LUIS GUILHERME PAVAM JUNIOR

COMPUTATIONAL INTELLIGENCE FOR LIGHTNING PREDICTION: A MACHINE LEARNING MODEL FOR CLOUD-TO-GROUND LIGHTNING NOWCASTING

(*pre-defense version, compiled at August 6, 2024*)

Dissertation presented as a partial requirement for the degree of Master of Sciences in Informatics in the Graduate Program in Informatics, Exact Sciences Sector, of the Federal University of Paraná, Brazil..

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dr. Marco Antonio Zanata Alves.

Coorientador: Prof. Dr. Luiz Eduardo Soares de Oliveira.

CURITIBA PR

2023

ABSTRACT

Lightning forecasting is a critical component of weather prediction, directly impacting public safety and infrastructure protection. This dissertation presents the development and evaluation of a Machine Learning (ML) model designed to predict Cloud-to-Ground Lightning (CG) lightning events with high spatial and temporal resolution. The model focuses on forecasting both the location and number of lightning occurrences for the Brazilian south-southeastern region for the next hour in five min. increments.

CG strikes are the focus of our prediction for these are the type of atmospheric electric discharges that occurs between the atmosphere and the planetary surface, and thus most capable of affecting our civilization. The selected Region of Interest (ROI) is due to the availability of lightning data in the region and the significant economic and human activities in the area. For this task, our proposal is to employ a ML method to build a CG lightning nowcasting system. Nowcasting is a mode of forecast in which predictions are made in the very-short term, in the range of a few hours.

The examination of the spatiotemporal patterns of lightning occurrences provided valuable insights into the behavior of this electric event and the definition of its seasonality. We verified that lightning is most present in the coast during the summer and in the central-western region of our ROI in the spring. This allowed for the development of curated datasets for training of the ML algorithm. The ML method relies on collections of data that describe the problem considered. Hence, we constructed a database containing the lightning occurrences declared by a local Lightning Detection and Location Network (LDLN) from 2018 to 2023.

Three ML algorithms were tested, with those being: Linear Regression (LR), Gradient Boosting Regressor (GB), and Random Forest (RF). The GB algorithm presented the foremost performance. The model predictions define the number of CG lightning occurrences for the next hour in five min. increments. Hence, not only the presence of lightning is predicted but also the severity of the event. Furthermore, the prediction of the quantity of CG strikes in five min. intervals allows for the monitoring of the thunderstorm evolution.

This work contributes to the advancement of lightning forecasting by providing a reliable and actionable predictive tool, offering a improvement over traditional forecasting methods. The findings underscore the model's potential to enhance public safety and disaster preparedness, paving the way for more accurate and timely lightning predictions in various operational contexts.

Keywords: 1. Lightning 2. Nowcasting 3. Machine Learning 4. Brazil

RESUMO

A previsão de relâmpagos é uma componente crítica da previsão meteorológica, com impacto direto para a defesa civil e na proteção de infraestruturas. Dessa forma, esta dissertação apresenta o desenvolvimento e avaliação de um modelo de Aprendizagem de Máquina (AML) projetado para prever eventos de relâmpagos Nuvem-Solo (NS) com alta resolução espacial e temporal. O modelo foca na previsão da localização e do número de ocorrências de relâmpagos para a região sul-sudeste do Brasil para a próxima hora em intervalos de cinco minutos.

Os relâmpagos NS são o foco da nossa previsão, pois estes definem o tipo de descargas eléctricas atmosféricas que ocorrem entre a atmosfera e a superfície do planeta e, portanto, mais capazes de afetar a nossa civilização. A região de interesse selecionada deve-se à disponibilidade de dados de relâmpagos na região e as significativas atividades econômicas e humanas na área. Para esta tarefa, nossa proposta é empregar um método de AML para construir um sistema de nowcasting de relâmpagos NS. Nowcasting é uma modalidade de previsão em que as previsões são feitas a curtíssimo prazo, na faixa de algumas horas.

O exame dos padrões espaço-temporais de ocorrência de relâmpagos forneceu valiosas informações sobre o comportamento desse evento elétrico e a definição de sua sazonalidade. Verificamos que os relâmpagos estão mais presentes no litoral durante o verão e na região centro-oeste da nossa região de interesse na primavera. Isto permitiu o desenvolvimento de conjuntos de dados curados para o treino do algoritmo de AML. O método de AML baseia-se em colecções de dados que descrevem o problema considerado. Assim, construímos uma base de dados contendo as ocorrências de relâmpagos declaradas por uma Rede de Detecção e Localização de Relâmpagos (RDLR) local de 2018 a 2023.

Foram testados três algoritmos de AML, sendo eles: Linear Regression (LR), Gradient Boosting Regressor (GB) e Random Forest (RF). O algoritmo GB apresentou o melhor desempenho. As previsões do modelo definem o número de ocorrências de relâmpagos NS para a próxima hora em incrementos de cinco minutos. Assim, não apenas a presença de relâmpagos é prevista, mas também a severidade do evento. Para além disso, a previsão da quantidade de NS ocorridos em intervalos de cinco minutos permite o monitoramento da evolução de uma tempestade.

