intra-cloud (IC) flashes [92]. The system measures in real time the angle of incidence and the arrival times of the radiation fields at a network of ground-based measurement stations using LS7001 sensors from Vaisala [93]. The signals are received in the low frequency (LF) band (1 kHz - 350 kHz) [93]. The accuracy of the locations mainly depends on the uncertainty of the arrival time measurements, the background noise level in the operating frequency band, and the number and positions of the stations used to obtain each solution. The arrival times are measured independently at each station using an accurate time base provided by a GPS receiver [94]. The system then combines the data received from all stations to provide a detailed analysis of individual flashes with 100 ms accuracy for the time of strike. It also provides the shape and size of the ellipse that can be said to contain the location of the strike with a 90% level of confidence [95]. In 2017, the median accuracy of detections was reported to be 100 m in Western Europe [93]. In this study, we do not aim to warn for each individual flash, but to warn of the risk of having lightning activity within a ten-minute interval (aka a dichotomous decision basis). Thus, we do not explicitly import the time stamp and location of each individual recorded flash into the model. Instead, we look at total lightning activity in

each ten-minute interval for which we have the meteorological observations available. To do that, based on the distance of each recorded flash from each MeteoSwiss station, the flashes were labeled as long-range lightning activity if they had occurred within an area of radius 30 km surrounding the station. The lightning activity corresponding to each ten-minute interval in the database was assigned to “Yes” (lighting-active class) if at least one flash was recorded in that interval and in the selected area around the station, otherwise it was assigned to “No” (lighting-inactive class). This labeling method also enables us to give lead times to the generated warnings. To do that, the data from preceding observations of meteorological parameters should be used to make the prediction for the following intervals.

CAPE data are retrieved from the ERA5 hourly reanalysis data on single levels provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) [96]. ERA5 is the fifth major global reanalysis produced by ECMWF where, every 12 hours, observations from across the world are combined with the previously generated forecast to produce the most accurate state of the atmosphere at a given point in time. Reanalysis estimations are made at each grid point around the globe over a long period of time with regular time steps and always using the same format. The provision of such estimates makes reanalysis data convenient to work with in this study, especially since we need the data to be uniformly estimated during 12 years. The horizontal resolution of ERA5 data is 0.25˚ x 0.25˚ and the temporal resolution is 1 hour. According to the ECMWF parameter database [97], CAPE is calculated by considering parcels of air departing at different model levels below the 350 hPa level. The maximum CAPE produced by the different parcels is the value retained. The calculation of this assumes: (i) that the parcel of air does not mix with surrounding air; (ii) that ascent is pseudo-adiabatic (all condensed water falls out), and (iii) other simplifications related to the mixed-phase condensational heating.

It can be seen that the spatial resolution roughly matches with the considered long-range activity (30 km) but the locations of the 12 meteorological stations do not necessarily match with the center of the grid for which the reanalysis data are available. In this regard, at each station, the data from 9 neighbor grids imported from the ERA5 are used to interpolate the value of CAPE for the area of interest corresponding to long-range activity in this study, i.e. 30 km surrounding each station. In addition, the temporal resolution of 1 hour for the

reanalysis data does not match with the one from meteorological data (i.e. 10 minutes). Hence, in addition to the data with the original resolution of 1 hour, another version of data is generated where the missing values for ten-minute steps are linearly interpolated. Both versions of data are used in the study and the findings are reported in Results.

Among the selected stations, the Säntis tower was the only one equipped with vertical electrostatic field measurements. The mountain summit has been used as a meteorology station since 1881. In the 1950s, the site was selected for the installation of an 18-m tall radio and TV transmitting antenna that was erected at its summit in 1955 and was replaced by a taller, 84-m tower in 1976. That tower was itself replaced in 1997 by the current tower, which is 124 m tall. Since 2010, this tower is instrumented for lightning current measurements using advanced equipment including remote monitoring and control capabilities [98]–[100]. An EFM-100C RS485 BOLTEK E-field mill has been installed since 15 July 2016 to measure the vertical electrostatic field in the immediate vicinity of the Säntis tower. This electro-mechanical device measures the amplitude of the vertical static electric field ($E(z)$) at the installation point. The distance between the installed field mill and the tower base is about 20 m. The system is set to record the field continuously with a sampling time of 50 ms. The highest range of electric field that can be recorded is ±20 kV/m [101].

The Electric Field Mill (EFM) sensor not being located over a perfect flat ground, the electric field measurements could be affected by the environment. In addition to that, the surrounding objects such as buildings and tall objects could partially shield the electric field which would consequently affect the measured electrostatic field values. Hence, a correction factor, k, was considered to correct the measured values due to these possible sources of error. In this study, the correction factor was determined by comparing the measured values of the vertical electrostatic field with simulated results obtained using COMSOL Multiphysics software for fair weather. The simulation model incorporates the exact terrain topography at the mounting location of the sensor. The modified data are then used as predictor by the E-FIELD model to provide a competitive baseline for the ML model. The value corresponding to the maximum recorded amplitude of the vertical electrostatic field during each ten-minute interval was considered and assigned to the electrostatic field parameter for the corresponding interval in the database.

Once the database was formed, we partitioned the data at each station into two parts: (i) Data Part 1, including the first four years of data (from 2006 to the end of 2009) and (ii) Data Part 2, including the data for the remaining 8 years (from 2010 to the end of 2017). We then used the first part to do the model search and tune the model and its hyperparameters (Stage #1), and withheld the second part to do the final evaluation (including both training and testing) using the information derived during the first stage (Stage #2). Doing this greatly decreases the risk of overfitting since the data used for the final performance evaluation (Data Part 2) remained independent from the part that was used for model search and tuning processes (Data Part 1). What follows describes the model selection, generation, tuning, training, and testing procedures.

2.4.2 Stage #1: Model selection, generation, and tuning

All gathered data subsets in this study are featured as high dimensional and multivariant datasets. In Figure 2-7, the data subset 6 using a parallel coordinates plot can be visualized. The plot maps each row of data as a line. The orange lines correspond to the lighting-active class and the blue lines are the data from the lighting-inactive class. Looking at the distribution of these two classes in each of the coordinates, the plot shows that the two classes are highly mixed in all coordinates and no explicit distinction could be found. Similar high complexity was found as well in other data subsets summarized in Table A-3. Further to this complexity, after labeling each piece of data using the aforementioned procedure in Section 2.4.1, the two classes turned out to be highly imbalanced at all stations. The imbalance was expected since lightning-active periods throughout the year are rare compared to periods devoid of lightning. Due to this high imbalance seen in the data, an extensive model search process was carried out to choose the most appropriate machine learning classification model based on Data Part #1 at each station. To do this, the TPOT Python Automated Machine Learning tool [102] was used (i) to choose the best-fit model, and (ii) to tune the hyperparameters of the model at each station. When applied to a certain dataset, the AutoML approaches automatically explore lots of possible machine learning pipelines and build the one with competitive classification accuracy for that specific task [103]. The results drawn from separate runs at each of the stations and for each of the three lead time ranges indicated that the best performance would be achieved using the XGBoost algorithm. XGBoost [48], [104]

stands for “Extreme Gradient Boosting” and it is a variant of the gradient boosting machine which uses a more regularized model formalization to control overfitting.

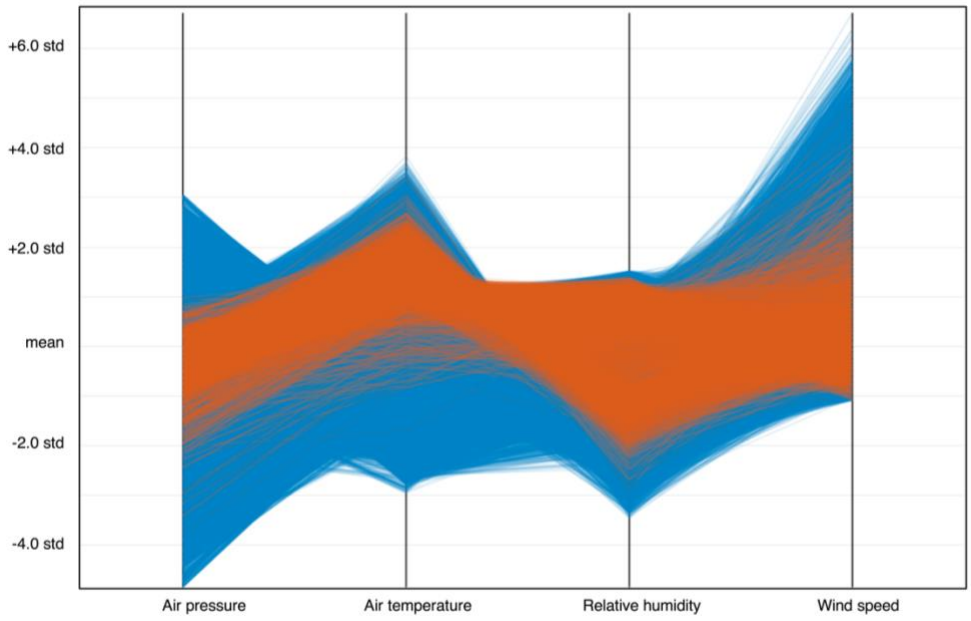


Figure 2-7 Parallel coordinates plot from data subset 10. The mean of each predictor is set to zero and the predictors are scaled by their standard deviations. Each line represents a recorded observation at the start of a ten-minute interval and is labeled according to lightning activity in that interval as either blue (without any long-range lightning activity) or orange (with at least one long-range lightning activity recorded).

To do the classification, the XGBoost algorithm generates an ensemble learner out of individual classification trees using a scalable tree boosting system. Ensemble learners use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone called weak learners [105], [106]. Weak learner is an algorithm that generates classifiers that can merely do better than random guessing. In order to design an ensemble system, three questions need to be answered: (i) How will the individual classifiers (base classifiers) be generated? (ii) What is the number of ensemble members? and (iii) what is the ensemble aggregation method? What follows briefly describes the framework for ensemble learning used in this study.

In this study, we used classification trees as the weak learners. Classification trees are decision trees which predict a response following the decisions in the tree from the root node down to the leaf nodes where the responses are. Figure 2-8 shows an individual decision tree made by the model at an arbitrary iteration number. The flow of data points is split at each node

based on the condition at each internal node. Each data point flows to one of the leaves following the direction on each node. When a data point reaches a leaf, a weight is assigned to it as the prediction score. The predictive algorithm would then combine the prediction scores that each data point gains from the ensemble members to make the final decision about the class to which it belongs, whether lighting-active or lighting-inactive. For ease of presentation, the maximum depth of the tree is limited to 3 in Figure 2-8. In the real training, however, the parameters are tuned using hyperparameter tuning skills.

Boosting is a machine learning ensemble algorithm that is based on the idea that a weak learner can be turned into a strong learner that generates a classifier that is arbitrarily well-correlated with the true classification. Most boosting algorithms consist of iteratively learning weak classifiers and adding them to a final strong classifier. At each iteration, the algorithm attempts to construct a new model that corrects the errors of its predecessor. Hence, the next weak learner will learn from an updated version of residual errors.

The XGBoost algorithm is called gradient boosting since the objective function is optimized using the gradient descent algorithm before each new model is added. The objective function consists of two terms: The loss function, which is put as a measure of the predictive power, and the regularization factor, which controls the complexity of the model which helps to avoid overfitting. At each iteration, the algorithm needs to solve two key problems: (i) How to define the structure of the next weak learner (decision tree) in the ensemble so that it improves the overall prediction skill, and (ii) how to assign the prediction scores to the leaves. The algorithm uses gradient descent to solve these two problems. To build a tree, the algorithm greedily enumerates the features and finds the best splitting point by calculating the split gains. After each split, it assigns the weight to the two new leaves grown on the tree. This process continues repeatedly until the maximum depth is reached. The algorithm then starts pruning the tree backwards and removes nodes with a negative gain.

More information about the XGBoost algorithm including the definition and calculation of the loss function, regularization function, and split gain can be found in Chen and Guestrin [104] and Chen and He [48].

For each subset of the data, some hyperparameters of the model such as the number of trees or iterations in the ensemble (number of learners), the rate at which the gradient boosting learns (learning rate), and the depth of the tree (maximum depth), were optimally selected using both manual and AutoML approaches.

In the manual approach, the model was first initialized with a set of hyperparameters. Second, using 4-fold cross validation [107], we repeatedly split Data Part 1 at each station into four folds (groups) in a way that each group contains the data from a specific year. The XGBoost model was fitted on the data from 3 years (training set) and evaluated on the data from the remaining one year (validation set). This process is repeated until each group (each year of data) had been assigned once as the validation set. At the end, the results from all four runs were summarized to give the overall classification skill. The hyperparameters of the model at each station were tuned in order to improve the summarized cross validation scores. The AutoML approaches, in turn, do an intelligent search inside the hyperparameter space sweeping a broad range of possible combinations to find the optimized set of parameters that perform best on the given data (Data Part 1 at each station) [103].

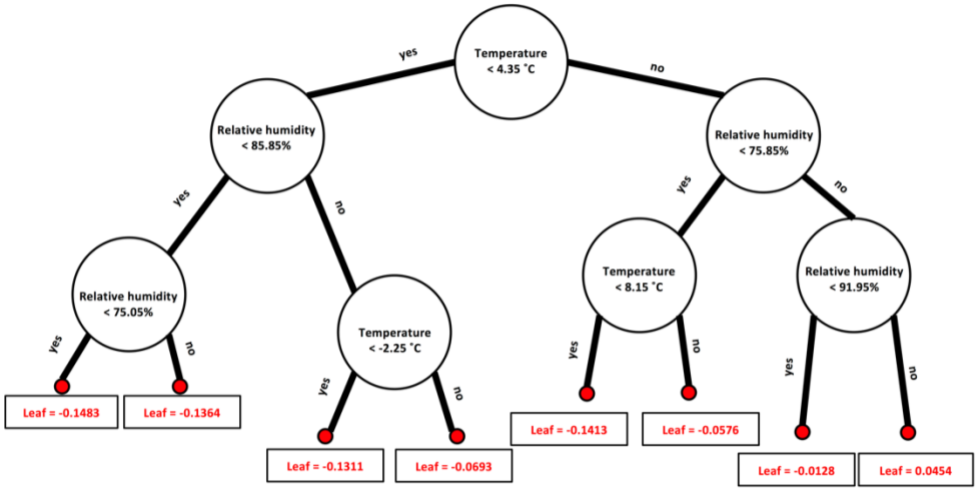


Figure 2-8 Sample of a decision tree grown by the ML model in the ensemble classifier. In this example, the maximum depth of the tree is set to 3 and subset 1 is used as the training set. The prediction score at each leaf would be assigned to its associated observations. The model then combines the prediction scores for each sample to predict its class as whether lightning active or lightning-inactive.

Given the large temporal coverage and high temporal resolution of the gathered data, it is common that the data contains noise and outliers due to for instance to measurement errors. Removing the noise and outliers allows the learning algorithm to learn more accurate

classification criteria and helps to provide better evaluation of the classification quality. We took advantage of the ML model evaluation results in Data Part 1 to find the conditions when the model has poor performance by looking at a random number of the misclassified instances. We then identified criteria that could explain these conditions and used them to identify samples in Data Part 2 with similar conditions to those of the outliers in Data Part 1. These samples were then removed from Data Part 2 with the presumption that the ML model would have no skill under those conditions. As a result of this filtering process, at some of the stations, a small portion of data (the size varied between 4% and 6% of the total data at each station) remained un-fitted and, hence, excluded from the final training and testing procedure based on Data Part 2. It is worth noting that the criteria to identify and filter the outliers on final evaluation were derived based on Data Part 1 and the filtering was done before the training and testing procedure based on Data Part 2. One should also note here that since this filtering process starts with selecting a random number of the misclassified samples in Data Part 1, different executions may lead to different results. More work is underway to optimize this process and to make it fully automatic.

2.4.3 Stage #2: Training and testing procedure

As mentioned in the previous section (Stage #1), to do the final evaluation, the predictive ML model was trained and tested based on Data Part 2. To do this, at each station, Data Part 2 was split into different groups in such a way that each group contained the data from an individual year. As a result, each observation in the dataset was assigned to an individual group and remained there for the duration of the training and testing process. For each unique group, the group was held out from the dataset as the test set and the training was done using the remaining groups as the training set. The XGBoost model with the hyperparameters already optimized based on Data Part 1 was then fitted on the training set and evaluated on the test set. The prediction results on the test set were evaluated using the evaluation metrics. The process continued until each individual group had been taken once as the test set. The evaluation results were combined over the rounds to summarize the model prediction skill. This validation method is similar to the k-fold cross validation whereas the folds are forced to be the data from individual years and are not randomly selected from the shuffled data. This splitting method would help to eliminate the leakage of correlated

samples from the training set into the test set due to the high temporal correlation of lightning data. The proposed approach in Stages #1 and #2 is summarized in Figure 2-9.

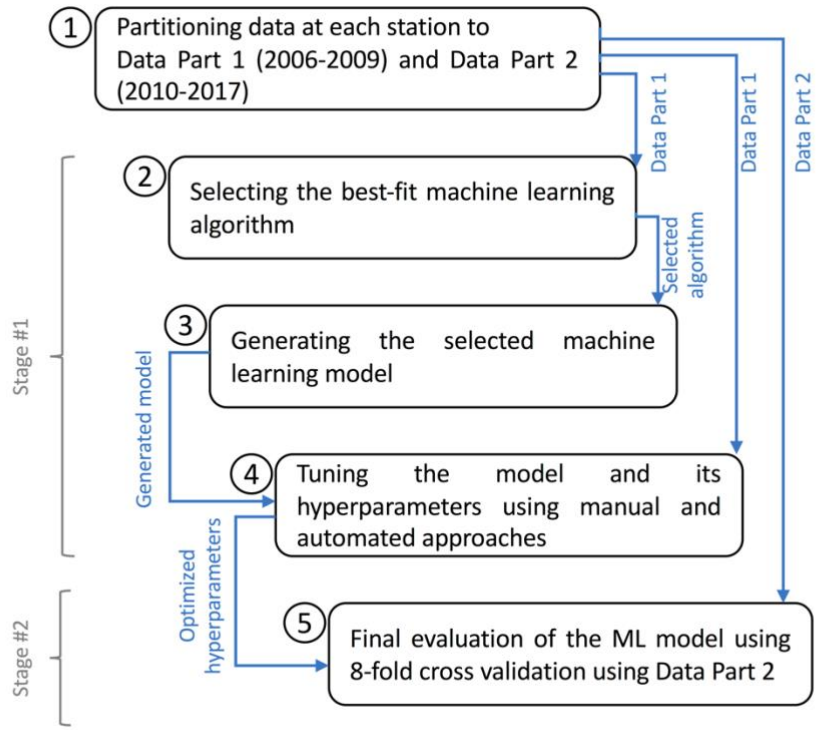


Figure 2-9 Summary of the model selection, generation, tuning, and evaluation.

In order to provide lead to the warnings, the observational data for meteorological parameters for a given ten-minute interval was used to do the prediction for the following intervals. To achieve this, instead of feeding the model with the labels of the same interval, the labels for the following intervals should be used. Given the fact that both the meteorological and lightning data were imported into the database with the granularity of ten minutes, the lead times would also be quantized in ten-minute ranges. For example, if the model is fed with the lightning labels corresponding to the same interval as the one for the meteorological data, then the lead time for the warnings would be 0-10 minutes, which corresponds to an imminent warning. However, if, instead, the lightning labels for the next interval were used, then the lead time of the prediction would be 10-20 minutes.

2.4.4 Model evaluation metrics

Even in high activity regions, lightning strikes are rare. It is important for the nowcasting scheme to correctly predict non-lightning events (lightning-inactive samples) which numerically dominate lightning events. However, while a low false alarm rate is desirable, it is not indicative of predictive skill when true alarms are rare. In other words, in imbalanced databases, neither the overall accuracy nor the false alarm rate may be able to correctly evaluate the significance with which the prediction scheme performs better than chance [108]. To bridge the gap, a couple of metrics to measure the skill in rare event forecasting are suggested in the literature which are mainly based on the values of the contingency table [109]. A sample of a contingency table for the two-class prediction scheme is given in Table A-2. The rows and columns correspond, respectively, to the observed and predicted alternatives. Giving a customized definition for the case of lightning prediction studied in this paper, for example, hit is the total number of times that at least one lightning activity (either CG or IC) occurred in a specific area in a specific time frame as it was correctly predicted by the predictive scheme. The specific area corresponds to the areas within the circular distance of radius 30 km (based on the adopted definition for long-range activity) around each of the 12 stations and the specific time frames are ten-minute time windows defined according to the desired lead time. Similarly, correct rejection denotes the total number of times that the predictive model responds that lightning will not occur when it indeed did not occur. The Miss parameter gives the number of cases that the predictive scheme actually misses the occurred events and False Alarm gives the number of cases when the model would predict lightning when it did not occur. Based on the four described entries in the contingency matrix, a couple of performance parameters are adapted and defined in Table 2-1.

Probability of Detection (POD) is defined as the ratio of the hits to the total number of observed events (lighting-active samples). This parameter shows how the prediction scheme was able to correctly predict the rare events (lighting-active samples). However, it does not provide any information on how the model performs in the majority of cases where no lightning has actually occurred. The False Alarm Ratio (FAR) indicates the fraction of lightning alarms issued by the model that were actually false. The Critical Success Index (CSI) is sensitive to both POD and FAR since it penalizes both misses and false alarms. It can be also regarded

as a measurement of accuracy when correct rejections are removed from consideration assuming that they are less important. Therefore, it concentrates on the fraction of hits to the total number of both forecast and missed events.

The Heidke Skill Score (HSS) is also used to evaluate the performance of the model. The score ranges from -∞ for no skill to 1 for perfect skill and it measures the performance of the prediction scheme after eliminating the correct predictions that would have been achieved purely by random chance. The Heidke Skill Score (HSS) is known to be usable in forecasting rare events since it gives credit to the correct rejections in a controlled way so that the false alarms are also considered. It is also known to take into account the correct random forecasts of both event and non-event cases [109]. The performances of the ML model and the competitive models are evaluated using the aforementioned metrics.

Chapter 3

Paper title: Machine Learning Based Lightning Localization Algorithm Using Lightning-Induced Voltages on Transmission Lines[2]

List of authors: Hamidreza Karami, Amirhossein Mostajabi, Mohammad Azadifar, Marcos Rubinstein, Chijie Zhuang, and Farhad Rachidi

Author contributions: H.K. and A.M. conceived the study. H.K. and M.A. performed the numerical calculations. A.M. performed the machine learning modeling and data anlysis. C.Z., M.R. and F.R. supervised the study and contributed to the interpretation of the results. A.M. led the manuscript preparation with input from all co-authors. All co-authors reviewed the manuscript.

Abstract: In this study, we present a Machine Learning based method to locate lightning flashes using calculations of lightning-induced voltages on a transmission line. The proposed approach takes advantage of the preinstalled voltage measurement systems on power transmission lines to get the data. Hence, it does not require the installation of additional sensors such as ELF, VLF, or VHF. The proposed model is shown to yield reasonable accuracy in estimating 2D geolocations for lightning strike points for different grid sizes up to 100 × 100 km^2. The algorithm is shown to be robust against

[2] Postprint version of the article published in IEEE Transactions on Electromagnetic Compatibility (DOI: https://doi.org/10.1109/TEMC.2020.2978429)

the distance between the voltage sensors, lightning peak current, lightning current rise time, and signal to noise ratio of the input signals.

3.1 Introduction

Knowing the exact geolocation of a lightning strike is important in a wide range of research and application domains, including geophysical research, lightning warning, aviation/air traffic, weather services, insurance claims, power transmission and distribution, etc**.** For example, some lightning warning systems rely on such data to indicate approaching thunderstorms [28], [29] and thus to prevent catastrophic effects of lightning strikes to critical infrastructure, sensitive equipment or systems, and outdoor facilities. Even in high-energy atmospheric physics, lightning location data are used to seek the lightning flashes associated with Terrestrial Gamma-ray Flashes (TGFs) seen from space [30].

Given its fundamental role in many aspects of lightning studies, appreciable attention has been given to accurate lightning localization. The most widely used lightning location techniques are the time-of-arrival method (ToA) and the magnetic direction finding method [31]–[33]. In [34]–[37], a method was proposed to locate the lightning strike point based on the Electromagnetic Time Reversal (EMTR) theory. In this technique, first the electric field waveforms of the lightning strike are measured by multiple sensors (this is the so-called “forward” or direct-time phase). Second, a time-reversed version of these waveforms is back-injected from the sensor points into the location domain, also called the computational domain since this phase is carried out by simulation (the so-called “backpropagation” or reverse-time phase). Finally, a criterion is used to determine the location of the lightning strike point. It is reported in [34] that at least 3 sensors are needed to locate the lightning strike using EMTR. Recently, Qin et al. [110] introduced a GPU-based algorithm to increase the performance of lightning geolocation networks. The algorithm has been effectively applied in, respectively, a six-station and a five-station networks for two-dimensional (2-D) and 3-D geolocation estimation of lightning flashes. All the above-mentioned approaches require the installation of appropriate sensor modules (ELF, VLF, or VHF).

A methodology based on the difference of the time of arrival of the induced voltages on a transmission line was presented in [111] to obtain the coordinates of the lightning strike. However,

in the derivation, the direct field from the lightning strike was ignored, which is not an acceptable assumption in most cases.

This paper presents a new machine learning based lightning localization algorithm in 2D that utilizes data from the preinstalled voltage measurement systems on power transmission lines. The algorithm requires at least two sensors to operate. It should be noted that using only two sensors can lead to an ambiguity in the lightning location on both sides of the transmission line. The ambiguity can be removed either by considering another voltage or current sensor on another transmission line, or by considering the terrain topography or other objects in the environment that could remove the symmetry of the problem.

Machine learning algorithms can give computers the ability to learn a skill (such as the prediction of the geographic coordinates of a passive or active object) from sets of archived data, with the final goal of applying the skill to new cases. They do that by automatically extracting an unknown underlying mapping function from the inputs to the outputs. In supervised learning, they are designed to learn a target function that best maps input features to the outputs given a dataset with an already known output (the so-called training set). They would use the learned function in future to estimate the outputs for new unseen examples of the input features (the so-called testing set). Recently, machine learning has been also shown to be a useful method for source localization. For example, Huang et al. [112] applied deep neural networks to acoustic source localization in shallow water environments. Vera-Diaz et al. [113] used deep learning to directly estimate the three-dimensional position of a single acoustic source using raw data from microphone arrays. Regarding the application of Machine Learning to lightning localization, we present in this paper a proof of concept on how machine learning algorithms could be used to locate lightning flashes by looking at their associated lightning-induced voltages on transmission lines. To achieve this, a machine-learning-based model is trained to estimate the 2D geolocation of a lightning strike point, given the real-time measured values of lightning-induced voltages measured by two sensors on the transmission line.

The paper is organized as follows. Section 3.2 explains different steps in the proposed algorithm including building the database, machine learning model selection and generation. A discussion is then given in Section 3.3 on how the generated model is trained and tested to estimate the lightning

strike point using numerical simulation results. Finally, conclusions and a final discussion are given in Section 3.4.

3.2 Methodology

3.2.1 Numerical Simulation and Data Acquisition

In this section, we aim to train a machine learning model to estimate the geolocation of an electromagnetic source, given the data of the lightning-induced voltages obtained by two sensors located on a power transmission line. The geometry of the problem, shown in Figure 1, consists of an infinite transmission line with two sensors that capture the lightning induced voltages on it. We defined 2000 uniformly random positions for the source within the presented geometry in Figure 1.

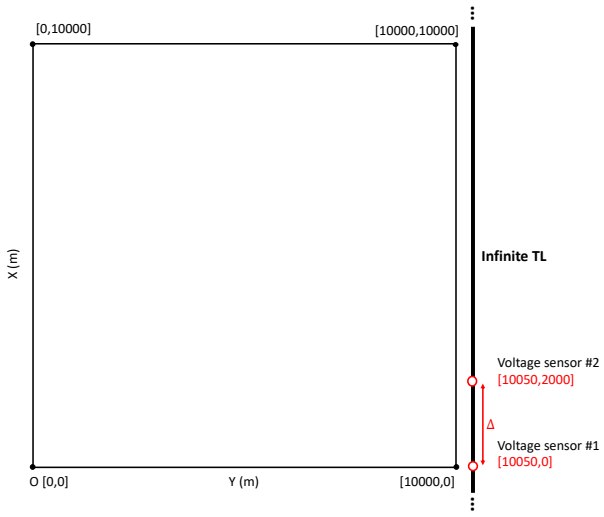


Figure 3-1 Sketch diagram of the detection region. The infinite transmission line is located 50 m beside the detection region. The two voltage sensors are located on the line and 2 km away from each other (Δ=2 km). The coordinates of the sensor positions are annotaed in red.