Este trabalho contribui para o avanço da previsão de relâmpagos ao fornecer uma ferramenta de previsão fiável e acionável, oferecendo uma melhoria em relação aos métodos de previsão tradicionais. Os resultados denotam o potencial do modelo para melhorar a proteção de vidas e infraestrutura, estabelecendo um método para previsões de relâmpagos mais confiáveis em contextos operacionais.

Palavras-chave: 1. Relâmpago 2. Nowcasting 3. Aprendizagem de Máquina 4. Brasil

LIST OF FIGURES

LIST OF TABLES

Glossary

AI Artificial Intelligence. iv, 17, 20, 21, 25–30, 33

AML Aprendizagem de Máquina. iii

AMS American Meteorological Society. 33

BD Big Data. 16, 27

CEMIG Minas Gerais Energy Company. 25, 39, 40

CG Cloud-to-Ground Lightning. ii, iv–vii, 12–19, 25, 30, 33, 35, 37–40, 43–55, 57, 59–68, 77, 78, 85–96

CNN Convolutional Neural Network. 35

CSI Critical Success Index. iv, 33, 53, 55, 61, 71

CSV Comma Separated Values. 40

DF Federal District. 39, 40

ES Espirito Santo. 39, 40

FN False Negative. 53, 61

FP False Positive. 52, 53, 61

FSS Fraction Skill Score. iv, 53–55, 62, 63, 67, 68, 78

GB Gradient Boosting Regressor. ii–iv, 52, 55, 56, 58–60, 62, 68, 77, 78

GO Goias. 39, 40

IC Intra-Cloud Lightning. 12, 13, 19, 25, 35, 37, 39, 40, 43, 49

IMRAD Introduction, Methods, Results, and Discussion. 20

INPE National Institute of Space Research. 25, 39, 40

kA Kiloamper. 12, 40, 41

LDLN Lightning Detection and Location Network. ii, 21, 25, 35, 38, 39, 43, 54, 65

LR Linear Regression. ii, iii, 52, 55, 78

MAE Median Absolute Error. vii, 51, 52, 55, 57

MCS Mesoscale Convective System. 45, 46, 58, 77

MG Minas Gerais. 39, 40

MIV Model Interpretation and Visualization. 29

ML Machine Learning. ii, iv, vii, 16–21, 26–33, 35–38, 44, 45, 47, 49–55, 57, 59, 62, 66–69, 77–79

MS Mato Grosso do Sul. 39, 40

MSE Mean Squared Error. 51, 52, 55

NA Neighbourhood Array. iv, vii, 49–51, 57–59, 67

NS Nuvem-Solo. iii

NWP Numerical Weather Prediction. 32–34, 67

PR Parana. 39, 40

R2 R2 Score. 51, 52, 55

R2O Research to Operations. 37

RDLR Rede de Detecção e Localização de Relâmpagos. iii

RF Random Forest. ii, iii, 35, 52, 55, 78

RINDAT National Integrated Network for Detection of Atmospheric Discharges. iv, vii, 25, 39–44, 54, 77

RJ Rio de Janeiro. 39, 40

RMSE Root Mean Squared Error. 51, 52, 55

ROI Region of Interest. ii, iv–vi, 13–16, 18, 19, 25, 32, 35, 38, 40, 41, 43–48, 50, 54, 60–62, 64–68, 77, 78, 85–96

SIMEPAR Parana Technology and Environmental Monitoring Service. 25, 39, 40

SOM Self Organizing Maps. 18, 35, 67, 68

SP Sao Paulo. 39, 40

TL Total Lightning. iv, 13, 16, 19, 25, 33–37, 39–43

TN True Negative. 52, 53, 61

TP True Positive. 52, 53, 61

WMO World Meteorological Organization. 21

CONTENTS

1 INTRODUCTION

At this moment, hundreds of thunderstorms are occurring on Earth. Typically, at any given time, around 10% of the planetary surface is under sway of a storm (METED, 2014). Thunderstorms act an important part for the Earth's system, distributing high volumes of water and electricity in the atmosphere (Wallace and Hobbs, 2006). A storm may happen without detrimental effects on the environment and human society but can also be sources of great disturbances.

These tempests give rise to tens of lightning events every second (Rakov, 2016). Lightning is one of the main hazards posed by storms. It can damage infrastructure, disrupting power and water distribution; initiate wildfires that may span thousands of kilometers and last for days; and ultimately be lethal for humans and animals. Lightning occurrences leads to thousands of human and animal deaths and billions of dollars in damage annually worldwide (Holle, 2014). It is necessary to acknowledge the dangers that lightning strikes pose and take action to prevent or mitigate its harmful consequences.

A electrical atmospheric discharge, i.e., lightning, is a sudden transfer of electric charges between charged zones in the atmosphere. It occurs all around the planet and throughout the year. This transient phenomenon is characterized by its intensity and duration. An ordinary lightning has a peak electric current in the range of tens of Kiloampers (kAs) and a lifespan in nanoseconds' timescale. Defined both at the mesoscale and microscale, lightning events establishes complex interactions with the surrounding atmosphere (Zhou et al., 2020).

Lightning is a cardinal component of Earth's atmosphere in several aspects (Lopez, 2016). It is an indicator and influencer of local and global climates (Rakov, 2016). This natural phenomenon is a relevant local source of nitrogen oxides and ozone, which alters the atmospheric chemical makeup. By affecting the ozone layer in the troposphere and the stratosphere, lightning influences Earth's climate (Lopez, 2016).

Mitigating the fallout from high-impact weather events such as storms can be achieved by predicting its occurrence and evolution. This enables agents to put in place protective measures and safeguard assets and personnel. Forecasting weather conditions on scales of years, months, days, or hours is one of the meteorological services' most significant and challenging tasks. Lightning forecast is essential for the population's safety and continuous operation of various industries.

Lightning is a form of transport of electric charges. This transfer of charges can happen between numerous types of charged zones in the atmosphere, e.g., clouds or plumes from a volcanic eruption. Lightning nomenclature comes from the location of the charged zones involved in the transference of electric charge.

There are mainly two kinds of lightning: **Intra-Cloud Lightning (IC)**, which takes place solely in the atmosphere; and **Cloud-to-Ground Lightning (CG)**, which happens between

a cloud and an object in the planetary surface (Rakov, 2016). When both IC and CG events are considered, the phenomenon is termed as **Total Lightning (TL)**. Figure 1.1 presents a common case of a CG lightning strike, with its multiple channels traveling through the path of least resistance in search of an attachment point, which culminated to be a broadcast tower.

Figure 1.1: Photograph of a CG lightning strike to a TV broadcast tower (V. A. Rakov, 2003).

1.1 THE PROBLEM OF LIGHTNING FORECASTING

The problem of this research is defined as the improvement in the quality of CG lightning forecasting for an Region of Interest (ROI). Lightning, as a weather variable of significant environmental and societal impact, must be studied and, where there is human activity, tracked. However, lightning prediction poses a challenge. Predicting future weather conditions requires modeling an intricate and chaotic system, the atmosphere (Rasp, 2018).

Lightning arises from a chaotic process of cloud electrification – a process which we still lack a holistic understanding. Factors such as the water particle collisions, wind shear, and even the air composition (including air pollution) influence it, making it an intricate system to model accurately. In addition, data about lightning occurrences are obtained from remote sources, complicating the capture of the fine-grained details needed to pinpoint and describe lightning events.

An event of small-scale by nature, lightning is characterized as a localized event, specially when compared to the storm. Anticipating the location of a lightning strike is a arduous task due to the complex cloud dynamics. Thunderstorms, responsible for large volumes of lightning, are fast evolving events, developing and dissipating quickly, possible even within an single hour (Hayward et al., 2020). This rapid evolution makes long-term forecasting difficult.

The forecast of CG lightning is the focus of any meteorological service concerning lightning prediction, for CG strikes poses a severe risk to many activities. An accurate forecast with sufficient lead time can mitigate and even prevent the most damaging effects of lightning. Given that CG lightning is a highly dynamic and very short-lived event, its forecast is done in the very short time range, i.e., the next few hours. In meteorological jargon, forecasting up to six hours in the future is denominated nowcasting (described in further detail in Section 2.1).

Forecasting lightning is crucial due to its significant impact on safety, infrastructure, and various economic activities. Lightning poses a direct threat to human life, with thousands of injuries and fatalities reported globally each year. Lightning deaths typically makes the news, for they are preventable and can possibly be traced to a lack of warning or disregard for said warning.

To illustrate, in an event that stood out, 11 people were killed when a lightning stroke an indigenous ceremony in northern Colombia (CNN, 2014a). Given that it was an indigenous community, an warning was never made. In this single piece of news, there are references to two other events in a span of one month that resulted in lightning deaths (CNN, 2014b), (CNN, 2014c). In those cases, an warning was sent but not heard.

A lightning death may even be characterized as a criminal offense. In an event occurred in early 2019, a cowboy was killed by lightning when working in a open field. Due to the fact the cowboy was made to work under dangerous conditions – the occurrence of lightning was announced – his employer was convicted and made to pay damages to the cowboy's family (Consultor Jurídico, 2023). A lightning forecast that was not heeded was later used in a justice process.

Brazil is an economic and population powerhouse of South America, and has one of the highest rates of lightning in the world (Peterson, 2021). These frequent lightning strikes pose significant dangers. We defined the ROI of this study in the south-southeastern region of Brazil – which houses both vast urban areas and agricultural zones. Figure 1.2 presents the South American continent with an inset featuring our ROI.