To build the required database for training and testing procedures, the lightning-induced voltages from lightning striking at each of these source positions are calculated, without loss of generality, using Rusck's formula [114], which is written as follows,

$$v(x,t) = \frac{Z_0 I_0 \beta}{4\pi} \cdot \left[\begin{array}{l} \frac{\gamma_-}{(d^2+\beta^2{\gamma_-}^2)} \left(1 + \left(r + \frac{\beta^2{\gamma_-}^2}{\sqrt{(\beta ct)^2+\rho^2/\delta}}\right)\right) + \\ \frac{\gamma_+}{(d^2+\beta^2{\gamma_+}^2)} \left(1 + \left(r + \frac{\beta^2{\gamma_+}^2}{\sqrt{(\beta ct)^2+\rho^2/\delta}}\right)\right) \end{array} \right] \tag{3-1}$$

where,

$\gamma_- = (ct - x)$, $\gamma_+ = (ct + x)$, $\delta = \frac{1}{1-\beta^2}$, and $\rho = x^2 + d^2$.

The amplitude of the return stroke current is I_0 and its velocity is equal to βc, in which c is the velocity of light in free space and β, set to 0.4 in this paper, is the ratio of the return-stroke speed to the speed of light. t denotes the time and it should be greater than $\sqrt{x^2 + d^2}/c$. d is the horizontal distance between the lightning channel and the transmission line and x is the distance along the transmission line (Figure 1). Z_0 is the characteristic impedance of free space. In Rusck's formula, the lightning-induced voltages are calculated for a lossless, single-wire transmission line above a perfectly-conducting ground. The excitation source (i.e., the lightning flash) is a step current ascending along the lightning strike channel.

To calculate the induced-voltages at the nearest point of the transmission line (x = 0) to the lightning stroke, the following formula is used,

$$v\,(0,t) = \frac{Z_0 I_0 h}{4\pi d^2} \cdot \frac{2\beta ct}{1+\left(\frac{\beta ct}{d}\right)^2} \cdot \left(1 + \beta^2 \left(\frac{ct/d}{\sqrt{1+\beta^2((ct/d)^2-1)}}\right)\right) \qquad (3\text{-}2)$$

valid only for $t \geq d/c$. All parameters have been defined before.

It should be noted that other, more accurate induced-voltage calculation approaches, either based on transmission line theory (e.g., [115]) or full-wave simulations (e.g., [116]) can be easily substituted in the proposed method. Using (3-1), we calculated the lightning-induced voltages in the time domain at the locations of the two sensors.

A. Data Preprocessing

In order to eliminate the effect of the lightning peak current, each sensor's signal was normalized to its maximum value. For each of the sensors, the calculation of the signal was continued until the amplitude of the voltage reached 10^{-7} of the maximum recorded value by that sensor. Figure 3-2 shows a zoomed view of two sample induced voltages recorded by the two sensors.

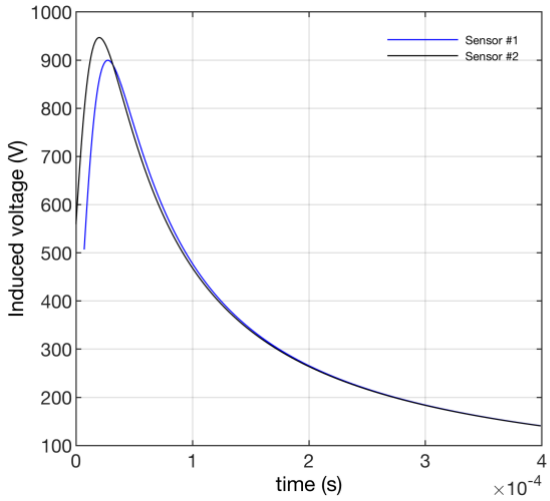


Figure 3-2 A zoomed view of a sample of the induced voltages measured by the two sensors. The 2D coordinates of the lightning strike point is [6996.9 m, 9459.3 m] and the excitation source is a step current ascending along the lightning strike channel. The voltages are calculated using the Rusck's formula and they are shifted so that the first signal starts from time equals to zero.

To remove the information of the absolute time in the signals recorded by the sensors, the sensor signals are time-shifted, so that the first one starts from zero. For example, in Figure 3-2, both signals are shifted in such a way that *time* equals to zero corresponds to the start of the first signal (i.e. signal from Sensor #2).

Once all random positions were considered, we ended up with 2×2000 different waveforms of the 'measured' transient voltages with each of them having a different number of samples. We performed two preprocessing steps before feeding the 'measured' signals into the model:

(i) Since the model will treat these samples as the features for its learning process, the differing number of samples would make the number of features be different from one observation to another. However, for the machine learning model, the number of features in the training set must be the same as the number of features in the test set. Therefore, a preprocessing step was required to make all transient signals be of the same length. To do that, we extended the length of the shorter signals by appending zeros to the end of each one of the shorter signals until it reached the maximum length, N_{max}, among the recorded signals for sensors 1 and 2.

(ii) The second preprocessing stage relates to how the data from the two sensors were merged before being used as the input features. We tried several linear and non-linear combinations using both, original and time-reversed signals. After several tries, the linear combination in (3-3) was seen to yield significantly higher accuracy:

$$V(t) = V_1(t) - V_2(\mathrm{T} - t) \quad (3\text{-}3)$$

where $V(t)$ is the new combined signal, $V_1(t)$ is the signal measured by sensor #1 after the first preprocessing step was applied (i.e. with the length of the signal extended to N_{max}=10000), T is the time window, and $V_2(T - t)$ is the time-reversed version of the signal measured by sensor #2 after the first preprocessing step was applied.

After the above preprocessing steps, we formed a tabular database with each row containing the values of $V(t)$ for each of the 2000 iterations mentioned above. In each row, the N_{max} samples of $V(t)$ were used as the predictors and the x and y coordinates of the source were used as the response.

3.2.2 Selection of the Machine Learning Model

Once the database is formed, a machine learning algorithm needs to be employed to identify regularities between predictors and responses using a portion of the data which, as mentioned above, is called the training set. The trained model can then use the explored correlations to predict the response for new cases (testing set). A model search process was performed to choose the most appropriate machine learning regression model. The model search process repeatedly looks for the best-fit model through several regression types including regression trees, support vector machines, neural networks, and different ensemble methods such as bagging and boosting. The results (Table 3-1) indicated that the best performance would be achieved using the XGBoost algorithm. XGBoost stands for “Extreme Gradient Boosting” and it is a variant of the gradient boosting machine which uses a more regularized model formalization to control overfitting [104]. More information on the XGBoost model description and generation is given in the following section.

3.2.3 Model generation

In this study, we generated an ensemble learner out of individual classification trees using a scalable tree boosting system. Ensemble learners use multiple learning algorithms to obtain better predictive

performance than could be obtained from any of the constituent learning algorithms alone called weak learners [105], [106]. A weak learner is an algorithm that generates classifiers that can merely do better than random guessing. What follows briefly describes the framework for the ensemble learning used in this study.

Table 3-1 Performance results (RMSE) for state-of-the-art algorithms in the case of two sensors as shown in Figure 3-1. The dataset included 2000 samples and the evaluation method was 5-fold cross validation.

Model Type	**RMSE for x-coordinate (m)**	**RMSE for y-coordinate (m)**	**Hyperparameters**
Regression tree	492	75	minimum leaf size: 4
Support Vector Machine (SVM)	658	225	kernel function: linear
Feed forward neural network	2155	469	optimizer: Adam activation function: rectified linear unit (ReLU)
Bagged trees	461	60	minimum leaf size: 8 number of learners: 200
Boosted trees	404	42	minimum leaf size: 8 number of learners: 200
XGBoost	88	37	maximum depth: 30 number of learners: 200

A. Preparing the weak learners

In this study, we used decision trees as the weak learners. Decision trees predict a response following the decisions in the tree from the root node down to the leaf nodes where the responses are. The flow of data points is split at each node based on the condition at each internal node. Each data point flows to one of the leaves following the direction on each node. When a data point reaches a leaf, a value is assigned to it as the prediction score. The predictive algorithm then combines the prediction scores that each data point gains from the ensemble members to generate the output.

B. Ensemble aggregation method

Boosting is a machine learning ensemble algorithm that is based on the idea that a weak learner can be turned into a strong learner. Most boosting algorithms consist of iteratively learning weak learners and adding them to a final strong learner. At each iteration, the algorithm attempts to construct a new model that corrects the errors of its predecessor. Hence, the next weak learner will learn from an updated version of residual errors.

The XGBoost algorithm is called gradient boosting since the objective function is optimized using the gradient descent algorithm before each new model is added. The objective function consists of two terms: The loss function, which is put as a measure of the predictive power, and the regularization factor, which controls the complexity of the model which helps to avoid overfitting. At each iteration, the algorithm needs to solve two key problems: (i) How to define the structure of the next weak learner (decision tree) in the ensemble so that it improves the prediction along the gradient, and (ii) how to assign the prediction scores to the leaves. The algorithm uses gradient descent to solve these two problems. To build a tree, the algorithm greedily enumerates the features and finds the best splitting point by calculating the split gains. After each split, it assigns the weight to the two new leaves grown on the tree. This process continues repeatedly until the maximum depth is reached. The algorithm then starts pruning the tree backwards and removes nodes with a negative gain.

In this study, we used the XGBoost python package to build the classifier [117]. An ensemble of 200 trees with depths capped to 30 and learning rate equal to 0.05 achieved the best result for estimating the x and y coordinates in this problem. More information about the XGBoost algorithm, including the definition and calculation of the loss function, regularization function, and split gain can be found in Chen and Guestrin [104] and Chen and He [48].

3.3 EVALUATION OF THE MACHINE LEARNING MODEL

In this study, the predictive ML model was evaluated using a 5-fold cross-validation described in what follows. The k-fold method divides the dataset into k equal parts. In this study, first, the dataset was shuffled and split into five different groups. As a result, each observation in the dataset was assigned to an individual group and remained there for the duration of the training and testing process. Each unique group was held out from the dataset as the test set and the remaining four groups were used as the training set. The model was then fitted on the training set and evaluated on the test set. The model estimation results on the test set were retained and the trained model was discarded.

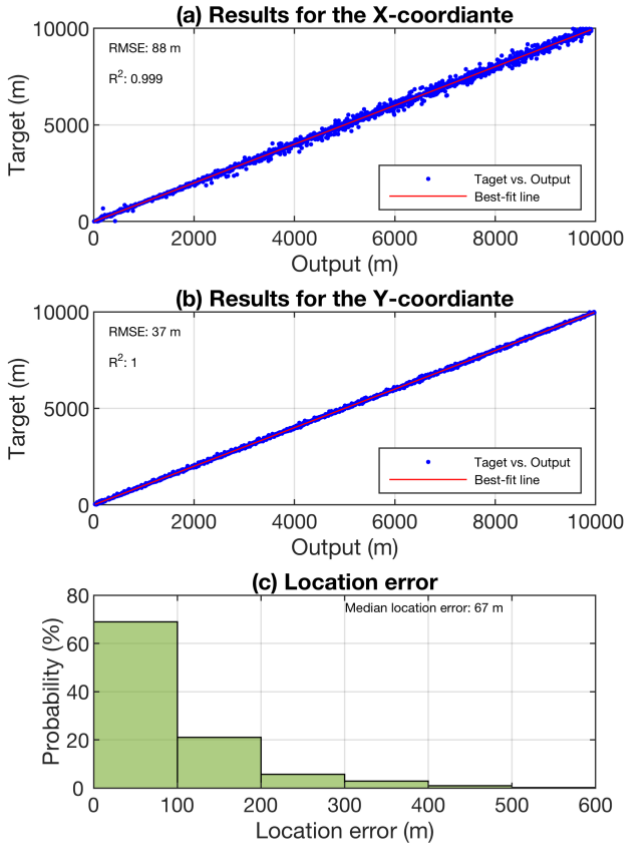


Figure 3-3 Model evaluation results for (a) the x-coordiante and (b) the y-coordiante. (c) is the histogram of the location error considering 2000 randomly selected lightning strike points inside.

The process was repeated until each individual group had been taken once as the test set. The model outputs were combined over the rounds. This splitting method helps to eliminate the leakage of correlated samples from the training set into the test set. Moreover, it avoids overfitting to a specific testing set since each one of the observations is considered as part of the testing set in one of the five rounds. Once the five rounds of training and testing were done, the model's prediction skill was evaluated by comparing the outputs with the target values for the x- and y-coordinates. The results presented in Figure 3-3a and Figure 3-3b, are the scatter plots of the target versus output values where the blue dots correspond to the 2000 observations in the database. The best-fit lines are calculated using the least-squares regression method. The very high values of the coefficient of determination (R^2) indicate that a high proportion of the variance in the data are explained by the fitting line. In order to visualize the error, 2.5% of the total dataset that had the largest localization errors compared to the remaining locations were considered outliers and excluded from the evaluation results (the same is done in the rest of the paper). Looking at the results in Figure 3-3a

and Figure 3-3b, the Root Mean Squared Errors (RMSE) are 88 m and 37 m for the *x* and *y* coordinates, respectively. These low error values mean that by just looking at the lightning-induced voltage waveforms measured by the two sensors on the transmission line, the model was able to accurately predict the location of the nearby lightning along the *x* and *y* axes. Finally, the relative probability of the location errors between the ground truth 2D source position and the estimated one by the model for each of the 2000 considered source positions is given in Figure 3-3c. The height of the bar in each bin is the relative number of observations that falls into that bin. According to the results, in more than 65% of the cases, the model was able to predict the source with less than 100 m location error.

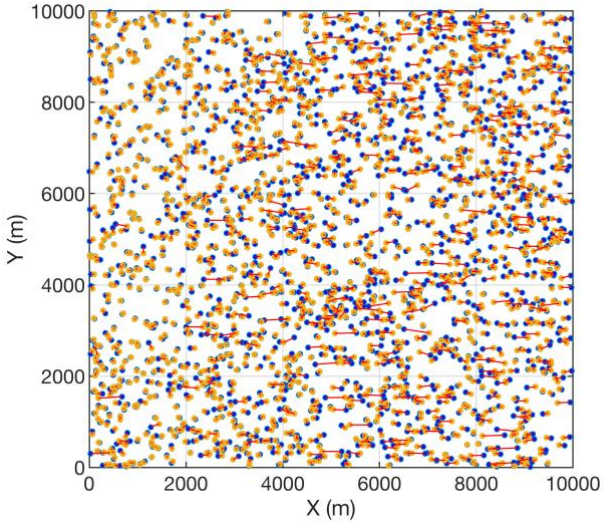


Figure 3-4 Scatter plot of the target (blue dots) versus estimated (yellow dots) 2D geolocations for the N = 2000 guest lightning strike points using the proposed ML based approach.

Figure 3-4 is the scatter plot of the true (blue dots) vs. estimated (yellow dots) lightning locations for the 2000 studied samples. In order to visualize the location error density based on the position relative to the transmission line, the heatmap of the location errors is shown in Figure 3-5. The figure shows, for each grid cell, the average of the location errors estimated by the model for the samples that fell into that grid. The colormap is the location error in meters.

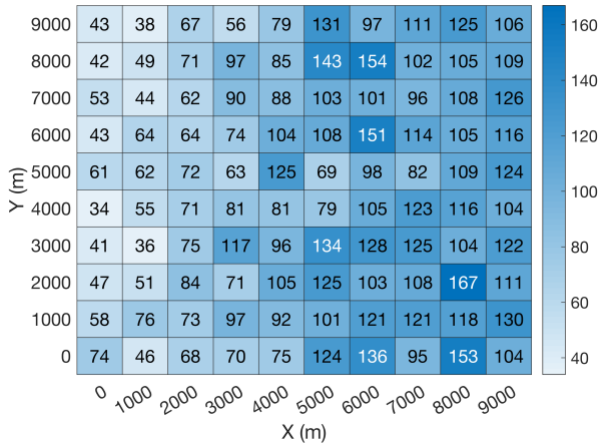
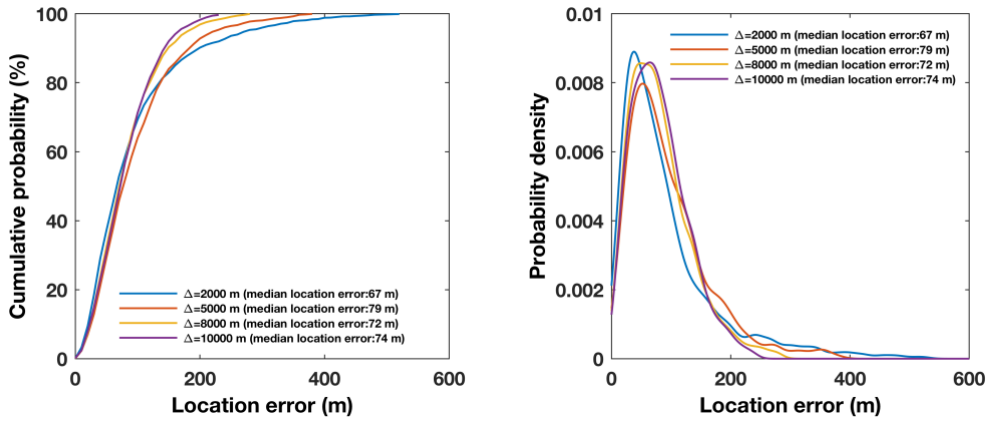


Figure 3-5 Average location error presented as heatmap chart inside the detection region. The colormap represent the location errors in m. The (0,0) point corresponds to the coordinate center (O) shown in Figure 3-1. The x and y labels for each of the cells are the coordinates of the left-bottom corner of the grid cell.

3.3.1 The Impact of the Sensors' Positions

In this subsection, we examine the effect of the distance separating the two sensors on the performance of the system. Specifically, we investigate how the geolocation accuracy changes when the distance between the two sensors varies along the transmission line. To do that, we changed the distance between the two sensors from 2 km to 5 km, 8 km, and 10 km. We did that by moving sensor #2 from [100050, 2000] to [10050, 5000], [10050, 8000], and [10050, 10000] in the detection region shown in Figure 3-1. For each of these new distances, we followed the same procedure described for the case of 2 km-distance in Section 3.2 using the same 2000 guest locations of lightning strike points. Figure 3-6 presents the cumulative probability and the Probability Density Function (PDF) of the location errors for the four different sensor positions. Looking at the results for the studied sensor positions, one can see that the median location error varies in a narrow range (between 67 m to 79 m) compared to the size of the detection region and the maximum location error is less than 600 m. Moreover, probability distribution functions are seen to have similar widths at the four sensors' positions. Given this, the proposed ML model seems to have low sensitivity to the distance between the deployed sensors. The reason comes back to the adaptability of the ML model to the change in the input data. In other words, once the sensors are moved to the new position, the ML model would then learn the new underlying relationships using the new input data and, hence, it would be able to deliver similar performance results.



(a) (b)

Figure 3-6 (a) Cumulative probability and (b) probability density estimate of the location errors for four different sensors' positions. Δ is the distance between the two deployed voltage sensors on th transmission line (see Figure 3-1).

3.3.2 Increasing the Number of Sensors

As stated in the Introduction, if only two sensors are used, an ambiguity exists between two sides of the transmission line. One potential way to resolve this ambiguity is to provide the data from an additional sensor to the ML model. In order to investigate the effectiveness of this solution, we added to the medium a second transmission line with a third voltage sensor. We then repeated the study for the case of two sensors explained in Section 3.2 but with the following differences: (i) we extended the grid size to 20 x 20 km^2 and (ii) we combined the three sensors signals as follows:

$$V(t) = V_1(t) - V_2(\mathrm{T} - t) - V_3(t) \quad (3\text{-}4)$$

where $V(t)$ is the new combined signal, $V_1(t)$ is the signal measured by sensor #1, T is the time window, $V_2(T - t)$ is the time-reversed version of the signal measured by sensor #2, and $V_3(t)$ is the signal measured by sensor #3. The configuration of the two transmission lines and the three sensors is given in Figure 3-7. The result for this case study shows median and mean location errors of 210 m and 288 m, respectively. The heat map of the location errors inside the medium is given in Figure 3-8.

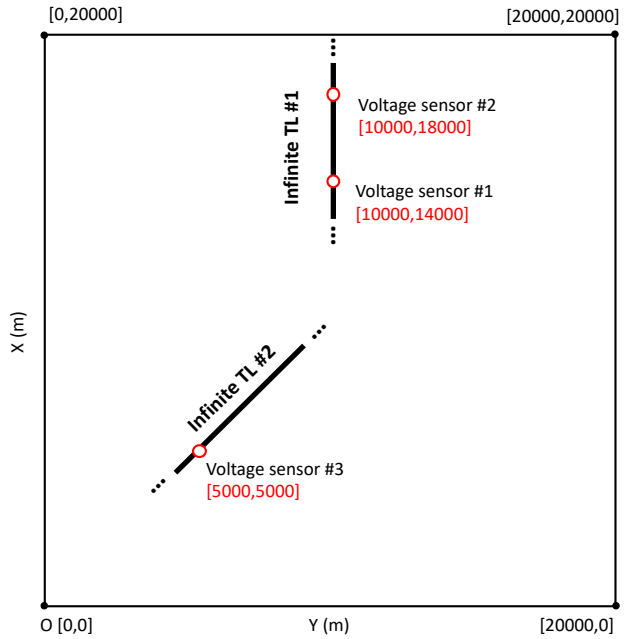


Figure 3-7 Sketch diagram of the detection region for the case of three voltage sensors. The coordinates of the sensor positions are written in red.

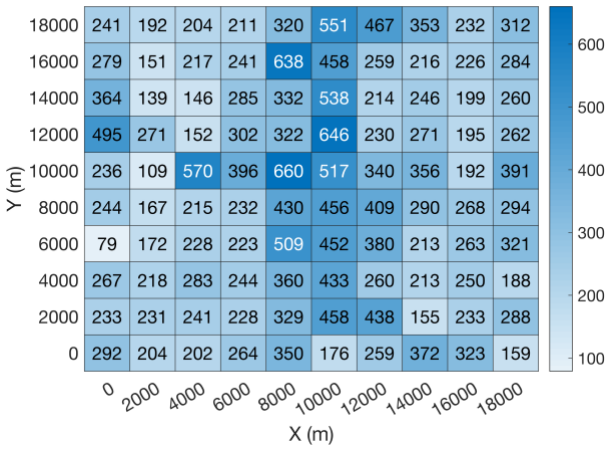
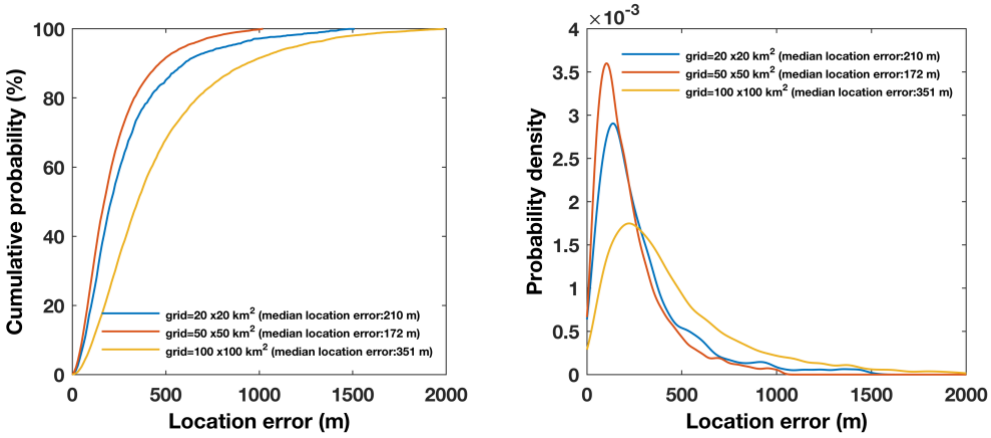


Figure 3-8 Same as in Figure 3-5 when the detection region is changed to the one shown in Figure 3-7 with three voltage sensors in use.

3.3.3 Impact of the Grid Size

At this step, we replicated the analysis in Section 3.3.2 expanding the grid size from 20 x 20 km^2 to 50 x 50 km^2 and to 100 x 100 km^2, increasing the number of guest locations for lightning strike points to 10000. The results for these new grid sizes are presented in Figure 3-9. Based on the results, although the median and maximum location errors increased by increasing the grid size, in more than 80% of the cases the location errors remained within 660 m. This result confirms that the algorithm is able to yield good location accuracy even for lightning flashes that occurred tens of

kilometers away from the transmission line. We have seen that larger location domains lead to greater statistical location errors. The exact relation between the distance to the lightning strike and the statistical location is left out for future work.



(a) (b)

Figure 3-9 (a) Cumulative probability and (b) probability density estimate of the location errors for three different grid sizes and three voltage sensors.

3.3.4 Sensitivity to the Noise Level and the Risetime of the Lightning Current

In this subsection we applied three changes to the methodology of Section 3.3.2 with three sensors in use:

First, as stated in Section 2.1, so far, the lightning current is modeled using a simple step function. To investigate the effect of the risetime on the performance of the model, we redid the analyses using a linearly rising ramp function as the excitation pulse (i.e. the lightning current). As a result, the transmission line (TL) model was used to represent the lightning return stroke, with a channel-base current expressed as follows

$$i(t) = \alpha t u(t) - \alpha(t - t_r)u(t - t_r) \qquad (3\text{-}5)$$

where $u(t)$ and t_r are, respectively, the Heaviside function and the risetime of the current. α is the slope of the linearly rising current.

At each of the 10000 guest locations, a random value for the risetime (t_r) was chosen and used to form the excitation source. The range for the rise time values was derived from the direct current measurements reported in [118]. The induced voltages at the three sensors (Figure 3-7) were calculated using the following analytical formula [119]:

$$v(x,t) = \alpha \frac{30h}{\beta c} \mathrm{Log}\left(1 + \left(\frac{\beta}{t_d}\frac{t^2 - t_0^2}{t + \beta t'}\right)^2 \qquad + 2\beta\ \mathrm{Log}\left(\frac{t + t'}{t_0 + t_0/\beta}\right)\right) \qquad (3\text{-}6)$$

where,

$$t_d = \frac{d}{c}, t_0 = \sqrt{t_x^2 + t_y^2 + t_z^2}, t_x = \frac{x}{c}, t_z = \frac{h}{c},\ t' = \sqrt{t^2 + \tau t_0^2}, \tau = (1 - \beta^2)/\beta^2.$$

Doing this, we investigated how the ML model performs when the lightning currents used during the training and testing procedures can have different risetimes, which is what happens in the real case.

Second, to make the situation closer to practical scenarios, the ML model's performance was evaluated considering noisy input signals. In this regard, a white noise was added to the calculated signals at each sensor with a Signal to Noise Ratio (SNR) varying randomly between 20 to 30 dB.

Third, in order to compensate for the increased difficulty of the problem, we extended the size of the database by increasing the number of data points from 2000 to 10000 guest locations.

The results are presented in Figure 3-10. The median and mean location errors are 329 m and 512 m, respectively. The obtained results show that the ML model still yields acceptable performance in realistic scenarios when the excitation sources have random values for the risetime and the sensor signals are affected by noise.

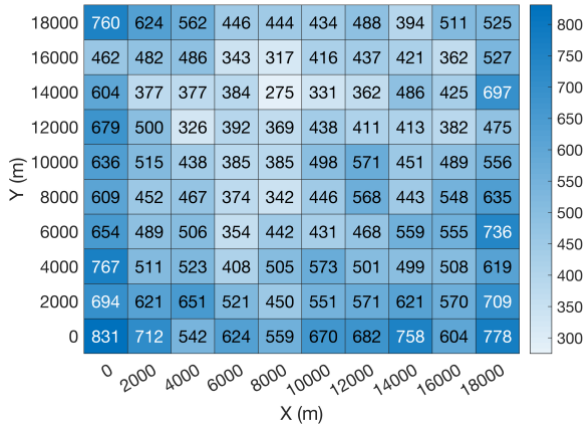


Figure 3-10 As in Figure 3-8 but for excitation sources with random values for the risetime. The sensor signals are noisy, and the number of data points is increased to 10000.