This region was selected for: the availability of lightning data; it houses approximately 30% of the country's entire population (Instituto Brasileiro de Geografia e Estatística, 2023);

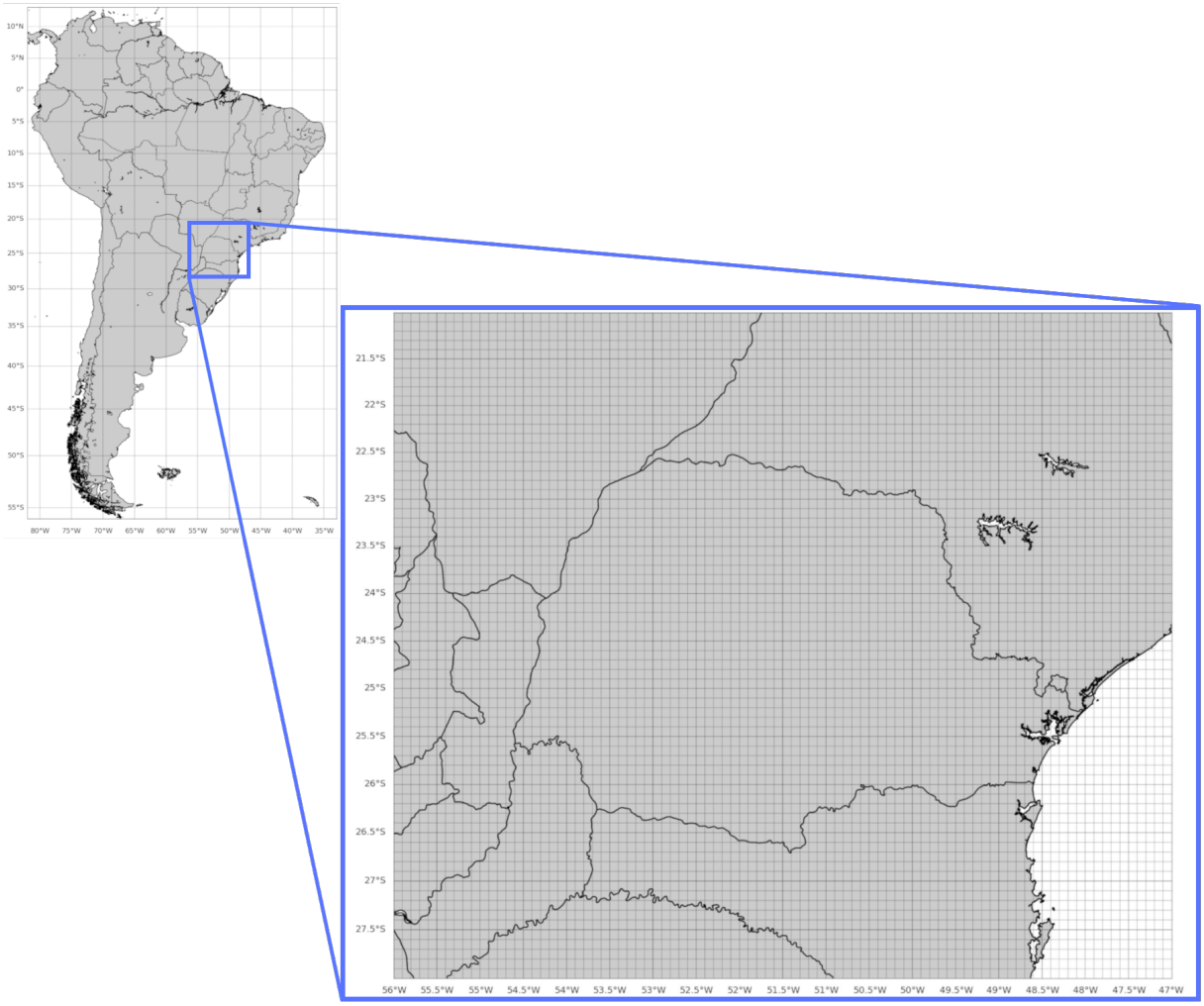

Figure 1.2: South America with an inset of our Region of Interest (ROI).

and its relevance for the economic scenario at a national level, with activities dependent on weather conditions. Adverse weather events are even more amplified in large urban centers, due to the vulnerability of the underprivileged communities and the lack of widespread response. Additionally, reliable lightning forecasts are essential for protecting the agricultural sector, which is a major component of the national economy, from lightning-induced fires and crop damage.

The forecast of CG events is of essence in this area, for the region's economic activities include lightning-sensitive industries such as farming, mining and energy generation. Our ROI contains more than 70 hydroelectric power plants, including the world's second-largest hydroelectric power plant, Itaipu (ANEEL, 2021). Itaipu is the result of cooperation between the governments of Brazil and Paraguay. It is responsible for supplying 8.6% of Brazil's energy demand and 86.3% of Paraguay's energy demand (Itaipu Binacional, 2023).

The power lines that pervade our ROI represent an essential asset in Brazil's energy distribution system. This asset is particularly dependent on the forecast of CG lightning, as it is one of the leading weather-related causes of power outages (Miyazaki and Okabe, 2010). The local power distribution company has atmospheric lightning and wind gusts as the two main

weather-related causes of outages, representing at least 23% of the known causes of energy disruption (Leite et al., 2011).

The forecast of TL or CG lightning as a product is offered by several enterprises, including public and private agencies. Different modes of lightning forecasting are employed by various actors, ranging from small civil protection agencies to multinationals corporations. Chapter 3 presents these various methods of lightning forecasting developed for operational and/or research purposes.

The solutions for lightning prediction offered by large private organizations are developed vying to provide for larger swaths of the planet and consequently, a large number of clients. Hence, the quality of the forecast tend to be substandard due to the scope of the area covered by the forecasting system, which will include regions of vastly different climatological conditions. Thus, to use their product for real-life purposes, clients need to develop and implement a form of bias correction so the product can be employed in their ROI.