3.4 Conclusions

A Machine Learning (ML) based method was proposed to locate lightning flashes using measurements of lightning-induced voltages on a transmission line. The algorithm builds up a database based on the solutions for lightning-induced voltages on transmission lines. It then uses part of the data to learn the underlying target function that best maps the inputs (measured lightning-induced voltages) to the outputs (geolocation of the lightning strike point). Once trained, the model can estimate the lightning strike point for cases that were not used in the training phase.

The model was tested using data from either two or three voltage sensors and it yielded reasonable accuracy for 2D geolocations in grids of 10x10 km^2, 20 x 20 km^2, 50 x 50 km^2, and 100 x 100 km^2. In case of using two sensors, the sensor's relative position along the transmission line was seen to have a negligible impact on the estimation results, with the model being able to deliver sufficient accuracy both when the sensors are moved closer or farther away from each other. The changes in the peak and risetime of the return stroke currents had negligible effect on the geolocation accuracies. Moreover, the model has shown to be robust against noisy input signals.

In this study, which is only intended to be a proof of concept, the conditions were highly idealized. The lightning return stroke current was modeled using a step function and linearly rising ramp function, the transmission line had an infinite length, was lossless and the ground was a perfectly conducting plane. Therefore, the model does not account for the effects of finite-length, lossy multiconductor transmission lines, the existence of the shield wire, a lossy ground, or the earth topography.

Research is in progress by the authors to investigate the performance of the model for finite length multi conductor transmission lines using advanced models based either on the transmission line theory or full-wave approaches.

Chapter 4

Paper title: Single-Sensor Source Localization Using Electromagnetic Time Reversal and Deep Transfer Learning: Application to Lightning[3]

List of authors: Amirhossein Mostajabi, Hamidreza Karami, Mohammad Azadifar, Alireza Ghasemi, Marcos Rubinstein, and Farhad Rachidi

Author contributions: H.K., A.M., M.A., and A.G. conceived the study. H.K. performed the numerical simulations. A.M. performed the machine learning modeling with support from A.G.. M. A. carried out the experimental recordings. M.R. and F.R. supervised the study and contributed to the interpretation of the results. A.M. led the manuscript preparation with input from all co-authors. All co-authors reviewed the manuscript.

Abstract: Electromagnetic Time Reversal (EMTR) has been used to locate different types of electromagnetic sources. We propose a novel technique based on the combination of EMTR and Machine Learning (ML) for source localization. We show for the first time that ML techniques can be used in conjunction with EMTR to reduce the required number of sensors to only one for the localization of electromagnetic sources in the presence of scatterers. In the EMTR part, we use 2D-FDTD method to generate 2D profiles of the vertical electric field as RGB images. Next, in the ML

[3] Postprint version of the article published in Nature Scientific Reports (DOI: https://doi.org/10.1038/s41598-019-53934-4)

part, we take advantage of transfer learning techniques by using the pretrained VGG-19 Convolutional Neural Network (CNN) as the feature extractor tool. To the best of our knowledge, this is the first time that the knowledge of pretrained CNNs is applied to simulation-generated images. We demonstrate the skill of the developed methodology in localizing two kinds of electromagnetic sources, namely RF sources with a bandwidth of 0.1-10 MHz and lightning impulses. For the localization of lightning, based on the experimental recordings in the Säntis region, the new approach enables accurate 2D lightning localization using only one sensor, as opposed to current lightning location systems that need at least two sensors to operate.

4.1 Introduction

Time reversal (TR) has received increasing attention in the field of source localization for applications in medicine, acoustics, electromagnetics, etc. In electromagnetics, for example, several studies have investigated the use of electromagnetic time reversal (EMTR) as a means of locating lightning (e.g. [34], [35], [120], [121]). Mora et al. [34] and Lugrin et al. [35] proposed an algorithm to locate lightning discharges that requires at least three field sensors and the accuracy of the lightning localization estimations was investigated under ideal and also lossy propagation conditions. Wang et al. [122] applied EMTR to estimate the direction of arrival of lightning radiation sources. Their proposed method was used to map the progression of the whole lightning discharge.

Although EMTR has been proved to have high accuracy in identifying the electromagnetic source locations, for it to yield a good performance, a sufficient number of sensors are required. For example, while the EMTR-based lightning localization approach developed by Mora et al. [34] has yielded excellent accuracy when 3 sensors are used, it does not successfully locate lightning if the number of sensors is lower than 3.

Machine learning algorithms such as neural networks could give computers the ability to learn a skill (such as the prediction of the geographic coordinates of a passive or active object) from sets of archived data and to apply the skill on new unseen data. Recently, machine learning has been also shown to be a useful method for source localization. For example, Huang et al. applied deep neural networks to acoustic source localization in shallow water environments [112]. Vera-Diaz et al. used deep learning to directly estimate the three-dimensional position of a single acoustic source using raw data from microphone arrays [113].

Here, we show as a proof of concept that EMTR and ML can be effectively combined to accurately localize electromagnetic sources using only one electric field sensor. In the EMTR localization technique, usually three steps are defined: (i) several transducers placed in different locations are used to record the electric or magnetic field from the source, (ii) the time domain signals recorded by each of the sensors are then time-reversed and synchronously retransmitted back into the medium, and (iii) the results from the back-propagation step are analyzed to determine the source location [123]. The main idea of this paper is to substitute the conventional methods that are used to make a localization decision based on the back-propagation results (e.g., energy and cross correlation) with machine learning techniques. To do that, the 2D electric field profiles from the output of the back-propagation phase in the EMTR technique are first converted to RGB images and then transmitted to the pretrained VGG-19 Convolutional Neural Network (CNN) in order to extract an efficient feature vector representing these images [124]. Two regressors are then trained, each to estimate the x and y coordinates of the source localization based on the aforementioned extracted features.

Moreover, in some of the studies related to EMTR [34]–[37], [121], [122], the 1/R dependence of the radiated wave in the back-propagation step is removed in order to deal with singularities at the source locations. Back-propagation of the fields by keeping their amplitude constant ensures that the total field will be maximum at the location of the original source and that the fields from the individual sensors will not vanish at very large distances. The maximum constructive interference will necessarily occur at the source point because it is only there that the relative phase delays will vanish if three or more sensors are used. However, forcing the amplitude of the propagating wave to be constant by eliminating the 1/R dependence has two main disadvantages: (i) the condition cannot be applied to numerical methods that are commonly used in many electromagnetic simulators, and (ii) the condition is not applicable in the presence of one or more scatterers in the computational domain. To avoid these disadvantages, some papers have abandoned the condition of constant amplitude and have used instead the entropy technique to find the optimum time slice and then obtain the source location [120], [125].

In this paper, we combine EMTR and Machine Learning to locate radiation sources. We call the combined methodology EMTR/ML. In the EMTR part of the methodology, we use the Two-Dimensional Finite Difference Time Domain (2D-FDTD) method to calculate the direct and back-propagated waves. The simulation results obtained with the full-wave 2D-FDTD method intrinsically

include the propagation losses in the forward and backward time and they improve the accuracy of the source geolocation. Moreover, scattering objects such as mountains can be readily included in the FDTD simulation space. While a 3D model is able to better represent the complexities inside the medium, such as the terrain topography, the actual shape of the scatterers, and the ground conductivity, it is computationally more expensive and time consuming compared to 2D simulations. Hence, in this paper, simple 2D modeling of the scattering objects is used to perform the analyses. We have shown that even using such a simplified model, the proposed method can provide reasonable localization accuracy.

In order to validate the methodology, two case studies are defined. In the first case study, the goal is to localize a Gaussian RF source with 0.1-10 MHz bandwidth based only on numerical simulation results. In the second case study, the methodology is used to find the location of lightning strikes using both numerical and experimental data. In both cases, we will show that only one electric field sensor is needed to reach a reasonable localization accuracy. The presence of mountains or other scatterers is required for the method proposed in this paper to locate the source, since it uses measured fields only at one location. It is shown in Section 4.2.1.C.6 that if the terrain were flat, the ML algorithm would need further information to determine the azimuth to the source due to the symmetry.

4.2 Methods and Results

4.2.1 Case Study 1

The proposed methodology is first applied to localize a Gaussian RF source with a 10 MHz bandwidth using only one electric field sensor. First, the EMTR and ML methods are applied separately to the problem. The results are then compared with that of the proposed combined approach to evaluate the obtained improvement. We used a simple yet realistic geometry as the detection region. It consists of one electric field sensor and two cylindrical scatterers with a radius of 500 m which are put inside the medium as shown in the schematic diagram in Figure 4-1.

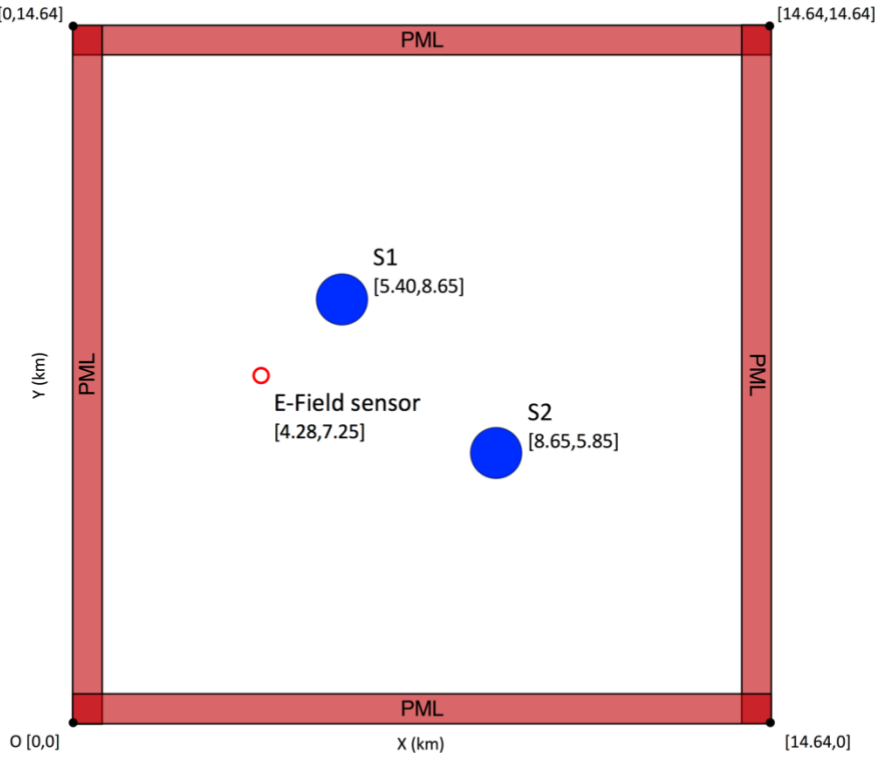


Figure 4-1 Geometry of the problem in the first case study. The blue/filled circles and the red/unfilled circle show the scatterers and the location of the electric field sensor, respectively. The solution space spans 13.44 x 13.44 km2.

A. Numerical Simulation Using EMTR

1) Description of the Procedure

Consider the 2D geometry shown in Figure 4-1. In the EMTR approach, first, a random position for the source was selected. The source was a z-axis dipole with a Gaussian pulse current source of 10 MHz bandwidth and 1 A/m current density amplitude. Then, the incident electric field at the location of the sensor was obtained by means of the 2D-FDTD method. Perfectly Matched Layers (PML) with a depth of 10 mesh cells were deployed as boundary conditions at the perimeter of the simulation domain. Equally spaced cells with the length of 60 m were used to mesh the solution space.

The electric field at the sensor's location was normalized to its maximum value. The end of the sensor signals was set to the time at which the derivative with respect to time of the normalized signal's energy becomes smaller than 10^{-9} 1/s. The recorded signal was then time-reversed and back injected into the medium. Figure 4-2 shows an example of the current density for the excitation

source as well as the corresponding time-reversed version of the electric field calculated at the sensor and normalized to its maximum value. The FDTD method was used again in the back-propagation to calculate the distribution of the electric field inside the medium. Using the maximum amplitude criterion, the EMTR method assumes the source position to be the point with the maximum amplitude of the vertical electric field over the whole simulation time window.

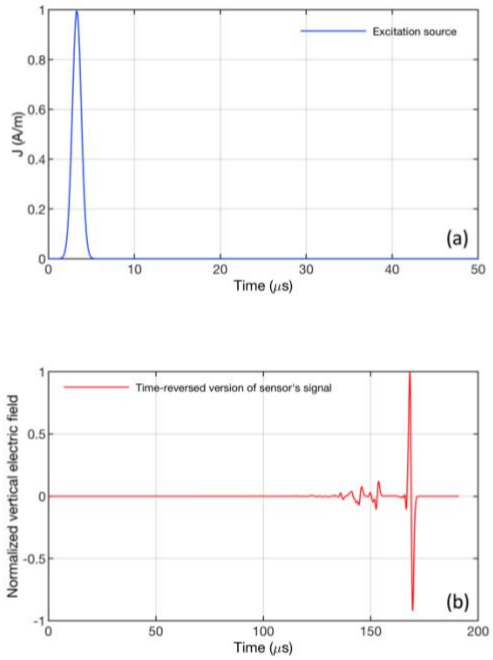


Figure 4-2 Example of the simulation results for the source and the field at a sensor. (a) The linear current density of the excitation source (Gaussian RF with 10 MHz bandwidth) and (b) the time-reversed version of the signal measured by the electric field sensor normalized to its maximum value.

2) Results

The normalized maximum amplitude of the electric field at all time steps is given in Figure 4-3a. As shown in that figure, the maximum amplitude of the electric field occurs at the location of the sensor, which is not the correct source position. This failure to locate the original source was also reported in Karami et al. [120] and it was attributed to the consideration of the 1/R propagation factor in the EMTR calculation.

Moreover, even when the 1/R propagation factor is ignored, the results from several previous studies (e.g. Mora et al. [34] and Lugrin et al. [35]) showed that by using less than three sensors, the EMTR algorithm leads to multiple positions in the time-reversed wavefront with the same value of

the electric field. Instead of getting one unique point, a line of multiple candidates of the source point is observed.

Instead of looking at the results over all of the time steps, some previous studies used the EMTR back-propagation stage results at a specific time slice during the simulation time window to find the source. For example, Karami et al. [120] used the entropy criterion to find the optimal time slice among the back-propagation time steps to locate the lightning strike point. The point with maximum electric field in the selected time slice would correspond to the source position. While this method was successful in localizing sources using 2 and more sensors, it was not applicable in our case study when only one sensor is involved as it led to an ambiguity in the response. To show this ambiguity, let us consider the simulation presented in Figure 4-3a but instead of plotting the maximum electric field over the whole simulation time, let us examine the 2D profile of the electric field at the last time step (i.e., the moment when the back-propagation ends). The resulting plot is shown in Figure 4-3b. In this figure, there are three regions of high electric field inside the 2D profile and each includes several candidates for the source position based on the EMTR approach. Therefore, instead of getting one unique point, a number of multiple candidates of the source point is identified. A similar behavior can be seen if the results from other time slices are considered. This ambiguity was also reported in Mora et al. [34] when less than 3 sensors were used to locate the lightning strike point using EMTR. Furthermore, none of the nominated points corresponds to the actual source position.

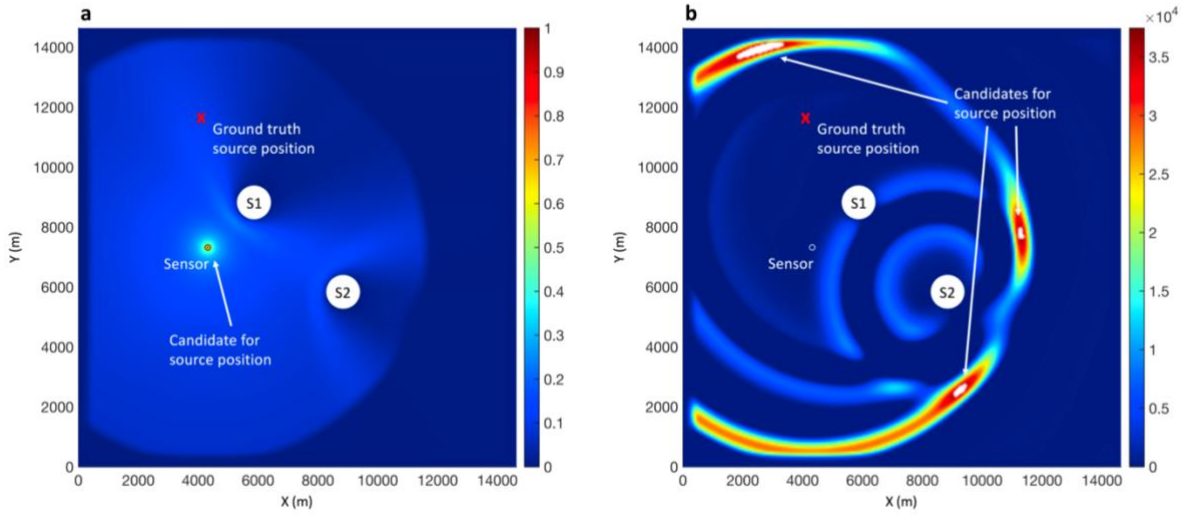


Figure 4-3 (a) Maximum amplitude of vertical electric field at all time steps normalized to its maximum value and (b) 2D profile of the vertical electric field at the last time step. The red cross is the ground truth of the source position, S1 and S2 are the scattering objects, and the white circle is the electric field sensor. The nominated point/points by EMTR are also annotated for on each panel.

B. Numerical Simulation Using Machine Learning

1) Building the Database

Machine learning is an effective approach for approximating a function from a finite set of its input-output pairs, when such approximation cannot be done through parametric estimation or the function has no closed form. In this section, we aim to train a machine learning model to estimate the geolocation of an electromagnetic source, given the data on the incident electric field at a single remote electric field sensor. To do that, we defined more than 1600 random positions for the source within the geometry presented in Figure 4-1. In order to form the required database for training and testing procedures, for each of these source positions we carried out the following steps:

(i) We placed the Gaussian RF source plotted in Figure 4-2a at the selected position and we performed the 2D-FDTD simulation described in Section 4.2.1.A.1 but only in the direct phase to calculate the vertical electric field in the time domain at the location of the sensor.

(ii) Next, the values were normalized to the maximum recorded value. This normalization was done to make sure the results would not be sensitive to the amplitude of the source signal. Similar to Section 4.2.1.A.1, we continued to record the sensor signal until the time derivative of the normalized signal's energy became less than a predefined threshold.

Once all random positions were considered, a preprocessing step was required to make all recorded sensor signals be of the same length. We did so by padding shorter signals with zeros up to a size of N_{max} (i.e., the maximum length among the recorded signals).

We formed a tabular database out of the processed recorded incident electric field values. In each row, the corresponding samples of the normalized electric field measured at the place of the sensor were used as the predictors and the x and y coordinates of the source were used as the response.

2) Machine Learning Modeling

Once the database was formed, a machine learning algorithm was employed to identify regularities between predictors and responses using a portion of the data which is known as the training set.

The model could then use the explored correlations to predict the response for the unseen cases (testing set). A model search process was conducted to choose the most appropriate machine learning regression model. The candidate models were (i) several regression types including regression trees, support vector machines, and gaussian process regression models, (ii) different ensemble methods such as bagging and boosting, and (iii) the neural networks. The results showed that for the prediction of both, the x and y coordinates, the best performance in the sense of the Mean Squared Error (MSE) was achieved using the XGBoost algorithm. Extreme Gradient Boosting (XGBoost) is a variant of the gradient boosting machine which uses a more regularized model formalization to control overfitting [104]. More information on the XGBoost model description and generation can be found in [104].

3) Model Training and Testing

In this study, the predictive ML model was evaluated using a 5-fold cross-validation described as follows. First, the dataset was shuffled and split into five different groups. As a result, each observation in the dataset was assigned to an individual group and remained there for the duration of the training and testing processes. Each unique group was held out from the dataset as the test set and the remaining four groups were used as the training set. The model was then fitted on the training set and evaluated on the test set. The process was repeated until each individual group had been taken once as the test set. The model outputs were combined over the rounds. This splitting method helps to eliminate the leakage of correlated samples from the training set into the testing set. Moreover, it avoids overfitting to a specific testing set since each one of the observations is considered as part of the testing set in one of the five rounds. Once the five rounds of training and testing were done, the model's prediction skill was evaluated by comparing the outputs with the target values for the x and y coordinates.

4) Model Evaluation Results

The model estimations and actual responses were compared to evaluate the model's prediction accuracy. In Figure 4-4a and Figure 4-4b, the evaluation results are presented by means of scatter plots of target versus predicted values for the x and y coordinates. A histogram of the Euclidean distance between the target and the ground truth for each of the considered source positions is given in Figure 4-4c. In order to visualize the error, 2.5% of the total dataset that had the largest localization errors compared to the remaining locations were considered outliers and excluded from

the evaluation results (the same is done in the rest of the paper). The best-fit lines are plotted using the robust least-squares regression method with bisquare weights [126], [127]. The method iteratively reweighted the least-squares algorithm to minimize a weighted sum of squares. At each iteration, the robust weights are computed and given to each data point based on how far it is from the fitted line. In this way, the data are fitted using the usual least-squares approach while the effects of outliers are minimized at the same time [75]. The complete procedure followed by the bisquare method can be found in MathWorks User's Guide [75]. Looking at the results in Figure 4-4, one can see that the model shows poor estimation skill on the data with a median localization error greater than 1300 m. The model's estimation skill is worse on the y coordinate compared to the x coordinate with the coefficient of determination (R^2) decreasing from 0.954 to 0.827 and the Root Mean Squared Error (RMSE) increasing from 688 m to 1302 m.

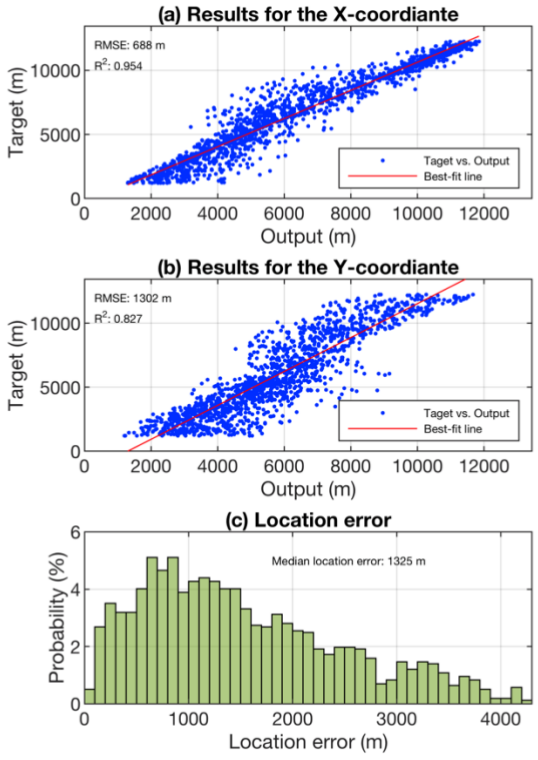


Figure 4-4 Model performance results for the case of numerical simulations using machine learning. (a) Estimation of the x coordinate and (b) the y coordinate of the randomly selected source locations. (c) Histogram of the location error.

C. Numerical Simulation Using the Combination of EMTR and Machine Learning

The results in Figure 4-3 and Figure 4-4 show that applying each of the described EMTR and ML methods separately to the localization problem can lead to inaccurate results when only one sensor

is used. However, by looking more closely into the results, one can see that it is possible to improve the performance of the methods. For example, the ML approach could be improved by feeding more relevant input features rather than just the transient electric field waveform. The back-propagation stage in EMTR is helpful in this regard as it transforms the 1D sensor's signal data into a sequence of 2D arrays (i.e., the profile of the electric field over the detection region and over the back-propagation time window), which includes details regarding the medium such as asymmetries in the location domain (caused by mountainous terrain, scatterers, etc.).

Therefore, teaming up these two methods can make up for the weaknesses of each one and hence increase the performance results. What follows explains the combinational approach proposed in this study.

1) Building the Database

The procedure starts with the building of a database. To that end, the following steps were carried out using the same geometry used in sections 4.2.1.A and 4.2.1.B. First, we generated a pool of random source positions (N>1200) and we followed the steps described in Section 4.2.1.B.1 to form a tabular database out of the sensor signals. Next, instead of simply using these signals as the input for a ML model, we back-propagated the time reversed signal from the location of the sensor using the 2D-FDTD method. At the end of the back-propagation stage, we looked at the 2D array of the vertical electric field associated with the last time step over the detection region. Second, we normalized the vertical electric field values by the maximum value in the 2D array, generated a 2D surface plot out of the data and saved it as an RGB image using the jet colormap array. We then continued with the next source position in the list to generate the next sample image. These images were treated as the sources of valuable visual features in the next steps in the algorithm. A sample of the produced images is shown in Figure 4-5. One should note here that in this figure, the positions of the scatterers, the excitation source, and the sensor are annotated only for the sake of presentation and were not included in the version given to the CNN.

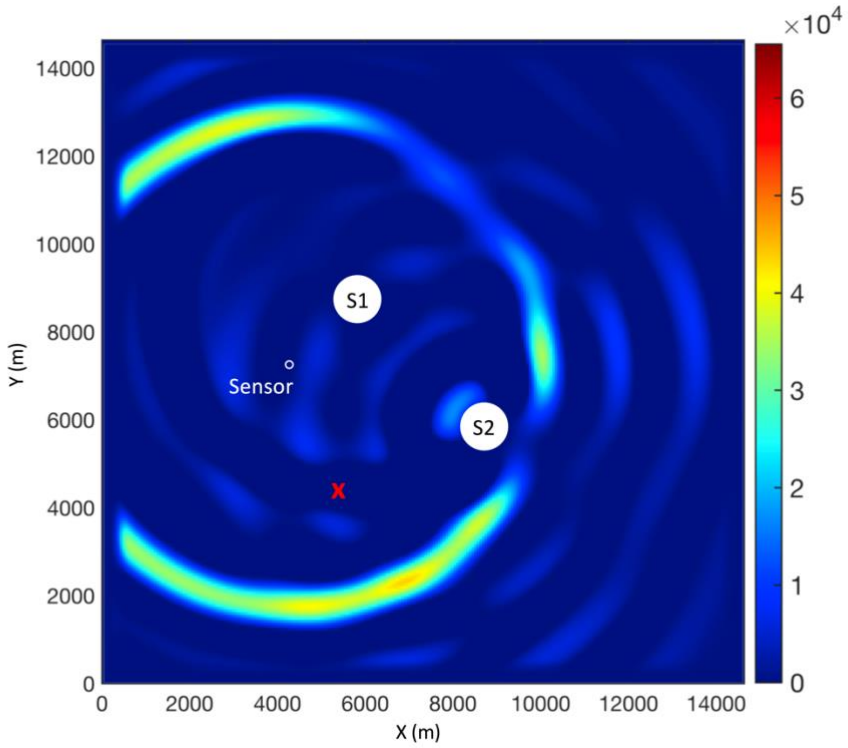


Figure 4-5 A sample of the generated images out of the 2D surface vertical electric field values in the detection region. The corresponding source position is shown as a red cross on the image. The images are produced using the jet colormap array with 216 elements.

2) Feature Extraction

Given our dataset of the images with a size of 224×224 pixels each, a traditional feedforward neural network would require 50176 input weights which would require a very long training time and a large amount of training data. Furthermore, a flattening of the image matrix of pixels to a vector loses the spatial structure in the image, which contains valuable information about scatterer positions, field distributions, etc. To bridge this gap, instead of using the pixel values as the features, Convolutional Neural Networks (CNNs) [128], [129] can be used to automatically learn semantically meaningful features from images. CNNs contain multiple hidden layers, each learning features in increasingly higher levels of semantic granularity from an image. This is achieved by applying various filters to the training image and transmitting the convolved image as the input to the next layer. More information about different types of layers in CNNs and their functionalities can be found in [130].

Although CNNs are proven to be powerful tools in working with data coming in the form of multi-dimensional arrays (e.g. [131]–[140]), there are some limitations if one wants to train the weights

in these networks from scratch: (i) a huge number of examples are needed for the network to understand the variation of features, (ii) the training process requires intensive computational resources and it is usually time consuming, and (iii) configuring the network architecture from scratch could be overwhelming due to the existence of many combinations of network layers. Looking at the architecture of CNNs reveals that they learn in a hierarchical manner [141]. This means that features detected by the first layers are more generic and can be reused in different problem domains, while features computed by the last layers are specific and depend on the chosen dataset and task. This has inspired researchers to take advantage of the available powerful pretrained CNNS, already trained to extract effective features from a huge number of training images, in order to solve tasks rather than the original target for which these networks are trained. In this regard, the internal stages of such pretrained CNNs can be used as a feature extractor tool to detect certain low-level features (that could be shared between the images), such as edges, shapes, corners, and brightness from new collections of images. These extracted features would then be transmitted to a second network to learn the target task. This process is an example of "transfer learning" (e.g. [142]–[144]), in which the representational power of pretrained CNNs can be used to extract some general features from images and hence accelerate the training procedure. This approach is especially of substantial help when the number of available training images is limited and/or the computational resources are not sufficient to train deep learning models from scratch.

In this study, we took advantage of transfer learning to extract features from the generated images using pretrained convolutional networks. As the feature extractor in this study, we imported the Oxford VGG-19 convolutional neural network [124] with the pretrained weights on the ImageNet database [145], [146] as implemented in the Keras library. The network is 19 layers deep with an image input size of 224 x 224. This network has been trained on more than a million images with the original goal of classifying images into 1000 object categories. In order to repurpose the network for our needs, we cut the top layer of the VGG-19 model (i.e., the classifier) and used the rest of the layers as the feature extraction mechanism. By testing each of the images with the repurposed pretrained model, we got a 4096-dimensional feature vector for each image. These vectors were then concatenated to form the dataset for the regression models which further learn to estimate the x and y coordinates of the source positions, as explained in the next section.

3) Regressors on Top of the Pretrained CNN

Once the features were extracted from the images by means of the pretrained VGG-19 network, custom regression layers were needed to do the source localization based on these extracted features. In this regard, we replaced the top layer of VGG-19 with several trainable layers to estimate the x and y coordinates of the source location. Parameter tuning was conducted through cross-validation, where the validation set was separated from both the test and the training sets. In order to make the optimization tractable, we approached the tuning in a greedy manner. We started with the number of layers and we concluded that the best results are achieved with a single hidden layer. We then optimized the number of nodes in the hidden layer, reaching an optimum for 20 nodes. The next step was fixing the gradient descent algorithm, choosing from the set of possible implementations provided by Keras [147]. Adadelta [147] proved the best here, surpassing others with a significant margin. The last step was finding the appropriate activation function; the rectified linear unit (ReLU) function achieved the best result and was chosen for the model.

In order to better understand the internal representation that the model has from an input image, the feature maps from the five main blocks of the VGG-19 model are presented in Figures B-1 to B-5 (Appendix B). The input image is the one presented in Figure 4-5 and the feature maps correspond to the first 64 outputs of the last layer in each block (i.e., layers 2, 5, 10, 15, and 20). As seen in the heat maps of the filter responses in layers 2 through 20 plotted in Figures B-1 to B-5, the filters in the VGG19 layers indeed look for rings of certain sizes and in various locations in the case of EMTR images. From a physical point of view, each ring represents the source response with a certain intensity. However, from a purely data-driven, statistical point of view, namely the way the neural architecture operates, some of the rings can be "computed" from the others and, hence, they are ignored by the network, thereby making the estimation with a finite number of nodes possible.

Continuing this process in consecutive layers, the response maps in the consecutive layers of the network are gradually consolidated into a single value, using a pooling layer in the later stages of the feature engineering network. The computed feature vector is then, very roughly, an indicator of whether a ring with a certain size, location, and intensity has been present in the image or not and, if that is indeed the case, how bright it has been. A super-linear combination of such responses, as learned by the single-layer neural network, estimates the location of the source.

4) Training, Validation, and Testing Procedure

With these new top layers, the predictive neural network was trained, validated and tested using a 5-fold cross validation, as described previously. Fifteen percent of the training samples (i.e., 12% of all samples) were separated and used as the validation set. The model was then fitted on the training set and the performance was evaluated on the validation set at each epoch. The model with the minimum loss on the validation set was used to retain the estimation results on the test set. The model outputs were combined over the rounds. The use of a validation set enables us to make an unbiased evaluation of the model fit while, at the same time, fine-tuning the model hyperparameters. It also helps to keep the test set as an independent subset of data which is only used once the model is completely trained using the training and validation sets. Moreover, the use of cross-validation as the evaluation method avoids overfitting to a specific testing set since each one of the observations is considered as part of the testing set in one of the five rounds. Once the five rounds of training and testing were completed, the model's prediction accuracy was evaluated by comparing the outputs with the target values for the *x* and *y* coordinates.

5) Model Evaluation

Figure 4-6a and Figure 4-6b present scatter plots of the target versus output values in which the blue dots correspond to the observations in the database. The best-fit lines are calculated using the least-squares regression method. The very high values of the coefficient of determination (R^2) indicate that a high proportion of the variance in the data is explained by the fitting line. The results in Figure 4-6a and Figure 4-6b show that the Root Mean Squared Errors (RMSE) are 303 m and 385 m for the x and y coordinates, respectively. These low error values reveal that despite having used only single-site electric field data, the model was able to accurately predict the location of the nearby RF source along the x and y axes. Finally, the relative probability of the location errors between the ground truth 2D source position and the estimated one by the model for each of the considered source positions is given in the histogram of Figure 4-6c. The height of the bar in each bin is the relative number of observations that fall into that bin. According to the results, in more than 75% of the cases, the model was able to predict the source with less than 600-m location error and the median location error was 389 m.

In order to visualize the location error as a function of the source position inside the detection region, a heatmap of location errors is shown in Figure 4-7. The figure shows, for each grid cell, the average of the location errors estimated by the model for the samples that fell into that cell. The

colormap scale indicates the location error in meters. The 0 values around the perimeter correspond to the PML marginal layer (see Figure 4-1).

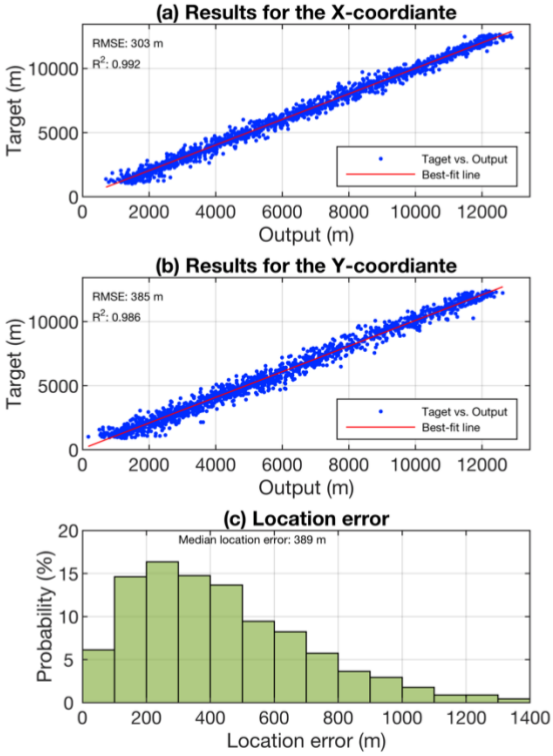


Figure 4-6 Model performance results for the case of numerical simulations using combinational EMTR/ML approach. The detection region is shown in Figure 4-1 including two scattering objects. (a) Estimation of the x coordinate and (b) the y coordinate of the randomly selected source locations. (c) Histogram of the location error.

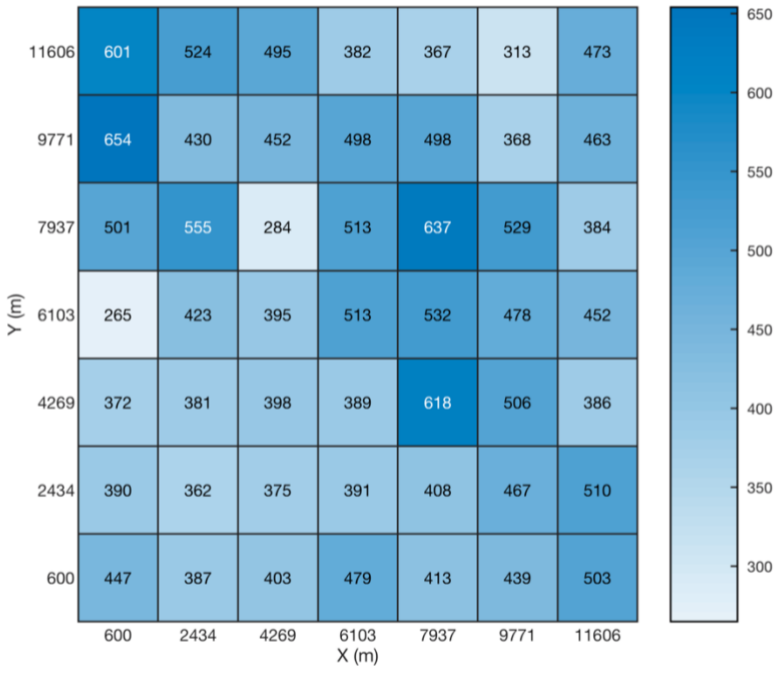


Figure 4-7 Average location error presented as a heatmap chart inside the detection region for the case of numerical simulations using the EMTR/ML combinational approach. The size of the scatterers, their number and locations, and the frequency of the excitation source remained fixed for simulations at each of the source locations. The colormap represents the location errors in meters. The (0,0) point corresponds to the coordinate center (O) shown in Figure 4-1. The x and y labels for each of the cells are the coordinates of the bottom-left corner of the grid cell.

6) Sensitivity to the source and medium parameters

So far, the size of the scatterers, their number and locations, and the frequency of the source remained fixed. In practical cases, knowledge of these parameters may be imperfect or incomplete. We therefore investigated the sensitivity of the model to variations in these parameters by redoing the analysis described in Section 4.2.1.C with the following changes to the detection region shown in Figure 4-1 before building the database: (i) either a third scatterer was added randomly to some of the cases with its center at [8.05, 8.05] (making the overall number of scatterers be 3), or the second scatterer, S2 in Figure 4-1, was removed randomly from some of the cases (i.e., only S1 exists in the medium), (ii) the radius of each of the scatterers was selected independently between 100-500 m, (iii) the frequency of the Gaussian RF excitation source was changed randomly to values between 100 kHz and 10 MHz, and (iv) the number of source locations was increased to more than 6700. The rest of the analysis, including the model generation, training, validation and testing methods were the same as the ones explained in sections 4.2.1.C.2 and 4.2.1.C.3.

The evaluation results showed the median and mean for the location error to be 258 m and 429 m, respectively. The obtained results in Figure 4-8 reveal similar location accuracies to the ones achieved when using fixed parameters for the source and the scatterers. The results show that the proposed approach still yields reasonable performance even when the source and the medium are only partially known and modeled.

Furthermore, the model finds its way to the source by looking at the disturbances caused by reflections from the scatterers inside the images of the 2D electric field profile. Our further analysis showed that the presence of such asymmetries inside the region is a requisite for the EMTR/ML approach to determine the 2D location of the source. So far, the study in Section 4.2.1.C has been done considering two scatterers inside the medium. Once reducing the number of scatterers to only one, in the general case there would be an ambiguity in determining the position of the source with the data from only one sensor. The reason is that, with only one symmetrical scatterer present inside the medium as shown in Figure B-6, the sensor would receive the same signal if the excitation source is put at point A or A', where A is an arbitrary point in the computational domain and A' is

the mirror of A with respect to the symmetry line. In the case of having only one scatterer, the aforementioned ambiguity could be resolved by adding a magnetic field sensor at the place of the existing electric field sensor to determine the direction of arrival. Furthermore, the ambiguity would not exist if the scatterer itself were asymmetrical. As a result, the minimum number of scatterers is one. In case of symmetrical scatterers, the minimum number to avoid ambiguities increases to two.

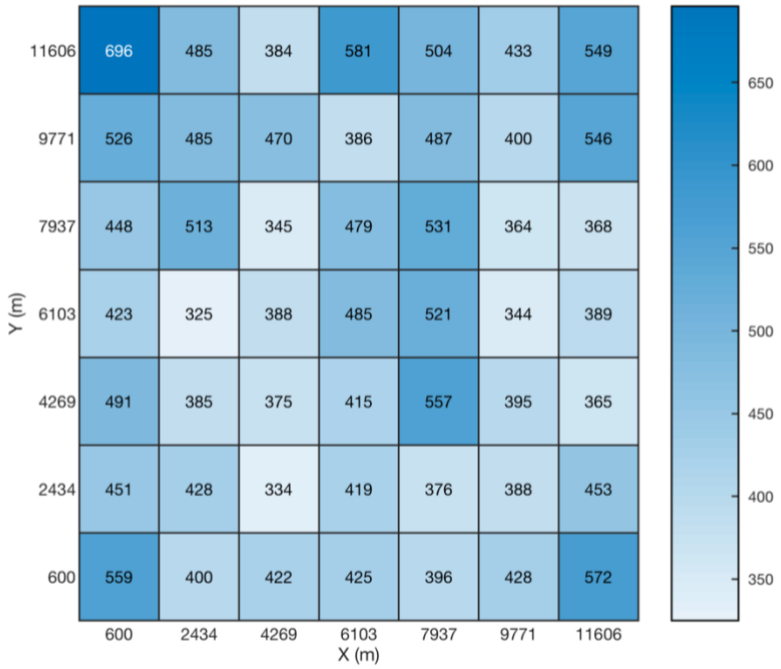


Figure 4-8 Average location error presented as a heatmap chart inside the detection region for the case of numerical simulations using the combinational EMTR/ML approach. A third scatterer, was added randomly to some of the cases with its center at [8.05, 8.05]. The size of the scatterers as well as the frequency of the excitation source were selected randomly for simulations at each of the source locations. The colormap represents the location errors in meters. The (0,0) point corresponds to the coordinate center (O) shown in Figure 4-1. The x and y labels for each of the cells are the coordinates of the bottom-left corner of the grid cell.

4.2.2 Case study 2: Application to Locate Lightning Flashes

Several methods allow to estimate the location of lightning using single-station measurements. They typically use the Poynting vector to determine the direction to the lightning and characteristics of the received electromagnetic fields to estimate the distance to the lightning. A review of single-station techniques is given in Rafalsky et al. [148].

Note that existing single-station techniques require at least two different sensors to operate and they suffer from poor accuracy. To obtain mean location accuracies of the order of a few hundred meters, an array of multiple stations is used [33]–[36], [120]. In what follows, we illustrate how the

proposed algorithm in Section 4.2.1.C can be used to locate cloud-to-ground lightning flashes in the Säntis region with acceptable accuracy using only data from a single electric field sensor at a single station. What follows will briefly describe the experimental electric field measurement system, the applied methodology, and the experimental validation results.

A. Electric Field Measurement System at Herisau

A study of the incidence of lightning to different towers in Switzerland showed the Säntis Tower is struck by lightning some 100 times a year [98],[99], making it far and away the most struck structure in the country. This telecommunications tower has a height of 123.5-m and it was erected in 1997 on the mountain from which it takes its name, the 2502-m Mount Säntis in the in northeastern part of Switzerland.

Sensors to measure the current and the current derivative from direct lightning strikes to the tower were installed at two heights and have been in operation since 2010 [100], [149], [150].

Sensors to measure the vertical electric fields and the horizontal magnetic field from lightning flashes to the tower were set up at a distance of 14.7 km in 2014, on top of a 25-m tall building in Herisau [150]. The wideband electric and magnetic sensors, including fiber optics for the signal transmission, were manufactured by Thomson CSF (currently Thales). The operation frequency of the vertical electric field measurement subsystem goes from 1 kHz to 150 MHz. Since the return stroke field waveforms are bandlimited to a few MHz [151], we used a digitization rate of 50 MS/s, which is sufficiently high to avoid aliasing. The magnetic field measurement system has the same upper cut-off frequency but its lower cutoff is at 2 kHz [150].

B. Implementation of the Combined EMTR and ML Approach

The experimental validation of the proposed combinational approach was carried out using data from a vertical electric field sensor installed at Herisau, at 14.7 km from the Säntis tower.

Consider a 15.78 x 15.78 km^2 geographical grid including the Säntis Tower as shown in Figure 4-9. We consider the electric field measurement system at Herisau to be the required sensor by the EMTR method. Four actual tall mountains around the Säntis were modeled as cylindrical scatterers. The scatterers had a radius $r = 240\ m$, and electrical parameters $\varepsilon_r = 10$, and $\sigma = 0.05\ S/m$. The geographical coordinates of the elements shown in Figure 4-9 are given in Table 4-1.

Figure 4-9 Geometry of the problem in the second case study, the blue/filled circles represent four actual tall mountains around the Säntis, modeled as cylinders. The red/unfilled circle shows the electric field sensor at Herisau. The red cross is the position of the Säntis Tower. The solution space spans 15.78 x 15.78 km2.

Table 4-1 Coordinates of the elements shown in Figure 4-9.

Item	Description	Latitude (˚N)	Longitude (˚E)
Säntis Tower	Lightning strike point	47.25	9.34
Herisau	Electric field sensor	47.39	9.27
Altenalp	Scatterer	47.27	9.38
Hochalp	Scatterer	47.28	9.25
Hundwiler Höhi	Scatterer	47.34	9.33
Kronberg	Scatterer	47.29	9.33

The European Cooperation for Lightning Detection (EUCLID) [91] is a cooperation of several national lightning location networks aiming at providing lightning data all over Europe. The Säntis area is covered by six EUCLID sensors. In 2016, Azadifar et al. [152] analyzed the performance of EUCLID for 269 upward lighting flashes striking the Säntis Tower from June 2010 to December 2013. The recorded flashes contain 2795 pulses (including return strokes or Initial Continuous Current pulses with a risetime lower than 8 μs and an amplitude greater than 2 kA). The results (see Figure 4-11) showed the median, mean and maximum pulse location errors to be 186 m, 447 m, and above 5000 m respectively.

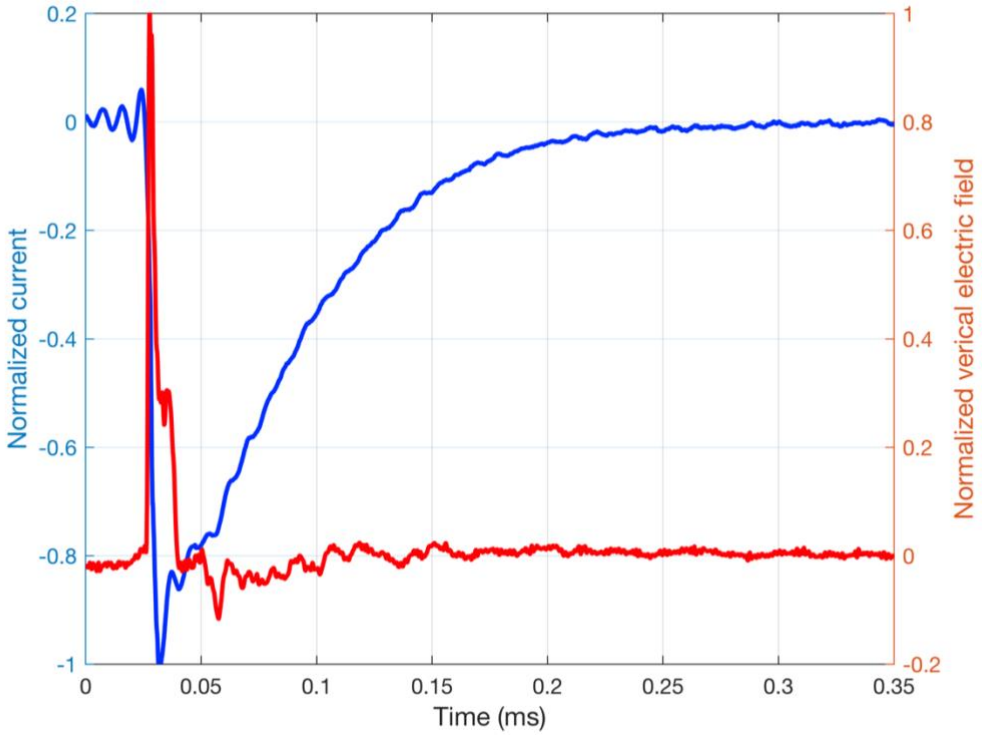


Figure 4-10 Simultaneous recordings of the vertical electric field at Herisau and the channel-base current associated with a return stroke of an upward lightning flash that occurred on on July 18th, 2017 at 18:31:35 (local time) at the Säntis Tower. The vertical electric field data were used for the experimental validation of the proposed approach.

In this study, the source positions used to build the database were taken from the EUCLID lightning location network in the region shown in Figure 4-9 between March 2016 to July 2017. The database was built from 1296 geolocations that were randomly selected out of all of the geolocations in that region and time period. The lightning was modeled as a z-axis dipole current source excited by a Heidler current whose parameters are those used in Karami et al. [153]. The ML model (which consists of layers from VGG-19 with pretrained weights and the added regressors at the top) were then trained on the database. As the testing set, we used the experimental records of the electric field at Herisau for six return strokes (hereafter called RS1-RS6) corresponding to two upward negative lightning flashes occurred on October 21st, 2014 at 20:23:22 (local time) and July 18th, 2017 at 18:31:35 (local time). Figure 4-10 shows the electric field waveform associated with RS1. More information about the characteristics of the current waveforms of these return strokes is given in Table 4-2 In order to make the experimental data compatible with the input of the model, we carried out the following steps: (i) denoised the signals using MATLAB's Wavelet Signal Denoiser [154] and normalized the signal to its maximum value, (ii) time-reversed the normalized signal numerically and back-injected it into the medium using the 2D-FDTD method described in Section 4.2.1.A.1 and the geometry shown in Figure 4-9, and (iii) generated an RGB image out of the 2D data

inside the detection region. Once the image was generated, the pretrained VGG-19 model was used to extract the feature vectors from it. This vector was then regarded as the testing sample and it was passed to the fitted model to evaluate its performance. In the second case study, the complications arising due to the sensor noise and weather conditions made the final regression problem have a higher complexity level. As a result, the accuracy of a neural network with a single point of non-linearity in the activation function deteriorates, making it unsuitable for solving the localization problem. Increasing the number of layers may allow incorporating more complicated models, but successful training of such a network requires much more training data than available. This led us to consider gradient boosting as an alternative, which makes it possible to estimate highly nonlinear functions with relatively fewer samples. An ensemble of 100 trees with depths capped to 3 to avoid overfitting achieved the best result for estimating the x and y coordinates in this problem.

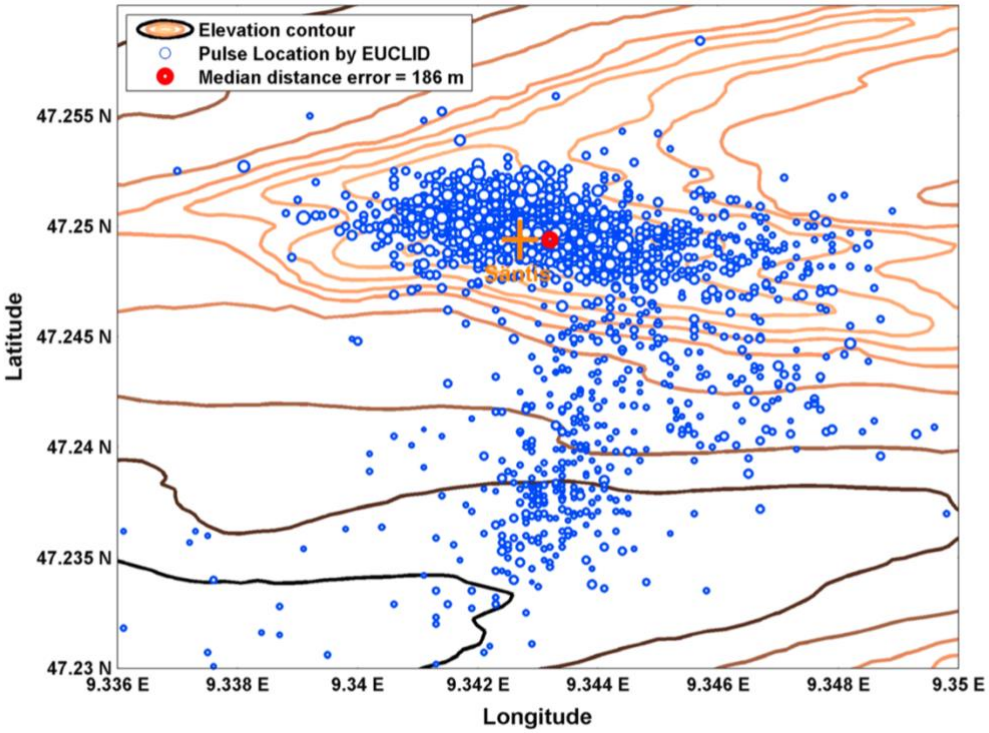


Figure 4-11 EUCLID pulse location errors for upward negative flashes striking the Säntis Tower recorded in the period of June 2010 until December 2013. The ground truth target for all pulses is the Säntis Tower and the estimated locations by EUCLID are presented as blue dots. The size of the dots is proportional to the pulse peak current amplitude. The length and width of the shown area are, respectively, 3.34 and 1.06 km. The figure is adopted from Azadifar et al. [152].

C. Experimental Validation Results

The location errors for each of the tested RSs are reported in table 4-2. The model's prediction can also be seen visually in Figure 4-12. As mentioned above, the task was to estimate the lightning strike point (i.e., the Säntis Tower) by looking at the single-sensor measurement of its associated electric field recorded 14.7 km away from the lightning channel. It can be seen that the model was able to locate the strike point with mean and median errors of 253 m and 328 m, respectively. The achieved result can be interpreted as a very good estimation skill for the model given the facts that (i) the detection area is mountainous, which highly affects the wave propagation inside the medium and (ii) only a simplified modeling of the geometry was used during the 2D-FDTD simulations.

A discussion is in order concerning the experimental validation and the conditions under which the combinational method presented here is applicable.

Although the performance of the combinational method in terms of location accuracy is excellent, both in the simulations and based on measured electric field data, further experimental confirmation is needed before any definitive conclusions can be drawn concerning the location accuracy of the method since lightning flashes at one source position were used for the experimental validation.

Finally, note also that, as mentioned in Section 4.2.1.C, the fact that the terrain contains mountains or other scatterers is a requirement for the proposed combinational method to locate the discharges. Indeed, if the terrain were flat, the symmetry would prevent the ML algorithm, in the absence of additional information on the source, from determining the azimuth to the lightning.

4.3 Discussion

In this paper, machine learning and electromagnetic time were paired to introduce a new single-sensor electromagnetic source localization technique. In the proposed combinational model, on the one hand, EMTR is used as a preprocessing stage to convert a measured transient electric signal into more relevant input data (i.e., corresponding more to the target of determining the location of the source) for the ML model. On the other hand, machine learning is used to fit the back-propagation results to the target without any explicit underlying teleconnections available in the data. The method requires the presence of at least one scatterer as long as its shape is asymmetrical. In case of symmetrical scatterers, the minimum number to avoid ambiguities increases to two.

Table 4-2 Characteristics of the current waveforms associated with the considered 6 return strokes (RS1 to RS6) and their corresponding location error using the EMTR/ML approach

Return stroke	Peak amplitude current (kA)	Zero to peak rise time (μs)	Full width at half maximum (FWHM) (μs)	Location error (m)
RS1	12.4	8.2	52.2	365
RS2	4.5	5.5	50.8	347
RS3	3.4	5.8	48.3	330
RS4	2.7	5.6	44.9	325
RS5	8.1	5	114	39
RS6	13.2	1.3	29.1	113

A 2D-FDTD method was used to calculate the electric field distribution on the detection region and generate the 2D profiles of vertical electric field as RGB images. Next, employing transfer learning, a pretrained VGG-19 Convolutional Neural Network (CNN) was used as the feature extractor tool and it was fed by the aforementioned simulation-generated images. Finally, two regressors were trained on the extracted features to do the final estimation of the source position inside the solution panel.

The method was first illustrated using a numerical example to localize a Gaussian RF source with 10 MHz bandwidth. A sensitivity analysis was presented using alternative parameters, such as the size and number of scatterers and the frequency of the excitation source.