Small organizations usually develop solutions considering local requirements, to meet the demands of specific clients or the needs of the people. These products may suffer from the same fault of the solutions offered by larger organizations, although by opposite reasons. While the large corporations may offer systems that are too inclusive, the smaller organizations may present products that are too niche, not meeting the client's necessity or also requiring bias correction. Besides, due to the small-scale of their operations, scalability can be a problem. The organization, although willing, may be unable to provide the solution to interested parties in a timely and uninterrupted manner.

Lightning forecasting is of essence, and so regions lacking an appropriate system make use of the broader models. This in turn demands the expertise of meteorologists to asses the forecast and correct it, which requires skilled human resources and time. Intent on providing a solution for lightning forecasting for our ROI to be used for practical purposes, we describe in this document the construction of a Machine Learning (ML) model for CG lightning forecasting.

1.2 MOTIVATION FOR THE STUDY OF ARTIFICIAL INTELLIGENCE FOR LIGHTNING FORECASTING

Lightning constitutes a relevant weather event, having an impact on multiple sectors of society. The accurate forecast of this event allows for the cessation of lightning sensitive activities, and the use of an automatic model for this forecast allows for the rapid dissemination of information in a continuous manner.

All around the globe, a myriad of sensors are working to acquire data describing a multitude of atmospheric variables. These sensors acquire different kinds of data in various space-time scales, so large amounts of weather data are created every day. These vast volumes of weather data constitute a form of Big Data (BD).

These large datasets hold relevant information about the weather, yet their sheer volume and complexity may impede forecasters' use to the fullest extent. Modern technological advances – such as increased processing power of computers, reduction in data storage costs, and better performance in managing vast amounts of information – stimulated the dissemination of Artificial Intelligence (AI) to several areas, including weather analysis and forecasting.

AI is a field of study that combines computer science and robust data sets to enable problem-solving. It relies on algorithms to build high-performance systems capable of making classifications or predictions based on collections of data (Russell and Norvig, 2010). The use of AI algorithms for weather forecasting allows for greater exploitation of the vast amounts of weather data available nowadays.

AI techniques enables the development of models based on statistical relations between large numbers of variables, considering previously known correlations and identifying entirely new ones. In sum, AI is a tool to increase efficiency, process and sort large volumes of data, and offload decision making.

The coupling of the knowledge of a human expert and AI has the prospect of generating more accurate forecasts. The physical understanding of the atmosphere by the meteorologist allied with AI's capacity to sift through and identify patterns in massive datasets facilitates model development.

One of the disciplines of AI is ML, a computational method capable of automating the construction of analytical or predictive models. ML was developed from the conception that computer systems can learn from data, identifying patterns and making decisions independently or with minimal human intervention (Géron, 2017).

The use of ML for model development should increase the efficiency of model parametrizations (Elsenheimer and Gravelle, 2019). An ML model in the weather forecast scenario must discover the rules and thresholds used in forecasting based on the available data. ML is well suited for atmospheric predictions, given that weather forecasts can be generated on demand and vast sets of weather data exist.

The ML model should perform equivalent or better than a human forecaster, with the advantage of being more consistent and automatic. ML models can deal with various problems and, provided sufficient data, be highly generic to be used in practical situations successfully (Michie et al., 1994).

1.3 OBJECTIVES AND HYPOTHESES

The purpose of this study is to develop an automatic CG lightning prediction model for the very short time range for the south-southeastern region of Brazil. This region was divided in a grid, hence the model must predict the number of lightning events per grid point for every designated time interval up to the defined lead time of one hour. Being so, this forecast task is a regression task.

For this end, our objectives include the investigation of the lightning conditions in this region; the identification of variables conducive to lightning occurrence; and the analysis of ML methods most apt for operational use.

Currently, the main lightning prediction method utilized in our ROI is a system based on Self Organizing Maps (SOM) – discussed in detail on Section 3.2. This model was trained using local lightning data, and has been employed for private and public interest purposes. Nevertheless, the model lacks interpretability and granularity, generating lightning occurrence maps of low resolution.

Thus, the hypothesis established in this proposal is that a different ML approach to lightning forecasting will yield a more interpretable model of better performance for our ROI. Model interpretability is a important characteristic for operational use of ML methods, given the explanation of the model inner workings is a desirable feature. The model's prediction may be used for decision making which involve cessation of industrial activities, which must be justified for the stake-holders.

The fundamental steps proposed for this study are as follows:

1. Review the literature pertinent to the problem, observing the state-of-the-art;
2. Collection and setup of the database, verifying gaps and examining the characteristics of the data;
3. Test different ML algorithms aiming at identifying the most suitable for the proposed task;
4. Evaluation of the model based on performance metrics;
5. Submission of a paper presenting our results in selected platforms;

The secondary objectives of this research are presented in the following list:

1. Verify if it is feasible to forecast the number of lightning events per area instead of just a binary lightning prediction (presence or absence of lightning);
2. Assess the quality of the forecasts for different lead-times, ranging from 5 minutes to up to one hour;
3. Experiments with different combination of features.