The proposed method was then applied to the localization of lightning discharges in the Säntis region in Northeastern Switzerland. The model was trained based on the simulation results and tested using experimental observations of lightning flashes in the Säntis region. The experimental validation results show a high estimation accuracy for the combinational approach in finding the 2D geolocation of the lightning strike point at the Säntis Tower using only one sensor. The comparison to the results from EUCLID (given in Section 4.2.2.B) reveals that the proposed approach can yield similar location accuracy with a significantly smaller number of sensors.

An extensive amount of work is in progress by the authors to further validate the proposed approach using more experimental cases and to consider 3D modeling to better simulate the medium and the excitation source details.

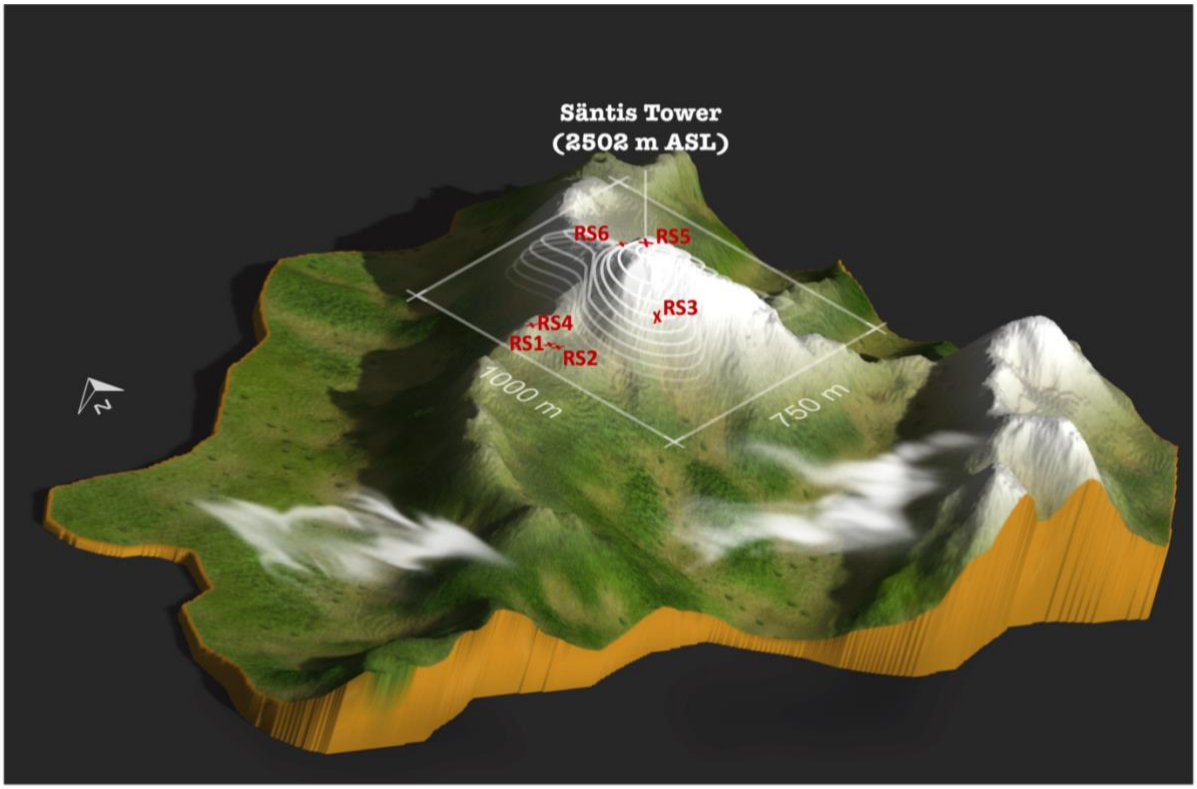


Figure 4-12 Experimental validation result for the proposed combinational approach: six return strokes (RS1-6) associated with two upward lightning flash that occurred at the Säntis Tower were the excitation source and the single-sensor recording of the associated electric fields 14.7 km away were the input data for the model. The estimated lightning strike point for each of the RSs is also shown as a red cross around the target (i.e., the Säntis Tower). The map is generated using 3D Map Generator–Atlas plugin (https://graphicriver.net/item/3d-map-generator-atlas-from-heightmap-to-real-3d- map/22277498) for Adobe Photoshop CC 2017.1.1 release.

Chapter 5

Conclusion

This thesis was focused on two major topics: prediction and localization of lightning flashes.

With regard to lightning prediction, we developed an ML-based lightning nowcasting system based on four commonly-available surface weather variables. The produced warnings were validated using the data from lightning location systems, showing that the model has statistically-considerable predictive skill for lead times up to 30 minutes. Our model also outperforms existing approaches, namely those based on the persistence method, or the widely-used method based on a threshold of the vertical electrostatic field. One of the main advantages of the developed nowcasting system is that it is independent of external sources of data such as numerical model outputs, satellite and radar. Specifically, the method can provide information in areas where radars are not present, where weather forecast resources are limited, or where nowcasting is not in operation, for instance in isolated areas in low-income countries in Asia, South America, and Africa.

Regarding the localization of lightning flashes, classical localization techniques, which are based essentially on the Time of Arrival (ToA) technique, suffer from a number of shortcomings such as sensitivity to the terrain profile, misclassification, and sometimes insufficient performance in terms of location accuracy and detection efficiency. We developed two novel lighting localization techniques:

The first technique is a machine learning based lightning localization algorithm that utilizes data from two preinstalled voltage measurement systems on power transmission or distribution lines to estimate lightning strike impact points. The algorithm was trained using synthetic data (solutions

for lightning-induced voltages on transmission lines) and it showed reasonable geolocation accuracy for grids up to 100 x 100 km^2.

The second model is based on the combination of Electromagnetic Time Reversal (EMTR) and Machine Learning (ML) for locating lightning impact points. Our studies have shown that the proposed combination allows to reduce the required number of sensors to only one. Furthermore, the terrain profile and the presence of inhomogeneities and scatterers does not impair the performance of the system. Our method was applied to the localization of lightning discharges in the Säntis region in Northeastern Switzerland. In that region, the Säntis telecommunications tower, which is often struck by lightning, is instrumented for lightning measurements. The model was trained based on the simulation results and tested using experimental observations of lightning flashes in the Säntis region. The experimental validation results confirmed a high location accuracy for the proposed hybrid approach in finding the 2D geolocation of the lightning strike point at the Säntis Tower using only one sensor. We also showed that the proposed approach yields similar location accuracies to those obtained with the European Lightning Detection Network (EUCLID), but with a significantly smaller number of sensors.

5.1 Future directions

5.1.1 Lightning nowcasting map

In this thesis, we deliberately restricted our efforts to the selection of surface data since we wanted the method to be easy to implement and widely applicable to a variety of vulnerable sites. In fact, our idea behind the choice of input variables for the developed nowcasting technique was to use types of predictors that are commonly available, have high temporal resolution, and are easy and fast to retrieve in real time. However, a huge amount of atmospheric data is available from numerical model outputs, atmospheric soundings, satellite, and radar observations. Furthermore, lightning activity is now readily detected with high spatiotemporal resolution by means of either space-borne instruments (e.g., Global Lightning Mapper (GLM) aboard GOES 17) or ground-based lightning location systems (e.g., EUCLID). The availability of both atmospheric and lightning data with high spatiotemporal resolution should allow the development of ML-based lightning nowcasting schemes that can generate a forecasting map over a large area. Three types of models can be devised by trying different feeding scenarios: (i) only atmospheric measurements (physics-

based), (ii) only the lightning data from previous time windows (data-driven), and (iii) fusing the atmospheric data with the lightning data (hybrid model). We have started the development of a fully data-driven approach with enhanced prediction speed based on deep neural networks. The developed lightning nowcasting model is based on a residual U-net architecture. Our dataset consists of post-processed data of recorded lightning occurrences in 15-minute intervals over 60 days obtained from the GOES satellite over the Americas. We have optimized the model using data from the northern part of South America, a region characterized by high lightning activity. The model was then applied to other regions of the Americas. We are using a 70-15-15% separation for the training, validation, and test datasets. Upon completion of the training process, the model can achieve an overall F1 score of 70% with a lead time of 30 minutes over South America in fractions of a second. This is more than a 25% increase in the F1 score compared to the persistent model, which is used as our baseline forecast method. We suggest to investigate the other two scenarios (physics-based and hybrid model) to define the best nowcasting scheme in future studies.

5.1.2 RF Machine Perception

Our work on single-sensor source localization allows intelligent sensing algorithms that can localize active objects (including RF sources such as routers, cell phones, harmful radiation zone) in either small rooms or big landscapes (large source of electromagnetic pulses such as a lightning strike) using only one sensor. We have started a research collaboration with Apple Inc. in California to use this technique as an electromagnetic scanner for certain Apple products. The outcome of the scanner will allow hardware designers to perform root cause failure analysis and complex EMC troubleshooting.

Moreover, different frequencies in the RF spectrum can have different interactions with the environment. Hence, extracting meaningful insights from these complex interactions is a challenging task. Fortunately, machine learning techniques, especially modern deep learning frameworks, have shown to be effective tools in pattern analysis of multiple data types. Exploring this confluence between multiple RF devices, leveraging advanced techniques in computational imaging such as EMTR, and applying state-of-the-art machine learning and pattern recognition techniques enable interesting sensing capabilities, ranging from scanning the environment even in the presence of walls and in poor lighting conditions, to autonomous perception in severe weather conditions, to non-invasive X-ray vision for security, to the detection of cancer biomarkers. The

proposed EMTR/ML combination can also be used to bring new perception capabilities and use cases to Mixed reality (MR) devices (like Microsoft HoloLens) and robots by adding a new intelligent layer that localizes and tracks sources/devices/agents in the environment. In the case of HoloLens, for example, this means that the headset would be able to spot the routers, cell phones, or any other electromagnetic sources in the scene. Such information could then be used by developers to create interesting features in games, add specific holograms for the spotted sources, or help users keep a safe distance from highly radiative devices.

Appendix A: Supplementary Information for Chapter 2

Table A-1 Geographical information of studied stations

No#	Station	Altitude ASL (m)	Canton
1	**Säntis**	2502	Appenzell Inner Rhodes
2	**Lugano**	273	Tessin
3	**Schaffhausen**	438	Schaffhausen
4	**Luzern**	454	Lucerne
5	**La Dôle**	1670	Waadt
6	**Glarus**	517	Glarus
7	**Engelberg**	1036	Obwald
8	**Basel / Binningen**	316	Basle-Country
9	**Bern / Zollikofen**	553	Bern
10	**Genève / Cointrin**	411	Geneva
11	**St. Gallen**	776	Saint Gall
12	**Zürich / Fluntern**	556	Zurich

Table A-2 A sample of contingency table for a two-class prediction scheme

		Event observed	
		Yes	**No**
Event forecast	**Yes**	Hit (H)	False Alarm (FA)
	No	Miss (M)	Correct rejection (C)

Table A-3 Specifications of the engaged data subsets

No#	Station	Included features				Labeling	Starting date	Ending date
		Air pressure (QFE)	Air tempearture	Relative humidity	Wind speed	Type of lightning activity		
1	**Säntis**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
2	**Monte San Salvatore**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
3	**Schaffhausen**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
4	**Luzern**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
5	**La Dôle**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
6	**Glarus**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
7	**Engelberg**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
8	**Basel**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
9	**Bern**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
10	**Geneva**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
11	**St. gallen**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017
12	**Zurich**	●	●	●	●	long-range	1-Jan-2006	31-Dec-2017

* considering the imminent lightning threat (lead time = 0-10 minutes)

Table A-4 List of measuring instruments for meteorological observations from SwissMetNet [95]

Parameter	Brand	Model	Measuring range*	Resolution*	Accuracy*
Surface pressure at the station level	Meteolabor	Ventilated thermo-hygrometer Thygan (VTP6, VTP37)	-50 ˚C to 50 ˚C	0.01 ˚C	± 0.15 ˚K from -20 ˚C to +50˚C ± 0.25 ˚K from -65˚C to -20 ˚C
Surface temperature at 2 m above ground	Vaisala	PTB330	500 to 1100 hPa	0.1 hPa	±0.4 hPa at 20 ˚C
Relative humidity	Meteolabor	Ventilated thermo-hygrometer Thygan (VTP6, VTP37)	-	0.1 %	-
Wind speed	Lambrecht	L14512	0.4 to 60 m/s	0.1 m/s	± 2% of the measured value
	Thies	US-Anemometer 2D	0.1 to 75 m/s	0.1 m/s	± 0.1 m/s for values ≤ 5 m/s ± 2% of the measured value for values ≥ 5 m/s

*values correspond to the basic available version/class of the device

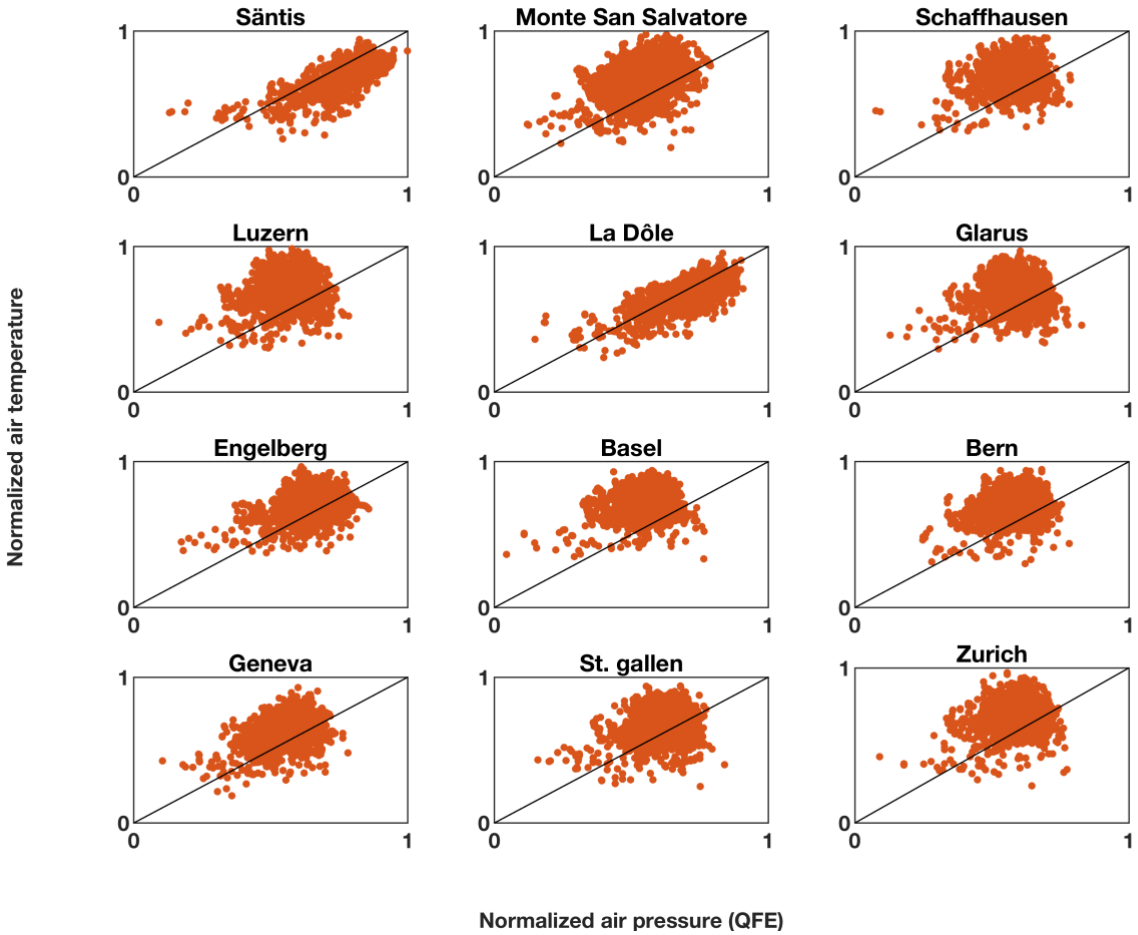


Figure A-1 Distribution of the lightning-active class samples for 12 stations during 2006 to 2017. In each subplot, the horizontal axis is the air pressure at the station level (QFE) and the vertical axis is the air temperature. For the sake of comparison, the parameter values at each station are standardized based on the local mean and standard deviation values. The standardizing function sets the mean of each parameter to zero and scales the parameters by their standard deviations. In this analysis, samples are classified as lightning-active or lightning-inactive based on the imminent long-range lightning activity (i.e. lead time = 0-10 minutes).

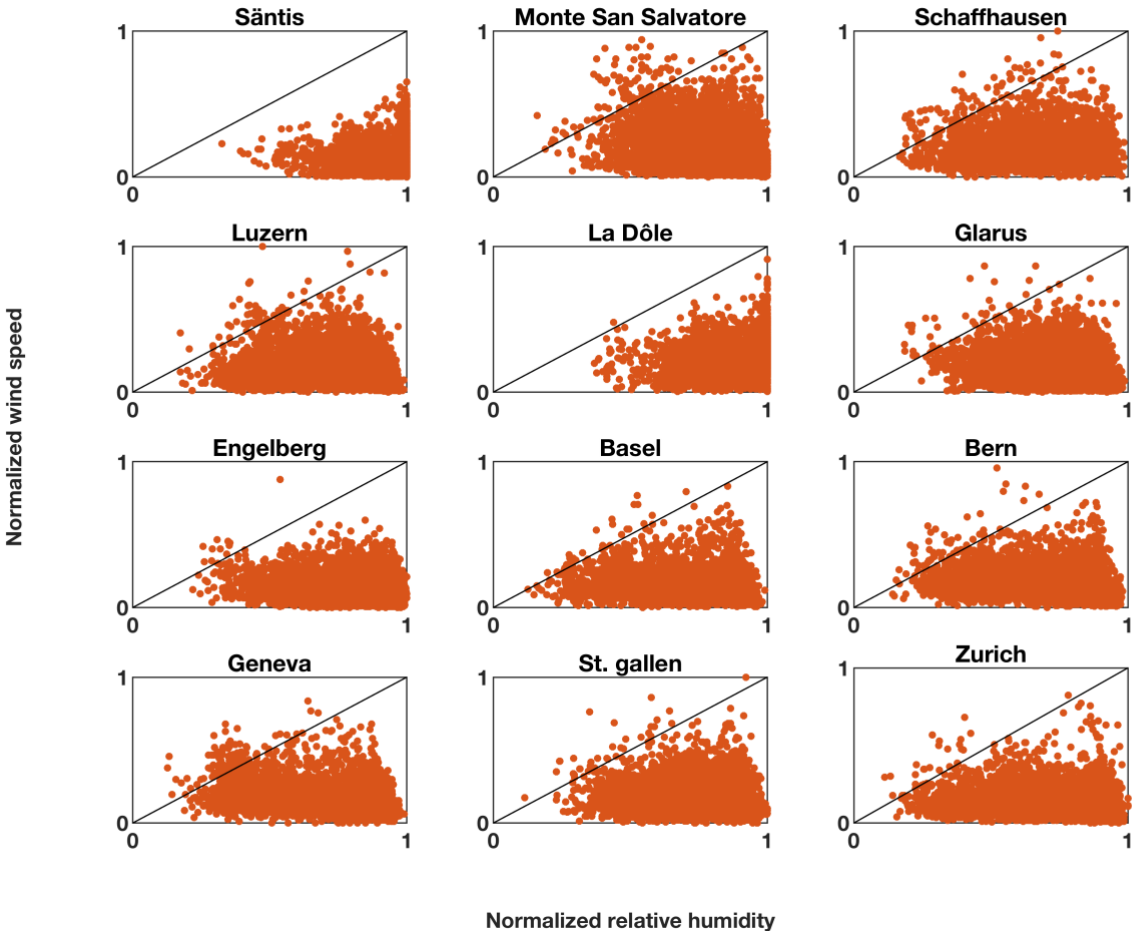


Figure A-2 Distribution of lightning-active class samples for 12 stations during 2006 to 2017. In each subplot, the horizontal axis is the relative humidity and the vertical axis is the wind speed. For the sake of comparison, the parameter values at each station are standardized based on the local mean and standard deviation values. The standardizing function sets the mean of each parameter to zero and scales the parameters by their standard deviations. In this analysis, samples are classified as lightning-active or lightning-inactive based on the imminent long-range lightning activity (i.e. lead time = 0-10 minutes).

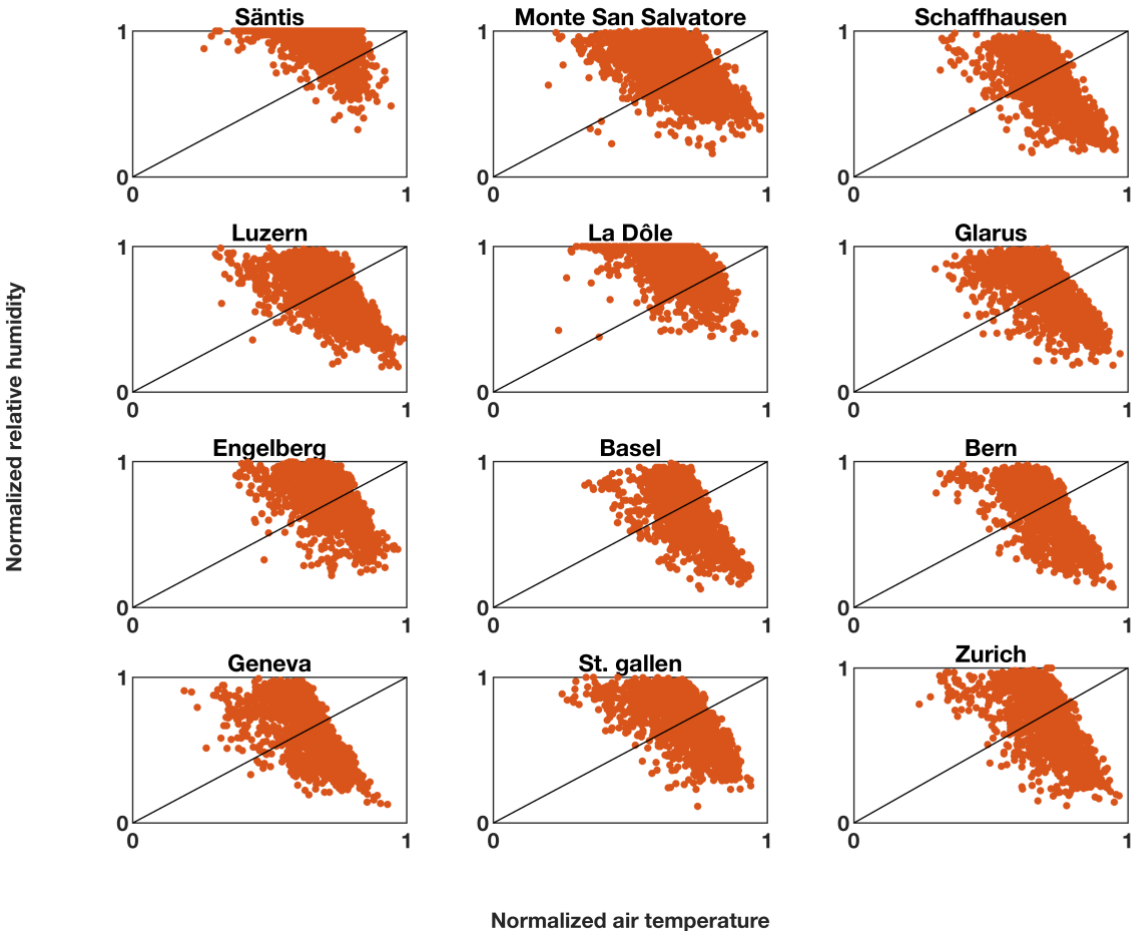


Figure A-3 Distribution of lightning-active class samples for 12 stations during 2006 to 2017. In each subplot, the horizontal axis is the air temperature and the vertical axis is the relative humidity. For the sake of comparison, the parameters values at each station are standardized based on the local mean and standard deviation values. The standardizing function sets the mean of each parameter to zero and scales the parameters by their standard deviations. In this analysis, samples are classified as lightning-active or lightning-inactive based on the imminent long-range lightning activity (i.e. lead time = 0-10 minutes).

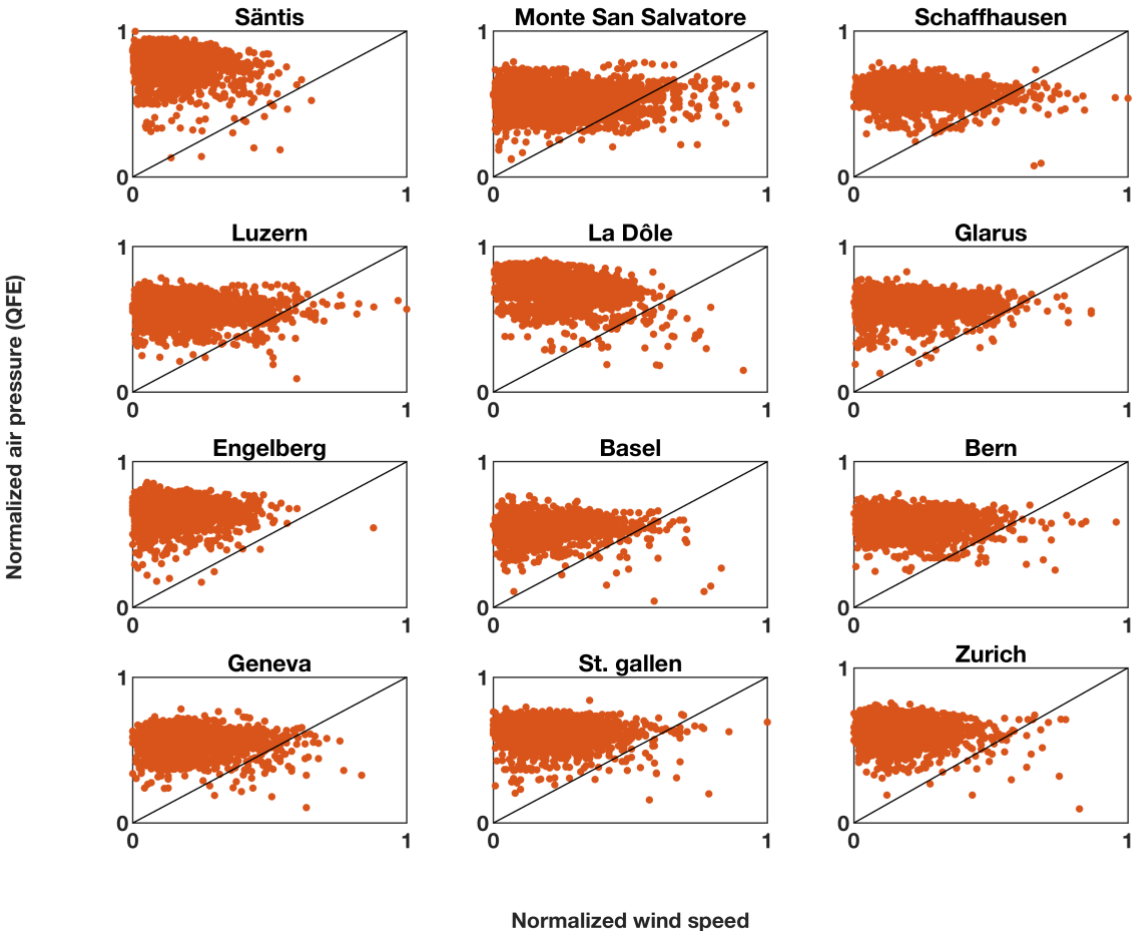


Figure A-4 Distribution of lightning-active class samples for 12 stations during 2006 to 2017. In each subplot, the horizontal axis is the wind speed and the vertical axis is the air pressure at the station level (QFE). For the sake of comparison, the parameters values at each station are standardized based on the local mean and standard deviation values. The standardizing function sets the mean of each parameter to zero and scales the parameters by their standard deviations. In this analysis, samples are classified as lightning-active or lightning-inactive based on the imminent long-range lightning activity (i.e. lead time = 0-10 minutes).

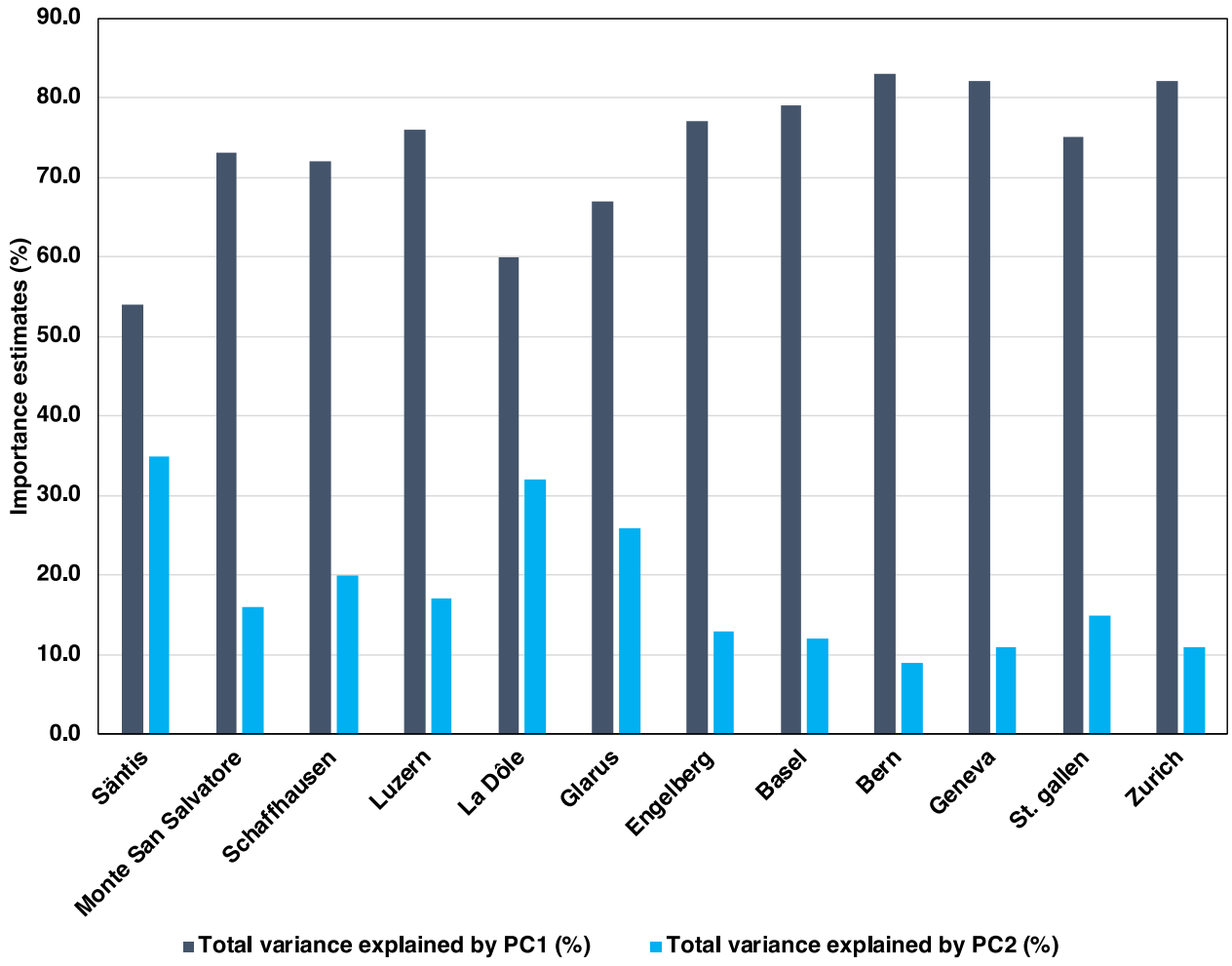
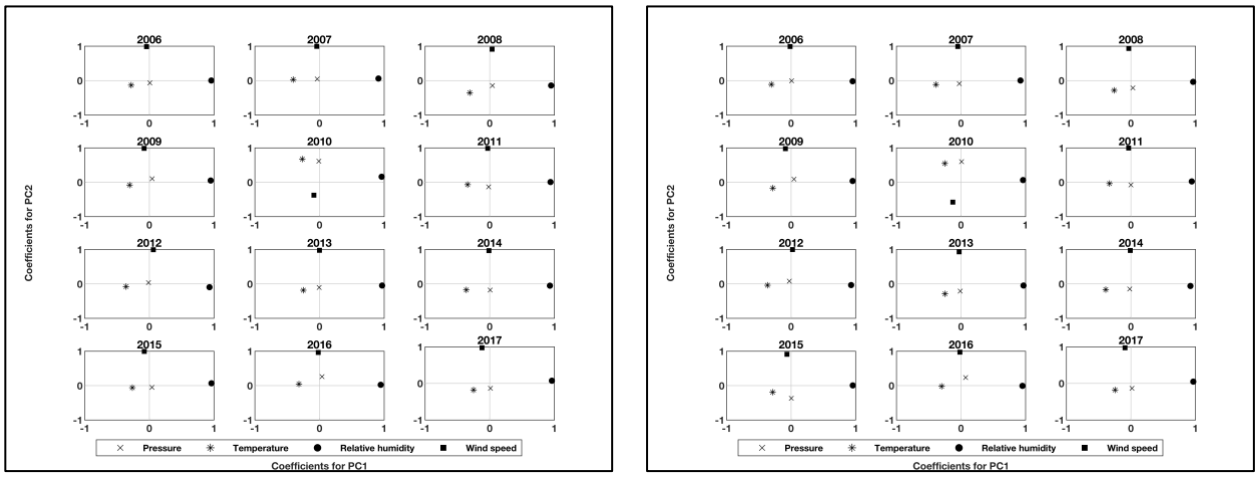


Figure A-5 Percentage of total variance explained by each of the first and second principle components at each station. The PCA analysis is done using data subsets 1 to 12 considering lead time range of 0-10 minutes.

(a) Zurich station (lead time: 10-20 minutes)

(b) Zurich station (lead time: 20-30 minutes)



(c) Säntis station (lead time: 10-20 minutes)

(d) Säntis station (lead time: 20-30 minutes)

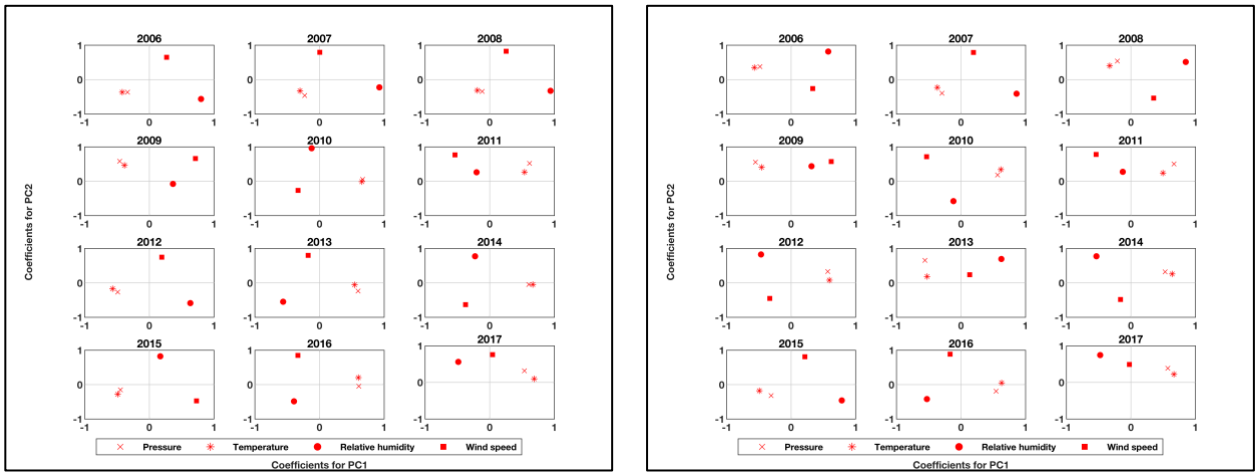


Figure A-6 The contribution of individual variables to the first and second principle components from 2006 to 2017 for **a** Zurich station (lead time: 10-20 minutes), **b** Zurich station (lead time: 20-30 minutes), **c** Säntis station (lead time: 10-20 minutes), **d** Säntis station (lead time: 20-30 minutes). In each subplot, the horizontal axis is the coefficients for PC1 and vertical axis is coefficients for PC2.

Appendix B: Supplementary Information for Chapter 4

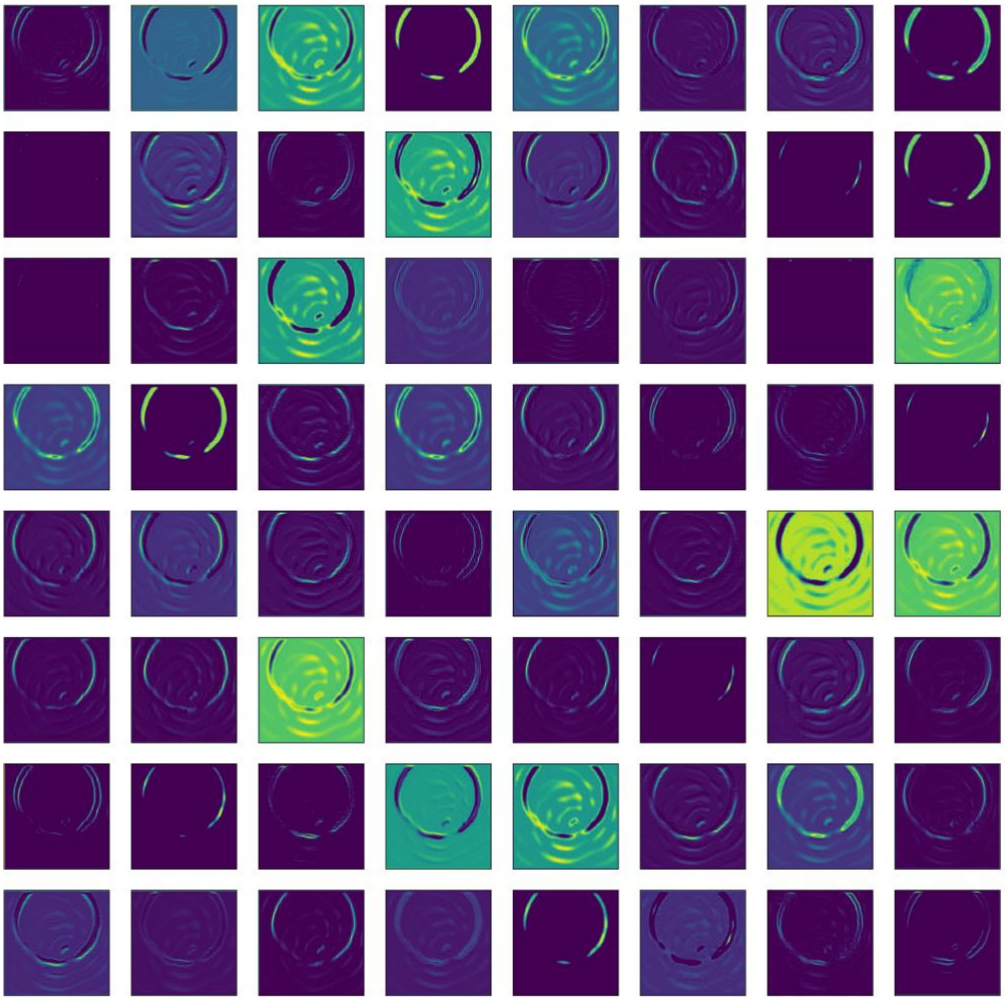


Figure B-1 Visualization of the first 64 feature maps of the 2nd layer in the VGG-19 model. The input image is the one presented in Figure 4-5.

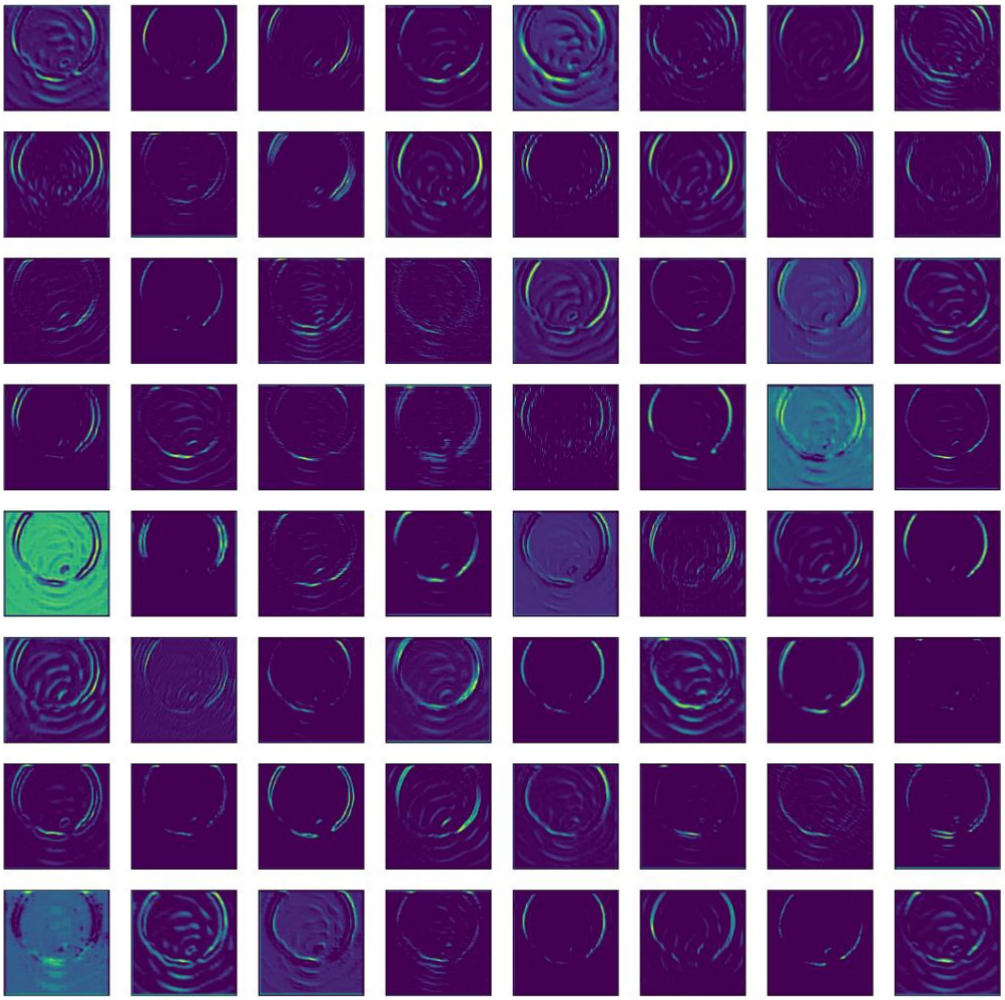


Figure B-2 Visualization of the first 64 feature maps of the 5th layer in the VGG-19 model. The input image is the one presented in Figure 4-5.

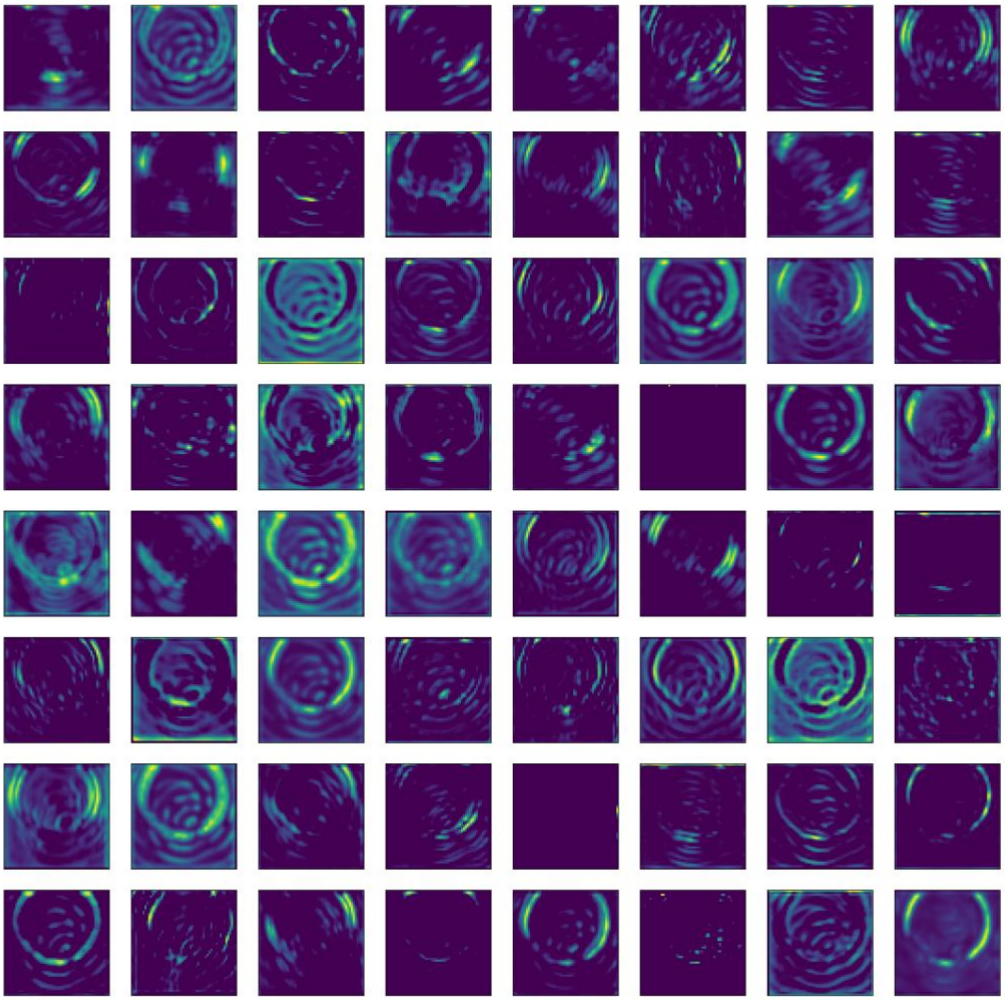


Figure B-3 Visualization of the first 64 feature maps of the 10th layer in the VGG-19 model. The input image is the one presented in Figure 4-5.

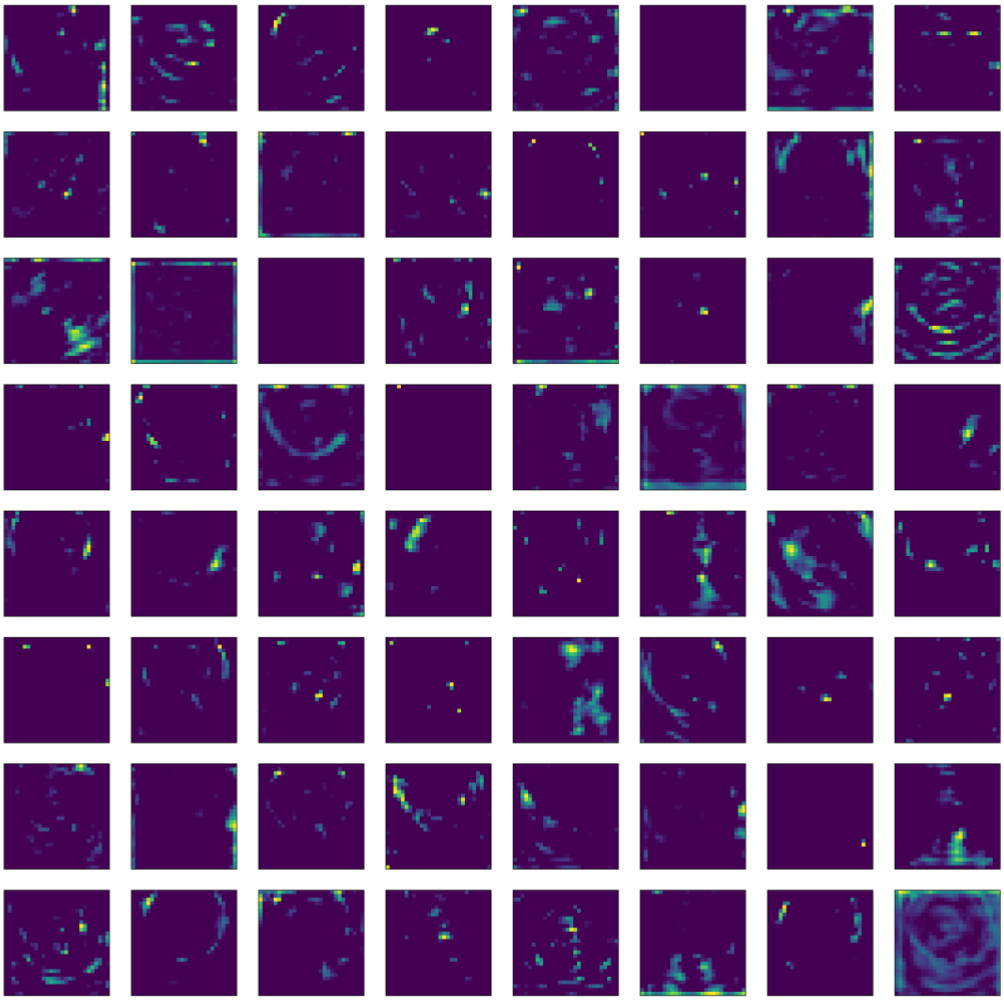


Figure B-4 Visualization of the first 64 feature maps of the 15th layer in the VGG-19 model. The input image is the one presented in Figure 4-5.

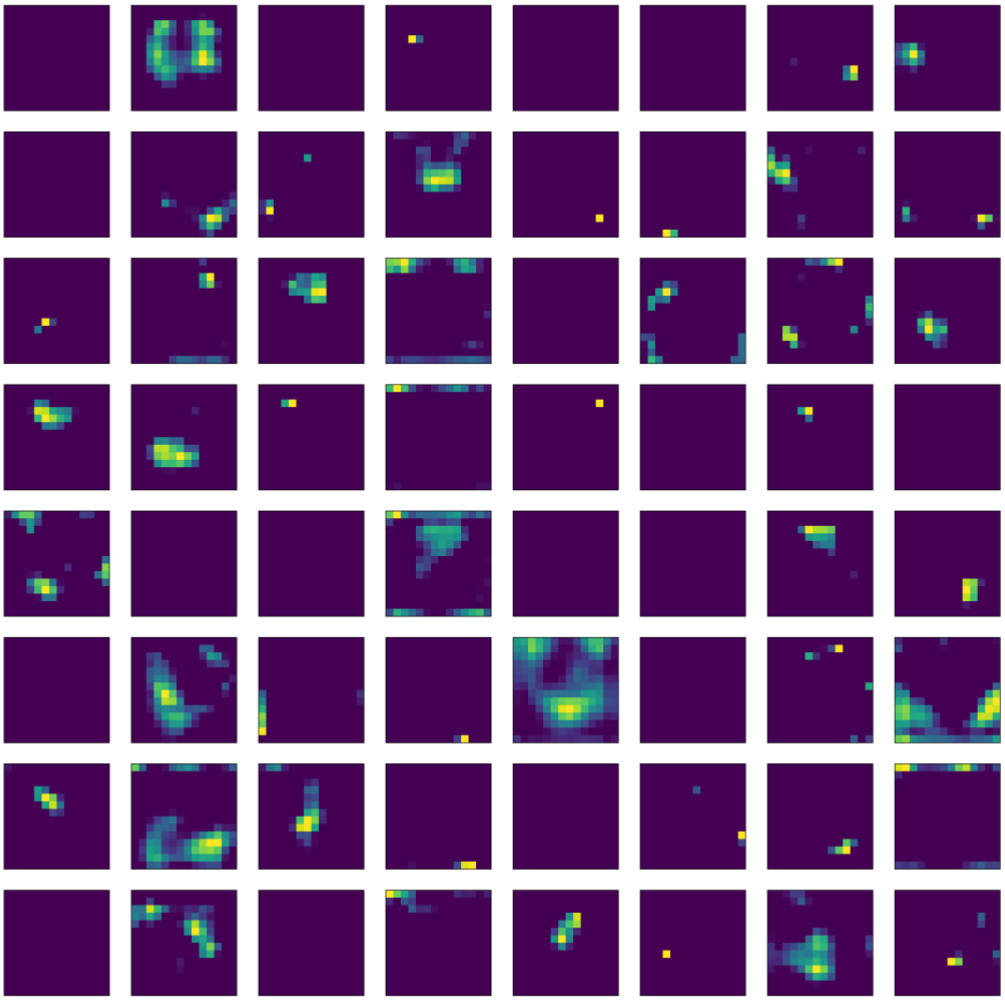


Figure B-5 Visualization of the first 64 feature maps of the 20th layer in the VGG-19 model. The input image is the one presented in Figure 4-5.

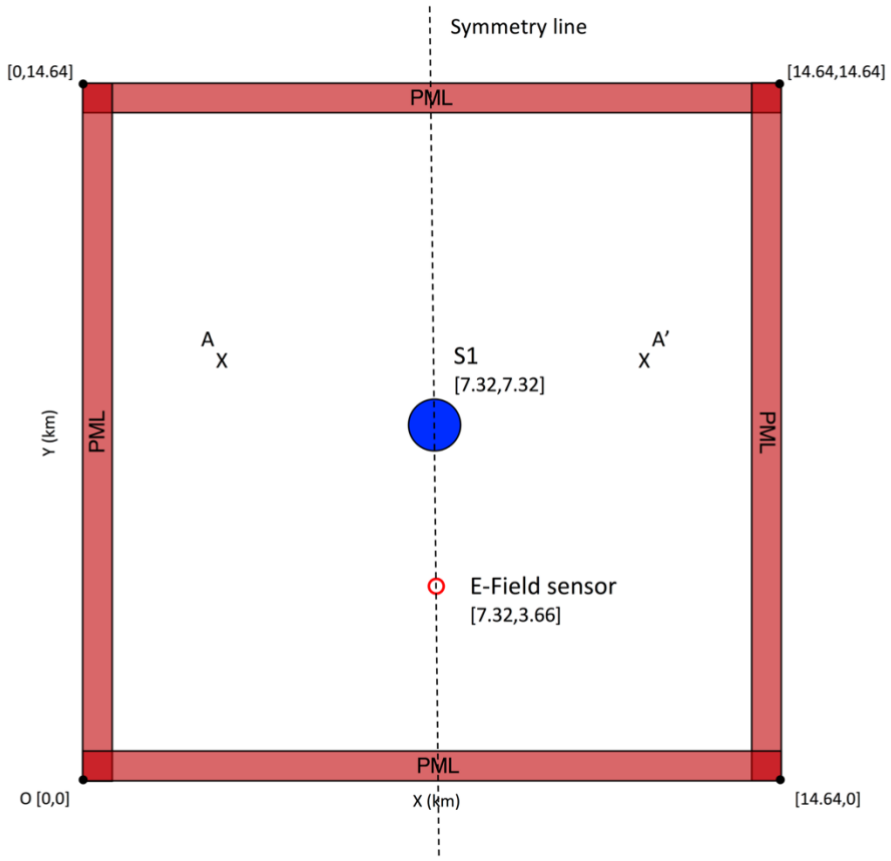


Figure B-6 Ambiguity of the solution in case of one symmetrical scatterer. A is an arbitrary point inside the medium and A' is the mirror of A with respect to the symmetry line.

Bibliography

[1] M. A. Cooper and R. L. Holle, “Current Global Estimates of Lightning Fatalities and Injuries,” Springer, Cham, 2019, pp. 65–73.

[2] R. S. Cerveny *et al.*, “WMO assessment of weather and climate mortality extremes: lightning, tropical cyclones, tornadoes, and hail,” *Weather. Clim. Soc.*, vol. 9, no. 3, pp. 487–497, 2017.

[3] R. L. Holle, “Annual rates of lightning fatalities by country,” in *20th International lightning detection conference*, 2008, vol. 2425.

[4] E. B. Curran, R. L. Holle, and R. E. López, “Lightning casualties and damages in the United States from 1959 to 1994,” *J. Clim.*, vol. 13, no. 19, pp. 3448–3464, 2000.

[5] W. S. Ashley and C. W. Gilson, “A reassessment of US lightning mortality,” *Bull. Am. Meteorol. Soc.*, vol. 90, no. 10, pp. 1501–1518, 2009.