1.4 CONTRIBUTIONS

In our study we developed a CG lightning prediction regressor model for the Brazilian south-southeastern region. The model receives as input the location and number of events (i.e., the

amount of CG strikes per grid point in past times) and outputs the number of CG lightning per grid point for the next 05, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55 and 60 min.

Our main contribution lies in the problem representation: forecast lightning based on the distribution of past lightning occurrences in an area. By analyzing the correlation of the temporal series of lightning occurrences in the ROI, we verified that lightning occurrence is highly correlated to the past lightning conditions, as presented on Section 4.2. This indicates that lightning data can be used as a predictor variable for lightning incidence.

Our developed model is a regressor, i.e., its output is a continuous value. Our model outputs represents the number of lightning events in a specific area during a time interval. This is in contrast to the related works in the area, which focuses on the development of a classification model. These classification models results in a binary prediction of just the presence or absence of lightning occurrence. Its desirable to predict the amount of lightning instead of just its incidence for this indicates the severity of the atmospheric conditions in an area.

The maximum lead-time achieved in this work is defined at one hour of advance notice. This corresponds to the maximum lead-time verified in most of the related works. In this aspect, the distinguishing feature of our work is the update rate of the model. The model generates predictions in five min. intervals, generating 12 predictions per hour. This allows for the monitoring of the evolution of a thunderstorm event, accompanying its development and dissipation.

Predicting specifically CG lightning rather than TL is particularly important due to the direct impact CG lightning has on life and property. CG events poses significant risks because it directly strikes the surface, causing fatalities, injuries, and substantial damage to infrastructure such as buildings, power lines, and communication networks. Unlike IC lightning, which primarily affects aviation, CG lightning has immediate, tangible consequences on the ground. By focusing on predicting CG lightning, we deliver a tool for meteorologists to provide more targeted and actionable warnings, enabling better preparedness and response strategies for mitigating these risks.

One of the unique aspects of our work is the analysis of multiple years of lightning data. In our assessment of the related works, we verified that they considered a period of months of lightning data for model training, selecting the months that contained most of the lightning activity of the year. This is well-suited for academic study, intent on verifying the applicability of a ML method for weather prediction. However, given our intention of developing a model applicable for operational use, the consideration of multiple years of lightning data allows for the model to be trained on different weather patterns.

Additionally, given the availability of five years of lightning data, an analysis of the lightning patterns in our ROI was made, developing a climatology of lightning. Hence, we identified the seasons and regions most prone to lightning incidence, described on Section 4.2.

1.5 DOCUMENT ORGANIZATION

This document is divided in seven chapters, written according to the Introduction, Methods, Results, and Discussion (IMRAD) framework (Sollaci and Pereira, 2004). The chapters were developed following the structure of introduction to the subject, methods employed, results obtained, and discussion of the results.

Chapter 2 presents the theoretical basis for this work, describing the concepts of nowcasting, atmospheric electric discharges, AI, and ML. Chapter 3 introduces related work on lightning forecasting and the state-of-art in lightning nowcasting. This Chapter presents different approaches for predicting this weather phenomenon in the established literature.

Chapter 4 denotes the methods describing the datasets used and the work done upon them to make them applicable to ML and to assure that the data is in correspondence to the expected. Additionally, Chapter 4 denotes the process to develop a functional ML model.

Chapter 5 contains the results, presenting the ML model's forecasts and performance according to various metrics, presenting an interpretation of its results. Chapter 6 concludes the text, presenting an analysis of the work done and future steps in this research. Additional results are presented in the Appendix in Chapter 7.

2 THEORETICAL FOUNDATION

This chapter presents the theoretical basis of this study, exploring the principles that underpin the application of Machine Learning (ML) in the realm of lightning forecasting. Our objective in this chapter is to delve into the underlying meteorological and electrical principles that govern lightning initiation and evolution. To appreciate the potential of ML in lightning forecasting, it is crucial to understand the intricacies of the meteorological processes and various factors contributing to lightning occurrence.

The main concepts involved in this work from the fields of weather and computer sciences are expounded. The mode of forecasting denominated nowcasting is presented, followed by a description of the physics involved in cloud electrification and lightning occurrence. From the computer science-related disciplines, the principal themes are Artificial Intelligence (AI) and their ML methods.

By establishing a strong theoretical foundation, we lay the groundwork for understanding the intricate interplay between meteorology and AI. With this knowledge, we aim to harness the power of ML to advance our ability to predict lightning occurrences with greater accuracy and, in doing so, contribute to enhanced safety and risk mitigation across various sectors.

2.1 NOWCASTING: FORECASTING IN THE VERY-SHORT TERM

Weather predictions made in the very short term are known as **Nowcasting**, a neologism that arose from the junction of the English words *"now"* and *"forecasting"*. In 2010, the World Meteorological Organization (WMO) published a definition of nowcasting as *"forecasting with local detail, by any method, at a time interval up to six hours into the future, including a detailed description of prevailing weather conditions"* (World Meteorological Organization, 2017). Nowcasting is motivated by the need to accurately predict high-impact weather phenomena for specific locations with sufficient lead-time (Browning and Monk, 1982).