[6] A. O. Fierro, E. R. Mansell, D. R. MacGorman, and C. L. Ziegler, “The Implementation of an Explicit Charging and Discharge Lightning Scheme within the WRF-ARW Model: Benchmark Simulations of a Continental Squall Line, a Tropical Cyclone, and a Winter Storm,” *Mon. Weather Rev.*, vol. 141, no. 7, pp. 2390–2415, 2013.

[7] A. Badoux, N. Andres, F. Techel, and C. Hegg, “Natural hazard fatalities in Switzerland from 1946 to 2015,” *Nat. Hazards Earth Syst. Sci.*, vol. 16, no. 12, pp. 2747–2768, Dec. 2016.

[8] D. Speranza III, “Lightning Prediction Using Recurrent Neural Networks,” AIR FORCE INSTITUTE OF TECHNOLOGY WRIGHT-PATTERSON AFB OH WRIGHT-PATTERSON …, 2019.

[9] A. I. Watson, R. E. López, R. L. Holle, and J. R. Daugherty, “The Relationship of Lightning to Surface Convergence at Kennedy Space Center: A Preliminary Study,” *Weather Forecast.*, vol. 2, no. 2, pp. 140–157, Jun. 1987.

[10] A. I. Watson, R. L. Holle, R. E. López, R. Ortiz, and J. R. Nicholson, “Surface Wind Convergence as a Short-Term Predictor of Cloud-to-Ground Lightning at Kennedy Space Center,” *Weather Forecast.*, vol. 6, no. 1, pp. 49–64, Mar. 1991.

[11] S. Uadiale, E. Urban, R. Carvel, D. Lange, and G. Rein, “Overview of problems and solutions in fire protection engineering of wind turbines,” *Fire Saf. Sci.*, vol. 11, pp. 983–995, 2014.

[12] S. Yokoyama, N. Honjo, Y. Yasuda, and K. Yamamoto, “Causes of wind turbine blade damages due to lightning and future research target to get better protection measures,” in *2014 International Conference on Lightning Protection (ICLP)*, 2014, pp. 823–830.

[13] H. Braam *et al.*, “Lightning damage of OWECS Part 3: Case studies,” 2002.

[14] D. R. MacGorman, J. M. Straka, and C. L. Ziegler, “A Lightning Parameterization for Numerical Cloud Models,” *J. Appl. Meteorol.*, vol. 40, no. 3, pp. 459–478, 2001.

[15] E. R. Mansell, D. R. MacGorman, C. L. Ziegler, and J. M. Straka, “Simulated three-dimensional branched lightning in a numerical thunderstorm model,” *J. Geophys. Res. Atmos.*, vol. 107, no. D9, p. ACL-2, 2002.

[16] E. R. Mansell, D. R. MacGorman, C. L. Ziegler, and J. M. Straka, “Charge structure and lightning sensitivity in a simulated multicell thunderstorm,” *J. Geophys. Res. D Atmos.*, vol. 110, no. 12, pp. 1–24, 2005.

[17] J. Helsdon, W. Wojcik, and R. Farley, “An examination of thunderstorm-charging mechanisms using a two-dimensional storm electrification model,” *J. Geophys. Res.*, vol. 106, no. D1, pp. 1165–1192, 2001.

[18] A. O. Fierro, E. R. Mansell, C. L. Ziegler, and D. R. Macgorman, “Explicit electrification and lightning forecast implemented within the WRF-ARW model,” in *XV International Conference on Atmospheric Electricity*, 2014, no. 7.

[19] P. R. Field, M. J. Roberts, and J. M. Wilkinson, “Simulated Lightning in a Convection Permitting Global Model,” *J. Geophys. Res. Atmos.*, vol. 123, no. 17, pp. 9370–9377, Sep. 2018.

[20] A. J. Dowdy, “Seasonal forecasting of lightning and thunderstorm activity in tropical and temperate regions of the world,” *Sci. Rep.*, vol. 6, no. 1, p. 20874, Aug. 2016.

[21] D. M. Romps, A. B. Charn, R. H. Holzworth, W. E. Lawrence, J. Molinari, and D. Vollaro, “CAPE Times P Explains Lightning Over Land But Not the Land-Ocean Contrast,” *Geophys. Res. Lett.*, vol. 45, no. 22, pp. 12,623-12,630, Nov. 2018.

[22] B. C. Bates, A. J. Dowdy, and R. E. Chandler, “Lightning Prediction for Australia Using Multivariate Analyses of Large-Scale Atmospheric Variables,” *J. Appl. Meteorol. Climatol.*, vol. 57, no. 3, pp. 525–534, Mar. 2018.

[23] P. Lopez, “A Lightning Parameterization for the ECMWF Integrated Forecasting System,” *Mon. Weather Rev.*, vol. 144, no. 9, pp. 3057–3075, Sep. 2016.

[24] C. Price and D. Rind, “A simple lightning parameterization for calculating global lightning distributions,” *J. Geophys. Res. Atmos.*, vol. 97, no. D9, pp. 9919–9933, Jun. 1992.

[25] B. H. Lynn *et al.*, “Predicting Cloud-to-Ground and Intracloud Lightning in Weather Forecast Models,” *Weather Forecast.*, vol. 27, no. 6, pp. 1470–1488, Dec. 2012.

[26] M. K. Tippett and W. J. Koshak, “A Baseline for the Predictability of U.S. Cloud-to-Ground Lightning,” *Geophys. Res. Lett.*, vol. 45, no. 19, pp. 10,719-10,728, Oct. 2018.

[27] Y. Geng *et al.*, “Lightnet: A dual spatiotemporal encoder network model for lightning prediction,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2439–2447.

[28] M. J. Murphy, N. W. S. Demetriades, and K. L. Cummins, “Probabilistic early warning of cloud-to-ground lightning at an airport,” in *16th Conference on Probability and Statistics in the Atmospheric Sciences,* 2000, pp. 126–131.

[29] A. Karagiannidis, K. Lagouvardos, and V. Kotroni, “The use of lightning data and Meteosat infrared imagery for the nowcasting of lightning activity,” *Atmos. Res.*, vol. 168, pp. 57–69, Feb. 2016.

[30] D. M. Smith *et al.*, “The rarity of terrestrial gamma-ray flashes,” *Geophys. Res. Lett.*, vol. 38, no. 8, p. n/a-n/a, Apr. 2011.

[31] V. A. Rakov, “Electromagnetic Methods of Lightning Detection,” *Surv. Geophys.*, vol. 34, no. 6, pp. 731–753, Nov. 2013.

[32] K. L. Cummins, M. J. Murphy, E. A. Bardo, W. L. Hiscox, R. B. Pyle, and A. E. Pifer, “A Combined TOA/MDF Technology Upgrade of the U.S. National Lightning Detection Network,” *J. Geophys. Res. Atmos.*, vol. 103, no. D8, pp. 9035–9044, Apr. 1998.

[33] K. L. Cummins and M. J. Murphy, “An Overview of Lightning Locating Systems: History, Techniques, and Data Uses, With an In-Depth Look at the U.S. NLDN,” *IEEE Trans. Electromagn. Compat.*, vol. 51, no. 3, pp. 499–518, Aug. 2009.

[34] N. Mora, F. Rachidi, and M. Rubinstein, “Application of the time reversal of electromagnetic fields to locate lightning discharges,” *Atmos. Res.*, vol. 117, pp. 78–85, Nov. 2012.

[35] G. Lugrin, N. M. Parra, F. Rachidi, M. Rubinstein, and G. Diendorfer, “On the Location of Lightning Discharges Using Time Reversal of Electromagnetic Fields,” *IEEE Trans. Electromagn. Compat.*, vol. 56, no. 1, pp. 149–158, Feb. 2014.

[36] N. Mora, F. Rachidi, and M. Rubinstein, “Locating lightning using time reversal of electromagnetic fields,” in *2010 30th International Conference on Lightning Protection (ICLP)*, 2010, pp. 1–6.

[37] G. Lugrin, N. Mora, F. Rachidi, M. Rubinstein, and G. Diendorfer, “On the use of the Time Reversal of Electromagnetic fields to locate lightning discharges,” in *2012 International Conference on Lightning Protection (ICLP)*, 2012, pp. 1–4.

[38] H. Braam *et al.*, “Lightning Damage of OWECS Part 3: ‘Case Studies,’” Netherlands, 2002.

[39] S. E. Reynolds, M. Brook, and M. F. Gourley, “THUNDERSTORM CHARGE SEPARATION,” *J. Meteorol.*, vol. 14, no. 5, pp. 426–436, Oct. 1957.

[40] C. P. R. Saunders, H. Bax-norman, C. Emersic, E. E. Avila, and N. E. Castellano, “Laboratory studies of the effect of cloud conditions on graupel/crystal charge transfer in thunderstorm electrification,” *Q. J. R. Meteorol. Soc.*, vol. 132, no. 621, pp.

2653–2673, Oct. 2006.

[41] L. D. Carey, K. M. Buffalo, L. D. Carey, and K. M. Buffalo, “Environmental Control of Cloud-to-Ground Lightning Polarity in Severe Storms,” *Mon. Weather Rev.*, vol. 135, no. 4, pp. 1327–1353, Apr. 2007.

[42] E. R. Mansell, D. R. MacGorman, C. L. Ziegler, and J. M. Straka, “Simulated three-dimensional branched lightning in a numerical thunderstorm model,” *J. Geophys. Res. Atmos.*, vol. 107, no. D9, 2002.

[43] A. O. Fierro *et al.*, “The Implementation of an Explicit Charging and Discharge Lightning Scheme within the WRF-ARW Model: Benchmark Simulations of a Continental Squall Line, a Tropical Cyclone, and a Winter Storm,” *Mon. Weather Rev.*, vol. 141, no. 7, pp. 2390–2415, Jul. 2013.

[44] D. M. Romps, A. B. Charn, R. H. Holzworth, W. E. Lawrence, J. Molinari, and D. Vollaro, “CAPE Times P Explains Lightning Over Land But Not the Land-Ocean Contrast,” *Geophys. Res. Lett.*, vol. 45, no. 22, pp. 12,623-12,630, Nov. 2018.

[45] P. Lopez, “A Lightning Parameterization for the ECMWF Integrated Forecasting System,” *Mon. Weather Rev.*, vol. 144, no. 9, pp. 3057–3075, 2016.

[46] J. Mecikalski, C. Jewett, L. Carey, B. Zavodsky, and G. Stano, “An Integrated 0-1 hour First-Flash Lightning Nowcasting, Lightning Amount and Lightning Jump Warning Capability.”

[47] J. P. Charba and F. G. Samplatsky, “Operational 2-h thunderstorm guidance forecasts to 24 hours on a 20-km grid,” in *Preprints, 23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction, Omaha, NE, Amer. Meteor. Soc. B*, 2009, vol. 17.

[48] T. Chen and T. He, “Higgs Boson Discovery with Boosted Trees,” in *HEPML’14 Proceedings of the 2014 International Conference on High-Energy Physics and Machine Learning*, 2015, vol. 42, pp. 69–80.

[49] G. N. Seroka, R. E. Orville, and C. Schumacher, “Radar Nowcasting of Total Lightning over the Kennedy Space Center,” *Weather Forecast.*, vol. 27, no. 1, pp. 189–204, Feb. 2012.

[50] S. Sumathi and S. N. Sivanandam, *Introduction to Data Mining and its Applications*. Springer, 2006.

[51] M. W. Libbrecht and W. Stafford Noble, “Machine learning applications in genetics and genomics,” *Nat. Publ. Gr.*, vol. 16, 2015.

[52] E. Alpaydin, *Introduction to Machine Learning*. MIT press, 2014.

[53] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, and F. Provost, “Machine learning for targeted display advertising: transfer learning in action,” *Mach. Learn.*, vol. 95, no. 1, pp. 103–127, Apr. 2014.

[54] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects.,” *Science*, vol. 349, no. 6245, pp. 255–60, Jul. 2015.

[55] C. Kerepesi, B. Daróczy, Á. Sturm, T. Vellai, and A. Benczúr, “Prediction and characterization of human ageing-related proteins by using machine learning,” *Sci. Rep.*, vol. 8, no. 1, p. 4094, Dec. 2018.

[56] A. Bracco, F. Falasca, A. Nenes, I. Fountalis, and C. Dovrolis, “Advancing climate science with knowledge-discovery through data mining,” *npj Clim. Atmos. Sci.*, vol. 1, no. 1, p. 20174, Dec. 2018.

[57] N. Jones, “How machine learning could help to improve climate forecasts,” *Nature*, vol. 548, no. 7668, pp. 379–380, Aug. 2017.

[58] A. McGovern *et al.*, “Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather,” *Bull. Am. Meteorol. Soc.*, vol. 98, no. 10, pp. 2073–2090, Oct. 2017.

[59] A. Manzato, “Hail in Northeast Italy: A Neural Network Ensemble Forecast Using Sounding-Derived Indices,” *Weather Forecast.*, vol. 28, no. 1, pp. 3–28, Feb. 2013.

[60] R. Lagerquist, A. McGovern, and T. Smith, “Machine Learning for Real-Time Prediction of Damaging Straight-Line Convective Wind,” *Weather Forecast.*, vol. 32, no. 6, pp. 2175–2193, Dec. 2017.

[61] G. R. Herman and R. S. Schumacher, “‘Dendrology’ in Numerical Weather Prediction: What Random Forests and Logistic Regression Tell Us about Forecasting Extreme Precipitation,” *Mon. Weather Rev.*, vol. 146, no. 6, pp. 1785–1812, Jun. 2018.

[62] C. D. Karstens *et al.*, “Development of a Human–Machine Mix for Forecasting Severe Convective Events,” *Weather Forecast.*, vol. 33, no. 3, pp. 715–737, Jun. 2018.

[63] S. E. Reynolds, M. Brook, M. F. Gourley, S. E. Reynolds, M. Brook, and M. F. Gourley, “THUNDERSTORM CHARGE SEPARATION,” *J. Meteorol.*, vol. 14, no. 5, pp. 426–436, Oct. 1957.

[64] J. Latham, W. A. Petersen, W. Deierling, and H. J. Christian, “Field identification of a unique globally dominant mechanism of thunderstorm electrification,” *Q. J. R. Meteorol. Soc.*, vol. 133, no. 627, pp. 1453–1457, Sep. 2007.

[65] V. Cooray, “Interaction of Lightning Flashes with the Earth’s Atmosphere,” in *An Introduction to Lightning*, Dordrecht: Springer Netherlands, 2015, pp. 341–358.

[66] I. L. Jirak, C. J. Melick, and S. J. Weiss, “Combining Probabilistic Ensemble Information from the Environment with Simulated Storm Attributes to Generate Calibrated Probabilities of Severe Weather Hazards,” in *Preprints, 27th Conf. Severe Local Storms*, 2014.

[67] “African Centres for Lightning and Electromagnetics Network : Home.” [Online]. Available: https://aclenet.org/. [Accessed: 05-Aug-2019].

[68] K. Berger, “Novel Observations on Lightning Discharges: Results of Research on Mount San Salvatore,” *J. Franklin Inst.*, vol. 283, no. 6, 1967.

[69] D. Li *et al.*, “On Lightning Electromagnetic Field Propagation Along an Irregular Terrain,” *IEEE Transactions on Electromagnetic Compatibility*, vol. 58, no. 1. pp. 161–171, 2016.

[70] A. Smorgonskiy, F. Rachidi, M. Rubinstein, G. Diendorfer, and W. Schulz, “On the proportion of upward flashes to lightning research towers,” *Atmos. Res.*, vol. 129–130, pp. 110–116, Jul. 2013.

[71] A. Mostajabi *et al.*, “LMA Observation of Upward Flashes at Säntis Tower : Preliminary Results,” in *Joint IEEE International Symposium on Electromagnetic Compatibility & Asia-Pacific Symposium on Electromagnetic Compatibility*, 2018, pp. 2–5.

[72] M. Antonio da Silva Ferro, J. Yamasaki, D. M. Roberto Pimentel, K. Pinheiro Naccarato, and M. Magalhães Fares Saba, “Lightning risk warnings based on atmospheric electric field measurements in Brazil,” *J. Aerosp.Technol. Manag. São José dos Campos*, vol. 3, no. 3, pp. 301–310.

[73] D. Aranguren, J. Montanya, G. Solá, V. March, D. Romero, and H. Torres, “On the lightning hazard warning using electrostatic field: Analysis of summer thunderstorms in Spain,” *J. Electrostat.*, vol. 67, no. 2–3, pp. 507–512, May 2009.

[74] A. Dewan, E. T. Ongee, M. Rafiuddin, M. M. Rahman, and R. Mahmood, “Lightning activity associated with precipitation and CAPE over Bangladesh,” *Int. J. Climatol.*, vol. 38, no. 4, pp. 1649–1660, Mar. 2018.

[75] “Mathworks. (2019). Least-Squares Fitting: User’s Guide (R2019a).” [Online]. Available: https://ch.mathworks.com/help/curvefit/least-squares-fitting.html. [Accessed: 17-May-2019].

[76] I. T. Jolliffe, *Principal Components in Regression Analysis*. Springer-Verlag New York, 1986.

[77] M. B. Richman, “Rotation of principal components,” *J. Climatol.*, vol. 6, no. 3, pp. 293–335, Jan. 1986.

[78] G. N. Seroka, R. E. Orville, and C. Schumacher, “Radar Nowcasting of Total Lightning over the Kennedy Space Center,” *Weather Forecast.*, vol. 27, no. 1, pp. 189–204, Feb. 2012.

[79] Q. Meng, W. Yao, and L. Xu, “Development of Lightning Nowcasting and Warning Technique and Its Application,” *Adv. Meteorol.*, vol. 2019, pp. 1–9, Jan. 2019.

[80] E. Brynjolfsson and T. Mitchell, “What can machine learning do? Workforce implications.,” *Science*, vol. 358, no. 6370, pp. 1530–1534, Dec. 2017.

[81] C. J. Schultz, W. A. Petersen, L. D. Carey, C. J. Schultz, W. A. Petersen, and L. D. Carey, “Preliminary Development and Evaluation of Lightning Jump Algorithms for the Real-Time Detection of Severe Weather,” *J. Appl. Meteorol. Climatol.*, vol. 48, no. 12, pp. 2543–2563, Dec. 2009.

[82] M. R. Smith and T. Martinez, “Improving classification accuracy by identifying and removing instances that should be misclassified,” in *The 2011 International Joint Conference on Neural Networks*, 2011, pp. 2690–2697.

[83] “Met Office WOW - Home Page.” [Online]. Available: http://wow.metoffice.gov.uk/. [Accessed: 20-Feb-2019].

[84] "Personal Weather Station Network | Weather Underground." [Online]. Available: https://www.wunderground.com/weatherstation/overview.asp. [Accessed: 20-Feb-2019].

[85] A. Smorgonskiy, F. Rachidi, M. Rubinstein, G. Diendorfer, and W. Schulz, "On the proportion of upward flashes to lightning research towers," *Atmos. Res.*, vol. 129–130, pp. 110–116, Jul. 2013.

[86] "Certification of monitoring stations - MeteoSwiss." [Online]. Available: https://www.meteoswiss.admin.ch/home/measurement-and-forecasting-systems/land-based-stations/automatisches-messnetz/certification-of-monitoring-stations.html. [Accessed: 22-Feb-2019].

[87] "Data preparation - MeteoSwiss." [Online]. Available: https://www.meteoswiss.admin.ch/home/measurement-and-forecasting-systems/datenmanagement/data-preparation.html. [Accessed: 22-Feb-2019].

[88] "The measurement values journey from the station to the customers - MeteoSwiss." [Online]. Available: https://www.meteoswiss.admin.ch/content/dam/meteoswiss/de/Mess-und-Prognosesysteme/Datenmanagement/doc/DWH_Weg_der_Daten_v1_0.pdf. [Accessed: 22-Feb-2019].

[89] "Lightning detection network - MeteoSwiss." [Online]. Available: https://www.meteoswiss.admin.ch/home/measurement-and-forecasting-systems/atmosphere/lightning-detection-network.html. [Accessed: 11-Jul-2018].

[90] M. Azadifar *et al.*, "Evaluation of the performance characteristics of the European Lightning Detection Network EUCLID in the Alps region for upward negative flashes using direct measurements at the instrumented Säntis Tower," *J. Geophys. Res. Atmos.*, 2016.

[91] W. Schulz, G. Diendorfer, S. Pedeboy, and D. R. Poelman, "The European lightning location system EUCLID – Part 1: Performance analysis and validation," *Nat. Hazards Earth Syst. Sci*, vol. 16, pp. 595–605, 2016.

[92] "Lightning detection network," 2016. [Online]. Available: http://www.meteoswiss.admin.ch/home/measurement-and-forecasting-systems/atmosphere/lightning-detection-network.html.

[93] "Our European detection network | Météorage." [Online]. Available: https://www.meteorage.com/who-are-we/our-european-detection-network. [Accessed: 04-Mar-2019].

[94] A. Mostajabi *et al.*, "LMA observation of upward flashes at Säntis Tower: Preliminary results," in *2018 IEEE International Symposium on Electromagnetic Compatibility and 2018 IEEE Asia-Pacific Symposium on Electromagnetic Compatibility (EMC/APEMC)*, 2018, pp. 399–402.

[95] "Measurement instruments - MeteoSwiss." [Online]. Available: https://www.meteoswiss.admin.ch/home/measurement-and-forecasting-systems/land-based-stations/automatisches-messnetz/measurement-instruments.html. [Accessed: 24-Feb-2019].

[96] "Copernicus Climate Change Service (C3S) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate . Copernicus Climate Change Service Climate Data Store (CDS)." [Online]. Available: https://confluence.ecmwf.int/display/CKB/ERA5+data+documentation#ERA5datadocumentation-HowtociteERA5. [Accessed: 21-Feb-2019].

[97] "ECMWF | Parameter details." [Online]. Available: https://apps.ecmwf.int/codes/grib/param-db/?id=59. [Accessed: 22-Jul-2019].

[98] C. Romero *et al.*, "A system for the measurements of lightning currents at the Säntis Tower," *Electr. Power Syst. Res.*, vol. 82, no. 1, pp. 34–43, 2012.

[99] C. Romero, F. Rachidi, M. Rubinstein, and M. Paolone, "Lightning currents measured on the Säntis Tower: A summary of the results obtained in 2010 and 2011," in *2013 IEEE International Symposium on Electromagnetic Compatibility*, 2013, pp. 825–828.

[100] M. Azadifar, M. Paolone, D. Pavanello, F. Rachidi, C. Romero, and M. Rubinstein, "An Update on the Instrumentation of the Säntis Tower in Switzerland for Lightning Current Measurements and Obtained Results," in *CIGRE Int. Colloquium on Lightning and Power Systems*, 2014.

[101] M. Azadifar, "Characteristics of Upward Lightning Flashes," Swiss Institute of Technology (EPFL), 2017.

[102] R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, and J. H. Moore, "Automating biomedical data science through tree-based pipeline optimization," in *European Conference on the Applications of Evolutionary Computation*, 2016, pp. 123–

137.

[103] R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, L. C. Kidd, and J. H. Moore, “Automating Biomedical Data Science Through Tree-Based Pipeline Optimization,” in *European Conference on the Applications of Evolutionary Computation*, 2016, pp. 123–137.

[104] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *KDD ’16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[105] R. Polikar, “Ensemble based systems in decision making,” *Circuits Syst. Mag. IEEE*, vol. 6, no. 3, pp. 21–45, 2006.

[106] L. Rokach, “Ensemble-based classifiers,” *Artif Intell Rev*, vol. 33, pp. 1–39, 2010.

[107] G. Casella, S. Fienberg, and I. Olkin, “An Introduction to statistical learning with Applications in R,” in *Springer Texts in Statistics*, Springer New York, 2013.

[108] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The Balanced Accuracy and Its Posterior Distribution,” in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 3121–3124.

[109] C. A. Doswell, R. Davies-Jones, and D. L. Keller, “On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables,” *Weather Forecast.*, vol. 5, no. 4, pp. 576–585, Dec. 1990.

[110] Z. Qin, M. Chen, F. Lyu, … S. C.-I. T., and U. 2019, “A GPU-Based Grid Traverse Algorithm for Accelerating Lightning Geolocation Process,” *IEEE Trans. Electromagn. Compat.*, 2019.

[111] Z. Abdul-Malek, Aulia, N. Bashir, and Novizo, “Lightning Location and Mapping System Using Time Difference of Arrival (TDoA) Technique,” in *Practical Applications and Solutions Using LabVIEW Software*, InTech, 2011.

[112] Z. Huang, J. Xu, Z. Gong, H. Wang, and Y. Yan, “Source localization using deep neural networks in a shallow water environment,” *J. Acoust. Soc. Am.*, vol. 143, no. 5, pp. 2922–2932, May 2018.

[113] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, “Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates,” *Sensors*, vol. 18, no. 10, p. 3418, Oct. 2018.

[114] S. Rusck, *Induced lightning over-voltages on power-transmission lines with special reference to the over-voltage protection of low-voltage networks.* Stockholm: KTH, 1958.

[115] F. Rachidi, “A review of field-to-transmission line coupling models with special emphasis to lightning-induced voltages on overhead lines,” *IEEE Trans. Electromagn. Compat.*, vol. 54, no. 4, pp. 898–911, 2012.

[116] F. Napolitano, A. Borghetti, C. A. Nucci, F. Rachidi, and M. Paolone, “Use of the full-wave finite element method for the numerical electromagnetic analysis of LEMP and its coupling to overhead lines,” *Electr. Power Syst. Res.*, vol. 94, pp. 24–29, 2013.

[117] “xgboost · PyPI.” [Online]. Available: https://pypi.org/project/xgboost/. [Accessed: 10-Dec-2019].

[118] “Lightning Parameters for Engineering Application,” no. 269. CIGRE WG C4.407, pp. 65–102, 2013.

[119] S. Sekioka, “An Equivalent Circuit for Analysis of Lightning-Induced Voltages on Multiconductor System Using an Analytical Expression,” in *International Conference on Power Systems Transients (IPST’05)*, 2005.

[120] H. Karami, A. Mostajabi, M. Azadifar, Z. Wang, M. Rubinstein, and F. Rachidi, “Locating Lightning Using Electromagnetic Time Reversal : Application of the Minimum Entropy Criterion,” in *Accepted in: International Symposium on Lightning Protection (XV SIPDA)*, 2019.

[121] H. Karami, F. Rachidi, and M. Rubinstein, “On the use of electromagnetic time reversal for lightning location,” in *2015 1st URSI Atlantic Radio Science Conference (URSI AT-RASC)*, 2015, p. 1.

[122] T. Wang, S. Qiu, L.-H. Shi, and Y. Li, “Broadband VHF Localization of Lightning Radiation Sources by EMTR,” *IEEE Trans. Electromagn. Compat.*, vol. 59, no. 6, pp. 1949–1957, Dec. 2017.

[123] F. Rachidi, M. Rubinstein, and M. Paolone, Eds., *Electromagnetic Time Reversal*. Chichester, UK: John Wiley & Sons, Ltd, 2017.

[124] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, 2015.

[125] P. Kosmas and C. M. Rappaport, “Time reversal with the FDTD method for microwave breast cancer detection,” *IEEE Trans. Microw. Theory Tech.*, vol. 53, no. 7, pp. 2317–2323, 2005.

[126] W. S. Cleveland, “Robust locally weighted regression and smoothing scatterplots,” *J. Am. Stat. Assoc.*, vol. 74, no. 368, pp. 829–836, 1979.

[127] P. W. Holland and R. E. Welsch, “Robust regression using iteratively reweighted least-squares,” *Commun. Stat. Methods*, vol. 6, no. 9, pp. 813–827, 1977.

[128] Y. LeCun *et al.*, “Handwritten Digit Recognition with a Back-Propagation Network,” in *Advances in Neural Information Processing Systems 2*, D. S. Touretzky, Ed. Morgan-Kaufmann, 1990, pp. 396–404.

[129] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[130] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[131] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. E. Barbano, “Toward automatic phenotyping of developing embryos from videos,” *IEEE Trans. Image Process.*, vol. 14, pp. 1360–1371, 2005.

[132] R. Vaillant, C. Monrocq, Y. L. C.-I. Proceedings-Vision, I. and Signal, and U. 1994, *IET computer vision*, vol. 141, no. 4. Institution of Engineering and Technology, 2007.

[133] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, “Efficient Object Localization Using Convolutional Networks,” in *Conference on Computer Vision and Pattern Recognition*, 2014.