Nowcasting is prevalent in crisis management and risk prevention. Nonetheless, its realization is a highly complex and integrated task since the forecast must be made and updated in small timeouts. This forecasting format is helpful for the analysis of local to mesoscale events of short duration, such as flash floods, as well as sub-events of large-scale systems, like lightning from storms.

Due to the small temporal and spatial scale considered in nowcasting, there is a need for high-resolution and rapidly updated observations. The main nowcasting tools are weather radars, satellites, Lightning Detection and Location Network (LDLN), ground weather stations, upper air observations (e.g., weather balloons and radiosondes), and global/regional numerical models (World Meteorological Organization, 2017).

Fundamental to nowcasting is the rapid detection of high-impact weather systems and the swift emission of warnings to concerned parties. Proper use of nowcasting systems in conjunction with an adequate response from the responsible agencies can significantly reduce material losses and even prevent the loss of life in regions affected by severe weather.

The goal of nowcasting is to provide timely and accurate information that can support decision-making, risk assessment, and resource allocation in rapidly changing environments. By leveraging the latest data and advanced analytical techniques, nowcasting enables individuals, organizations, and policymakers to respond effectively to imminent events.

2.2 CLOUD ELECTRIFICATION AND LIGHTNING GENESIS

Clouds are atmospheric phenomena composed of water particles, known as hydrometeors, suspended in the air in different phases of matter. Cloud genesis occurs when a parcel of warm, moist air rises and cools in an adiabatic expansion process. Part of the liquid water in a parcel of air begins to freeze when the parcel rises above the 0 °C isotherm. However, part remains liquid, usually smaller particles, even when subjected to temperatures below 0 °C. These are denominated super-cooled particles.

In an atmospheric layer below -40 °C, virtually all the water in the air parcel is frozen. Between the isotherms of 0 °C and -40 °C, there is a zone of coexistence between the liquid and frozen particles of water, the mixed phase region – an essential zone in the electrification of the cloud (Rakov, 2016).

Cloud electrification requires a large-scale process that spatially segregates charged particles by polarity and a small-scale process that electrically charges the hydrometeors. Individually, these particles have little charge, yet large amounts of charged hydrometeors can generate zones with high voltage values (Williams (1985), Warner (2020)). The two main hypotheses that describe the generation and maintenance of cloud electrification are the **Convective Charging Theory** and the **Precipitation Theory**.

In the **Convective Charging Theory**, electrical charges come from a source external to the cloud – such as cosmic rays or coronas near the Earth's surface – and are transported through the cloud by the convective displacement of air (Rakov, 2016). The convective movement of the air is a natural way of distributing heat in the atmosphere.

A cold air parcel has a higher density than a hot air parcel, and thus naturally subsides, while the hot air parcel arises. Warm cloud constituents (e.g., rain drops) carry a negative charge, while cold cloud constituents (e.g., ice crystals) carry a positive charge. In this way, the particles are vertically distributed across the cloud, stimulating the cloud electrification process.

Precipitation Theory dictates that graupel (small hydrometeor composed of ice enveloped by liquid water) tends to accumulate in the lower layer of the cloud, acquiring a negative charge. In contrast, the smaller, lighter particles of ice crystals become positively

charged. These particles separate in the cloud by the force of gravity, i.e., the heavier particles fall to the lower regions of the cloud, and the lighter particles arise with the ascending air currents.

According to this theory, the charging of the particles occurs through two possible mechanisms: charge transfer by electrical induction; and charge transfer by collisions between hydrometeors of different phases and physical properties. In the latter mechanism, graupel particles and ice crystals collide in the presence of super-cooled water droplets. The cloud becomes significantly electrified when there is rapid charge transfer without aggregation, accretion, or coalescence of the hydrometeors (Beneti, 2012).

Furthermore, it is possible that cloud electrification is the result of a combination of mechanisms, with the process changing as the cloud becomes more electrified. The electrification process may be initiated by the collision of hydrometeors and intensified through the convective activity of the air that occurs in the storm (Rakov, 2016).

Every cloud is electrified to some degree, given the innate presence of electric charged particles in the atmosphere. The atmosphere is never neutrally charged, due to the continual electrification of the air arising from natural radioactivity and ionization from cosmic rays (Saunders, 2008).

Convective clouds, i.e., clouds originated by the convective displacement of air, tend to display higher values of electric charge. Intense convection facilitates the charge separation process, enabling greater values of electric charge to be harbored. These clouds can produce high enough electric fields to surpass the dielectric capacity of air and subsequently generate lightning (Wallace and Hobbs, 2006).

Of the various types of clouds, cumulonimbus are the main lightning source. Cumulonimbus are clouds characterized by their vast dimensions and distinct anvil-shape. These clouds manifest as inherent electrical structures, harboring high electric charges, being commonly associated with high levels of rain and lightning (Wallace and Hobbs, 2006).