[134] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, “Pedestrian Detection with Unsupervised Multi-Stage Feature Learning,” in *International Conference on Computer Vision and Pattern Recognition*, 2013.

[135] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[136] S. J. Nowlan and J. C. Platt, “A convolutional neural network hand tracker,” *Adv. Neural Inf. Process. Syst.*, pp. 901–908, 1995.

[137] M. Osadchy, Y. Le Cun, and M. L. Miller, “Synergistic Face Detection and Pose Estimation with Energy-Based Models,” *J. Mach. Learn. Res.*, vol. 8, no. May, pp. 1197–1215, 2007.

[138] C. Garcia and M. Delakis, “Convolutional face finder: a neural architecture for fast and robust face detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1408–1423, Nov. 2004.

[139] S. C. Turaga *et al.*, “Convolutional Networks Can Learn to Generate Affinity Graphs for Image Segmentation,” *Neural Comput.*, vol. 22, no. 2, pp. 511–538, Feb. 2010.

[140] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, “Multi-column deep neural network for traffic sign classification,” *Neural Networks*, vol. 32, pp. 333–338, Aug. 2012.

[141] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018.

[142] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[143] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[144] S. J. Pan and Q. Yang, “A Survey on Transfer Learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[145] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *CVPR09*, 2009.

[146] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[147] “Optimizers - Keras Documentation.” [Online]. Available: https://keras.io/optimizers/. [Accessed: 23-Sep-2019].

[148] V. A. Rafalsky, A. P. Nickolaenko, A. V. Shvets, and M. Hayakawa, “Location of lightning discharges from a single station,” *J. Geophys. Res.*, vol. 100, no. D10, p. 20829, Oct. 1995.

[149] C. Romero *et al.*, “A system for the measurements of lightning currents at the Säntis Tower,” *Electr. Power Syst. Res.*, vol. 82, no. 1, pp. 34–43, Jan. 2012.

[150] M. Azadifar *et al.*, “Fast initial continuous current pulses versus return stroke pulses in tower-initiated lightning,” *J. Geophys. Res. Atmos.*, vol. 121, no. 11, pp. 6425–6434, 2016.

[151] J. C. Willett, J. C. Bailey, C. Leteinturier, and E. P. Krider, “Lightning electromagnetic radiation field spectra in the interval from 0.2 to 20 MHz,” *J. Geophys. Res.*, vol. 95, no. D12, p. 20367, Nov. 1990.

[152] M. Azadifar *et al.*, “Evaluation of the performance characteristics of the European Lightning Detection Network EUCLID in the Alps region for upward negative flashes using direct measurements at the instrumented Säntis Tower,” *J. Geophys. Res. Atmos.*, pp. 595–606, 2015.

[153] H. Karami, F. Rachidi, and M. Rubinstein, “On practical implementation of electromagnetic models of lightning return-strokes,” *Atmosphere (Basel).*, vol. 7, no. 10, p. 135, 2016.

[154] “Wavelet Toolbox MATLAB” [Online]. Available: https://ch.mathworks.com/products/wavelet/features.html#denoising. [Accessed: 19-Sep-2019].

Amirhossein Mostajabi

/ Extended CV

+41 (0) 78 717 10 16

amirhossein.mostajabi@epfl.ch
amirhossein.mostajabi@roche.com

ahmostajabi

Route de Chavannes 17
CH-1007, Lausanne
Switzerland

Applied AI Researcher / Data Scientist

Highlights

- **Applied AI researche**r inspired by the opportunity to **cross-fertilize AI and Science**: (i) applying AI to solve tough challenges in diverse range of areas including **healthcare, physics**, **climate**, and **hardware design/test**; (ii) applying insights from **computational sensing** and **RF/acoustic imaging** to problems in AI with an emphasis on **computer vision**, **mixed reality**, **health monitoring**, and **HCI**
- **4+ years of experience** applying machine learning techniques (deep learning frameworks, CV, AR/MR, sequence modeling, generative modeling, AutoML) to solve **real-world problems** in collaboration with leading companies such as **Apple**, **Roche**, and **CSEM**
- Experience in **mining complex datasets** to derive **data-driven solutions** and draw **actionable insights**
- **Pattern analysis** of multiple data types, including **RWD sources** (clinical patient-level data, omics, imaging), **time series of sensors** data, **biomedical/RGB/multispectral/remote sensing satellite/simulation-generated imagery** data, and **structured** data types
- Experience in **2D/3D image restoration** (denoising, segmentation, resolution enhancement) of **medical images,** including **fluorescent/brightfield microscopy** images, **MRI/CT** data
- **Predictive ML modelling, signal processing,** and **advance analytics** for **anomaly detection**, **fault detection** and **intelligent maintenance** using heterogeneous **multivariate industrial-scale time series** data (wireless, electrical, EM field, ECG)
- Proficient at tackling **imbalanced classification** problems and leveraging **deep transfer learning** to solve multidisciplinary problems
- Hands-on experience from conception and design to test and verification of RF/EMC/EMI/HV products
- Highly passionate about delighting people by creating **innovative and robust AI-based products** with immediate real-world impact

Education

2017.05-present

(PhD) Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
Electrical Engineering | Communication and Computer Science
- Advisors: Prof. Farhad Rachidi (EE), Prof. Marcos Rubinstein (CS)

2007.09-2014.01

(BSc/MSc) University of Tehran (UT), Tehran, Iran
Electrical and Computer Engineering
- Advisors: Prof. H. Mohseni, Prof. Kaveh Niyayesh, Dr. Shayegani

Working Experience

2021.03-2021.08

Roche, Basel, Switzerland
Machine Learning Intern at Pharmaceutical Research and Early Development (pRED)
- RiSE PhD internship on **ML-powered high-throughput angiogenesis screening**
- **ML-powered** and automated analysis of **vascular network images** supplemented with **RNA-seq** data to assess the **drug's anti-/angiogenic potential**

2017.05-present

EMC Lab, EPFL, Lausanne, Switzerland
Doctoral Assistant, Applied AI Researcher
- Developing **intelligent machines** for solving real-world problems in areas such as **computational imaging** (source localization, tracking, and reconstruction for RF machine vision and EMC troubleshooting applications), **healthcare** (digital disease diagnostic, accelerated biomedical imaging, non-invasive deep brain tumor ablation, microscopy image restoration, 3D MRI/CT segmentation), and **weather prediction** (lightning nowcasting)
- Developing **new perception capabilities** for **mixed reality devices** and **autonomous systems** harnessing RF/microwave/acoustic sensing techniques
- Actively involved in H2020 EU project **Laser Lightning Rod** (LLR)

2020.08-2021.01 **Pro-Tech,** Boston, United States
Machine Learning Researcher (part-time contract)
- **Biomarker discovery** on ECG leads using **digital signal processing** and **deep learning** for early detection of heart arrhythmia

2020.03-2020.09 **Apple,** Cupertino, California, United States
Research Collaborator (remote co-op)
- Research on **AI-empowered super resolution imaging and sensing techniques** for **complex EMC troubleshooting** scenarios

2019.10-2020.04 **Swiss Center for Electronics and Microtechnology (CSEM),** Neuchatel, Switzerland
Computer Vision Research Intern
- Implementing Convolutional Neural Networks for pattern recognition and big data analysis of multispectral images coming from geostationary satellites.

2018.02-2019.02 **Montena Technologies SA,** Rossens, Switzerland
HV Engineer
- Responsible for HV measurements of new liquid dielectrics (DC, lightning, and transient regimes) which are the candidates for the next generation of HV insulation liquids in fast transient impulse generators (applications ranging from EMC tests to medical treatments)

2009.06-2017.03 **University of Tehran,** Tehran, Iran
Research Assistant
- Scientist/experimentalist with 8+ years of fundamental and hands-on industrial experience in EMC/EMI, power system and high voltage engineering

2016.03-2017.03 **Tehran Power Distribution Co.,** Tehran, Iran
Distribution Generation Professional
- Scientific consultant on protection of power networks due to the installation of new Distribution Generation (DG) plants

2015.10-2017.03 **Niroo Research Institute,** Tehran, Iran
Research Assistant
- *FEM analysis using Multi-physics simulations and experimental data from outdoor HV insulators*

2012.06-2013.10 **Nirou Trans Co.,** Shiraz, Iran
Research Scientist
- Designing new optical voltage measurement product suitable for HVDC (+400 KV) lines using on Pockels cells.

2010.09-2010.10 **Technical University of Dresden,** Dresden, Germany
DAAD Funded IAESTE Internship
- Project title: Electromagnetic design of 6-phases electrical machines (Prof. Wilfried Hofmann)

Core Experience

AI for Health Intelligence

ML-powered high-throughput angiogenesis screening [Roche, 2021-present]
- RiSE PhD internship on using **machine learning** for **biomedical image processing** and **automated analysis** of vascular networks in the group of investigative safety
- **Medical Computer Vision** research for morphological phenotyping of 2D/3D microscopy images
- AI-powered discovery of **imaging biomarkers** and **phenotypic features** to assess the drug's anti-/angiogenic potential and to provide insight into the angiogenic and vasculogenic processes
- Automated predictive pipeline from image acquisition to reporting
- Research translation in collaboration with biologists and pharma specialists

Deep brain tumor ablation using EMTR and Machine Learning [EPFL, 2020-2021]
- Using Electromagnetic Time Reversal (EMTR) to generate focused fields for the **ablation of deep-seated brain tumors**
- Reconstruction of **3D electromagnetic brain model** using **Microsoft InnerEye 3D segmentation Toolkit**
- Using Machine Learning to optimally design the EMTR antenna array, their excitation pulses, and the metalens parameters

Heart arrhythmia analysis from ECG signals using ML [Pro-Tech, 2020-2021]

- Devise robust **end-to-end ML algorithms** to detect different types of arrhythmia from ECG leads by replacing conventional signal processing techniques (Matrix Pencil and Wavelet scattering) with end-to-end ML-based solutions
- Using **1D Convolutional models** for denoising and extracting robust features form time-series ECG signals

Image restoration for light and electron microscopy using ML [course project, Dr. Florian Jug (MPI-CBG)]

- Devising deep learning models for multiple image restoration tasks (denoising, segmentation, resolution enhancement) using Content Aware Image Restoration (CARE) self-supervised deep learning models (e.g., noise2noise and noise2void)

AI for Machine perception . Sensing . Imaging

RF source localization using deep transfer learning for radio frequency machine vision [EPFL, 2019-present]

- Accurate localization of RF sources in free space using a **single** EM field sensor by hybridizing traditional **Electromagnetic Time Reversal (EMTR)** imaging technique with **machine learning**
- Applying pretrained **Convolutional Neural Networks** (VGG, Inception) to extract features from **synthetic (simulation-generated) images** and localize electromagnetic sources in a medium with comparable localization accuracy to those of conventional methods which need > 3 sensors to operate
- Bringing new perception capabilities and use cases to **AR/MR** devices and **Autonomous Systems** by adding a **new AI layer** that localizes and tracks sources/devices/agents in the environment
- Published in Nature Scientific Reports: www.nature.com/articles/s41598-019-53934-4

Enhanced autonomous perception via RF/acoustic imaging [EPFL, 2019-present]

- Developing **intelligent sensing, tracking, and imaging** tools by combining **wireless/acoustic sensing techniques** with advanced signal and image processing algorithms and advances in Computer Vision and ML - such as **image recognition**, **image-to-image translation**, multiple **sensor fusion**, etc.
- Leveraging **Domain Randomization** to deploy ML models trained only on synthetic data but with demonstrated high generalizability to real world conditions (good **Sim2Real translation**)
- Developing **physics-based neural networks** to enhance the physical realism of the predicted solution using appropriate auxiliary losses
- Enables **advance sensing** for variety of applications in Computer Vision, AR/MR/XR, and Autonomous Systems (3D perception and action/gesture recognition even through-walls and obstructions and in poor light conditions, enhanced SLAM), Human-Computer Interaction, EMC Troubleshooting (super resolution imaging in complex scenarios), and health monitoring

Partial Discharge (PD) localization using Time Reversal Cavity (TRC) [EPFL, 2019-2020]

- Locating one or more partial discharge sources inside power transformers or Gas Insulated Substations (GISs) using either, acoustic, UHF or combined sensors leveraging Time Reversal (TR) focusing technique (location accuracy better than $\lambda/10$)

Data-driven fault localization in power system transmission lines [EPFL, 2019]

- Accurate localization of faults in transmission lines by hybridizing EMTR and ML using the data from pre-installed voltage sensors
- **End-to-end** implementation including signal pre-processing, EMTR forward and backward stages, and neural regression layer

AI for Climate

ML-based lightning nowcasting scheme [EPFL and CSEM, 2018-present]

Using Ground-based Meteorological Measurements

- Using Gradient Boosting algorithms (**XGBoost**) to successfully nowcast the lightning hazard in the next 30 minutes from commonly-available meteorological parameters
- Published in Nature Climate and Atmospheric Science: www.nature.com/articles/s41612-019-0098-0
- Broad media coverage by e.g., Yahoo News, Daily Mail, Swiss National TV (RTS), ScienceDaily

Using Higher Atmosphere Measurements

- Using the **multispectral images** form the Advanced Baseline Imager (ABI) onboard of GOES-R satellite to hindcast lightning hazard over Americas with high spatiotemporal resolution
- Investigating both **data-driven** and **physics-based** approaches using variety of models including **CNN (U-Net)**, **ConvLSTM**, and **Pix2Pix GAN**

H2020 European Laser Lightning Rod project (Blog, Paper, Euronews) [EPFL, 2017-2021]

- Actively involved in developing a novel type of **lightning protection** based on the use of upward lightning discharges initiated through a **high repetition rate multi terawatt laser**

Additional Experience

Analyzing the initiation of upward lightning flashes [EPFL, 2017-2018]

- Studying the various processes involved in the initiation of upward lightning flashes from tall structures including the influence of the structure characteristics and terrain topography, the impact of meteorological conditions, and the effect of the cloud height and charge distribution

Energy optimization algorithms for data centers [Data Center Energy/Environmental Audit (Contract Part-time), 2016-2017]

- Data center energy efficiency improvement by modifying the electrical infrastructures design
- Feasibility study and economic analysis of integrating renewable and low-carbon energy solutions in data centers
- Comprehensive review of data center standards (ASHRAE, ANSI BICSI, etc.)

Design and construction of an electromagnetic field exposure system for studying biological effects of extremely low-frequency electromagnetic fields [UT, 2015]

In Collaboration with Department of Biology of Shiraz University

- Adjustable magnetic field intensity and time on/off to be suitable for different exposure conditions

Modeling Impulse Breakdown in Liquid Dielectrics (Master Thesis) [UT, 2014]

Design and construction of 3 High-Power Low-Rise Time Pulse Generators (Bachelor Thesis) [UT, 2011]

- Application: **Insulator aging** and **medical applications** (skin cancer therapy)
- Using fast transistors in avalanche mode to engage low rise times (controllable 10-60 ns)
- V-far about 15 kV with a K type antenna
- Two pulse forming lines: Blumlein and Bipolar coaxial PFL.
- 3 energy storage systems: compact Marx generator, especially designed Tesla Transformer, and Voltage Multiplier.

Hands-on experimental skills ………………………………………………………………………………………………………

RF/EMC/EMI/ESD

- Proficient with **RF test equipment** and calibration methods such as VNA, spectrum analyzer, oscilloscope, signal generators, ESD generators, anechoic chambers, and RF enclosures
- Actively involved in **5 experimental campaigns on source sensing and imaging** which brought me experience in working with **field probes** (B-dot and D-dot sensors), **Broadband VHF interferometer** systems, and **Lightning Mapping Arrays** (LMA)

High Voltage/Current Engineering

- Hands-on experience of testing **high voltage apparatus including** routine and type tests of CT, CVT, bushing, transformer, insulators, and capacitors, **Impulse test** of circuit breakers and transformers, **partial discharge measurement** of high voltage apparatuses, Insulation test of **transformer oil,** tracking and aging test of **polymeric insulators, ferro-resonance test** of CVTs, **high current test** of circuit breakers and bus ducts, **calibration** of high voltage measurement instruments

Other Hands-on experiences

- **Building several instruments** including LED aging cages and control, high voltage DC source, high current transformer (up to 3000 A), high voltage resonance reactor prototype for testing medium voltage cables, Insulation interface circuit using infra-red sensors, plasma loud speaker.
- Helping to **complete and start new laboratories in UT** e.g., the **first LED aging laboratory**, the **high current laboratory** (up to 100 kA), **power system scaled model laboratory**, and preliminary laboratory for electromagnetic and electrical machines.
- Actively involved in design (FEM analysis), construction, and test of low voltage bus ducts (joint project between UT and MAPNA Group).

Specialties

Computer Science (CS)

Applied Machine Learning | **Deep Learning** | **Computer Vision** (RF machine vision, pattern recognition, segmentation, object detection, tracking, sensor fusion) | **AR/MR** (semantic understanding, depth sensing, pose estimation) | **Sequence Modeling** (LSTM) | **DL for Signal Processing** (denoising, feature extraction) | **Causal Inference** | **Self-supervised Learning** | **Transfer Learning** | **Domain Randomization** | **Optimization** | **Big Data Analysis** | **Cloud Computing** (Azure, GCP, AWS) | **Interpretable ML** | **AutoML**

Electrical Engineering (EE)

Computational Electromagnetics | **Microwave Engineering** | **RF/Acoustic Sensing and Imaging** | **3D Multiphysics Simulations** | **RF/EMC/EMI** | **High Voltage and Power System Engineering**

Programming & Software skills ………………………………………………………………………………………………

Python (TF, PyTorch, Scikit-learn, OpenCV) | **MATLAB** (CV, ML, NN, Signal processing, Wavelet, Simulink toolboxes) | **Github** | **Computational electromagnetics** (FDTD, FEM simulations) | **COMSOL Multiphysics** (AC/DC, RF, Plasma modules) | **CST Studio**

Soft skills

Creative, curious, and self-motivated | Problem solving and analytical thinking skills | Organized with strong attention to detail | Experienced collaborating with x-functional teams | Excellent interpersonal skills, cross-group, and cross-culture collaboration | Enthusiastic to work in fast-paced environments | Ability to rapidly master new concepts and technology

Teaching and Mentoring Experience

Spring 2020	**Deep Learning (EE-559), EPFL,** Lausanne, Switzerland *Teacher assistance (Instructor: Prof. François Fleuret)*
Fall 2019	**Machine Learning (CS-433), EPFL,** Lausanne, Switzerland *Final project mentor (Instructor: Prof. Martin Jaggi)*
Fall 2019 Fall 2018	**Electromagnetic Compatibility (EE-576), EPFL,** Lausanne, Switzerland *Teacher assistance (Instructor: Prof. Farhad Rachidi)*

Previous teaching experience (UT, Sadra Institute of Higher Education)

I have cooperated as a lecturer or teaching assistant for several graduate and undergraduate courses/labs including High voltage engineering, Electrical Machinery I & II, and Electromagnetic Preliminary Laboratory.

Master's and bachelor's thesis

I have supervised 3 Master's and 3 Bachelor's theses since 2014.

Grants, Honors, and Awards

2020	**EPFL Enable Grant for Advancing Early-Stage Technology,** Lausanne, Switzerland • Project title: "Partial discharge localization in transformers and GIS using Electromagnetic Time Reversal" Funding: CHF 30K.
2020	**Excellent innovation recognition** by **EU Innovation Radar** Platform, European Commission *In collaboration with Ariane Space Group* • Innovation title: Machine Learning based nowcasting of lightening
2017	**DAAD Scholarship Award** for Ph.D. study at University of Stuttgart, Stuttgart, Germany (not taken)
2012	**Graduate Study Fellowship**, Exceptional Talent Office of University of Tehran, Tehran, Iran • Offered to the top 10% of the ECE undergraduates
2012	**National Elites Foundation** Scholarship, Tehran, Iran

Language Skills

English (Fluent) | French (A2) | German (A1) | Persian (Fluent)

Hobbies

Hiking | Music | Handicraft art | Making short films

Personal Details

Born in Shiraz, Iran on September 21st, 1989 | Swiss Residence Permit B since 2017

Amirhossein Mostajabi

/ Publications List

Applied AI Researcher / Data Scientist

+41 (0) 78 717 10 16
amirhossein.mostajabi@epfl.ch
ahmostajabi
Route de Chavannes 17
CH-1007, Lausanne
Switzerland

Publications [Google scholar]

Peer-reviewed Journal publications .....

Single-sensor source localization using electromagnetic time reversal and deep transfer learning: application to lightning, A Mostajabi, H Karami, M Azadifar, A Ghasemi, M Rubinstein, F Rachidi
Nature Scientific reports 9 (1), 1-14, 2019.

Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques, A Mostajabi, DL Finney, M Rubinstein, F Rachidi
Nature Partner Journal (npj) Climate and Atmospheric Science 2 (1), 1-15, 2019.

Machine Learning-Based Lightning Localization Algorithm Using Lightning-Induced Voltages on Transmission Lines, H Karami, A Mostajabi, M Azadifar, M Rubinstein, C Zhuang, F Rachidi
IEEE Transactions on Electromagnetic Compatibility, 2020.

The Laser Lightning Rod project, T Produit, P Walch, C Herkommer, A Mostajabi, M Moret, U Andral, et al.
European Physical Journal Applied Physics (EPJ AP), 2020.

Partial Discharge Localization Using Time Reversal: Application to Power Transformers, H Karami, M Azadifar, A Mostajabi, M Rubinstein, H Karami, et al.
Sensors, 2020.

Localization of electromagnetic interference sources using a time reversal cavity, H Karami, M Azadifar, A Mostajabi, P Favrat, M Rubinstein, F Rachidi
IEEE Transactions on Industrial Electronics, 2020.

Measurement and Modeling of both Distant and Close Electric Fields of an M-component in Rocket-triggered Lightning, Q Li, F Rachidi, M Rubinstein, J Wang, L Cai, M Azadifar, A Mostajabi, et al.
Journal of Geophysical Research: Atmospheres, 2020.

LMA observations of upward lightning flashes at the Säntis Tower initiated by nearby lightning activity, A Sunjerga, M Rubinstein, N Pineda, A Mostajabi, M Azadifar, D Romero, et al.
Electric Power Systems Research 181, 2020.

Meteorological aspects of self-initiated upward lightning at the Säntis tower (Switzerland), N Pineda, J Figueras i Ventura, D Romero, A Mostajabi, M Azadifar, et al.
Journal of Geophysical Research, 2019.

Numerical and Experimental Validation of Electromagnetic Time Reversal for Geolocation of Lightning Strikes, H Karami, M Azadifar, A Mostajabi, M Rubinstein, F Rachidi
IEEE Transactions on Electromagnetic Compatibility, 2019.

Analysis of the lightning production of convective cells, J Figueras i Ventura, N Pineda Rüegg, N Besic, J Grazioli, A Hering, et al.
Atmospheric Measurement Techniques 12 (10), 5573-5591, 2019.

Polarimetric radar characteristics of lightning initiation and propagating channels, J Figueras i Ventura, N Pineda, N Besic, J Grazioli, A Hering, et al.
Atmospheric Measurement Techniques 12, 2881-2911, 2019.

Analysis of a bipolar upward lightning flash based on simultaneous records of currents and 380-km distant electric fields, A Mostajabi, D Li, M Azadifar, F Rachidi, M Rubinstein, G Diendorfer, et al.
Electric Power Systems Research 174, 105845, 2019.

Analysis of a bipolar upward lightning flash based on simultaneous records of currents and 380-km distant electric fields, A Mostajabi, D Li, M Azadifar, F Rachidi, M Rubinstein, G Diendorfer, et al.
Electric Power Systems Research 174, 105845, 2019.

Reactor failure due to resonance in Zahedan-Iranshahr parallel EHV lines, analysis and practical solutions, MH Samimi, M Abedini, A Hossein, AS Akmal, H Mohseni

Acta Polytechnica Hungarica 11 (1), 63-78, 2014.

Performance evaluation of insulators using flashover voltage and leakage current, MH Samimi, AH Mostajabi, I Ahmadi-Joneidi, AA Shayegani-Akmal, et al.
Electric Power Components and Systems 41 (2), 221-233, 2013.

Effect of humidity on the flashover voltage of insulators at varying humidity and temperature conditions, MH Samimi, AH Mostajabi, M Arabzadeh, AAS Akmal, H Mohseni
J. Basic. Appl. Sci. Res 2 (4), 4299-4303, 2012.

Peer-reviewed conference publications

A data-driven approach for lightning nowcasting with deep learning, A Mostajabi, E Mansouri, P Pad, M Rubinstein, L A Dunbar, F Rachidi
European Geoscience Union (EGU), 2021.

X-rays observations at the Säntis Tower: Preliminary results, A Sunjerga, P Hettiarachchi, D Smith, M Rubinstein, V Cooray, M Azadifar, A Mostajabi, F Rachidi
European Geoscience Union (EGU), 2021.

Localization of electromagnetic interference source using a time reversal cavity: Application of the maximum power criterion, H Karami, M Azadifar, A Mostajabi, M Rubinstein, F Rachidi
IEEE International Symposium on Electromagnetic Compatibility & Signal/Power Integrity (EMCSI), 2020.

Initial Results from a 2nd generation Broadband VHF Interferometer Imaging System, M A Stanley, P R Krehbiel, X Fan, H E Edens, D Rodeheffer, W Rison, M Azadifar, A Sunjerga, A Mostajabi, F Rachidi and M Rubinstein
American Geosciences Union (AGU), 2020.

Nowcasting Lightning Occurrence Using Machine Learning Techniques: The Challenge of Identifying Outliers (*invited paper),* A Mostajabi, DL Finney, M Rubinstein, F Rachidi
European Geosciences Union (EGU), 2020.

Locating Lightning Using Electromagnetic Time Reversal: Application of the Minimum Entropy Criterion, H Karami, A Mostajabi, M Azadifar, Z Wang, M Rubinstein, F Rachidi
2019 International Symposium on Lightning Protection (XV SIPDA), 1-4, 2019.

Meteorological conditions during self-initiated upward lightning at the Säntis tower (Switzerland), J Montanya, N Pineda, JF i Ventura, D Romero, A Mostajabi, M Azadifar, et al.
American Geosciences Union (AGU) Fall Meeting, 2019.

Measurement and Analysis of the Breakdown Strength of Different Liquid Dielectric Materials, N Mora, A Mostajabi, B Daout, F Rachidi
Asian Electromagnetic International Conference ASIAEM, 2019.

On the Impact of Meteorological Conditions on the Initiation of Upward Lightning Flashes from Tall Structures, A Mostajabi, A Sunjerga, M Azadifar, A Smorgonskiy, M Rubinstein, et al.
2018 34th International Conference on Lightning Protection (ICLP), 1-5, 2018.

On the Classification of Self-Triggered versus Other-Triggered Lightning Flashes, A Sunjerga, A Mostajabi, F Rachidi, N Pineda, D Romero, et al.
2018 34th International Conference on Lightning Protection (ICLP), 1-5, 2018.

LMA observation of upward flashes at Säntis Tower: Preliminary results, A Mostajabi, N Pineda, D Romero, M Azadifar, O Van der Velde, et al.
IEEE International Symposium on Electromagnetic Compatibility and IEEE Asia-Pacific Symposium on Electromagnetic Compatibility (EMC/APEMC), 2018.

LMA Campaign for the Observation of Upward Lightning at the Säntis Tower During Summer 2017: Preliminary Results, A Mostajabi, N Pineda, A Sunjerga, D Romero, M Azadifar, et al.
XVI International Conference on Atmospheric Electricity, 2018.

An Analysis of Lightning Activity in the Säntis Region through the Big Hiatus in Global Warming, A Mostajabi, A Smorgonskiy, F Rachidi, M Azadifar, M Rubinstein, et al.
2018 International Lightning Detection Conference (ILDC), 2018.

Simultaneous records of current and 380-km distant electric field of a bipolar lightning flash, A Mostajabi, M Azadifar, F Rachidi, M Rubinstein, G Diendorfer, W Schulz, et al.
2017 International Symposium on Lightning Protection (XIV SIPDA), 183-187, 2017.

Patents

Partial Discharge Localization Using Time Reversal: Aplication to Power Transformers and Gas-Insulated Substations,
H Karami, A Mostajabi, F Rachidi, M Azadifar, M Rubinstein
International Patent PCT/IB2020/061271, November 2020.