The dynamics and distribution of electrical charges in a thunderstorm are complex and evolve continuously (Krehbiel, 1986). In a simplified manner, the electrical structure of a cloud can be defined as a vertical electric dipole, illustrated in the cumulonimbus on Figure 2.1. There is a positive charge center in the upper section of the cloud near the -40 °C isotherm and a negative charge center below, in the mixed phase region (Williams, 1985).

Cloud electrification has yet to be a fully understood and physically described phenomenon. Their vast size, intricate structure, and fleeting existence make them difficult to study comprehensively. This electrical process is intricately linked to two key aspects of the cloud: its overall movement (dynamics); and the makeup of its particles (microphysics) (Krehbiel (1986), Saunders (2008)). Our understanding of these critical aspects of storm dynamics remains incomplete. However, a thorough comprehension of these very elements is essential to elucidating how thunderclouds crackle with electricity.

One of the obstacles for the comprehension of the microphysics of a storm cloud lies in the fact that only indirect measurements of the cloud's characteristics from remote sensors such as

Figure 2.1: Charge distribution in an electrically charged cloud (Rakov, 2016).

radars and satellites are feasible. Remote sensors can provide data about the macro-characteristics of a cloud, but are unable to capture the more specific intricacies of this phenomena.

On-site sensors like radiosondes and weather balloons can profile specific micro-parameters of a cloud, but are unable to provide enough data for a climatological analysis of the microphysics of thunderclouds. To this end, multiple experiments of data acquisition must be made, which is unpractical given the high cost of these kind of sensors, which are typically disposable.

In addition, the lack of control of the measurements (e.g., control over the trajectory of a weather balloon or the storm itself) effectively prohibits the complete capture of a storm cloud by a sensor. Furthermore, the atmospheric conditions inside a thundercloud are severe. Few sensors can withstand such harsh conditions and those that can, can only do so for a limited amount of time, being inevitably targeted by hydrometeors or even a lightning strike. These factors inhibits the elucidation of how a cloud becomes electrified.

Lightning begins in the cloud electrification process. It is both an effect and an agent of the electrodynamic transport of charge and energy that occurs during the onset of a cloud's electrification (Jr. and Domingues, 2002). It primarily reduces the electrical energy of these systems by carrying negative electric current from the higher regions of the atmosphere to the lower regions (Rakov, 2016).

An atmospheric electric discharge occurs when the electric field locally exceeds the dielectric insulation of the air. Hence, it causes an electrical stress between two regions of

opposite charges so significant that dielectric collapse occurs and a lightning strike begins (Beneti, 2012). Lightning generates, deposits, and redistributes vast amounts of charge in a thunderstorm.

An electric storm typically begins with the presence of Intra-Cloud Lightning (IC), given the intense electric field that develops within a convective airstream (Elsenheimer and Gravelle, 2019). It is postulated that this occurs mainly for two reasons: the electric field above the negative charge region in an updraft is stronger than the electric field below; and the electric field required to surpass the dielectric insulation of air is smaller at the lower atmospheric pressure found in higher altitudes (Krehbiel, 1986).

In any given storm, near to two-thirds of lightning strikes are IC, with most of the rest being Cloud-to-Ground Lightning (CG) (Rakov, 2016). This arises from the smaller distance between the charges' centers in the cloud than the centers in the cloud and on the planetary surface. Furthermore, the turbulent motions near the updraft core where non-inductive charging from graupel–ice crystal collisions occur are favored.

The challenge in lightning nowcasting arises from the fact that lightning is a highly transient event, and direct measurements in-situ of this phenomenon are not possible – except mainly for rocket-triggered lightning, which does not constitute a natural lightning event. As mentioned, the lack of knowledge about how the cloud electrification process begins and develops also constitutes barriers to the complete understanding of this phenomenon.

The dissemination of atmospheric sensors capable of acquiring various kinds of data with satisfactory reliability and high spatiotemporal resolution stimulated the researchers to venture into forecasting more specific phenomena, such as CG lightning. LDLNs are now present in large portions of the world, allowing for accurate identification of lightning climatology.

In respect to our Region of Interest (ROI), the first lightning sensors were installed in Brazil in 1986 by the Minas Gerais Energy Company (CEMIG). After that, other agencies concerned with the occurrences of lightning also developed LDLNs for aiding their lightning-sensitive activities, such as power generation and civil defense.

In the year 2000, several LDLNs present in Brazil were integrated to form the National Integrated Network for Detection of Atmospheric Discharges (RINDAT), Brazil's national LDLN. RINDAT is the result of a collaboration between CEMIG, Eletrobas Furnas, the National Institute of Space Research (INPE) and Parana Technology and Environmental Monitoring Service (SIMEPAR) (Beneti et al., 2000). RINDAT provides data about Total Lightning (TL), discriminating between IC and CG events since 2018, for our ROI.

2.3 ARTIFICIAL INTELLIGENCE

Since before the inception of machines capable of doing advanced computation, there has been the idea of developing automatic machines capable of higher reasoning, in other words, AI. In modern times, the first scientific forays into AI date back to the 1950s, with Alan Turing's

pioneering work *"Computational Machinery and Intelligence"* (Turing, 1950). In this paper, Turing delves into the question *"Can machines think?"*.

In 1956, just a few years after the conception of the Turing Test, a summer research program at Dartmouth College became the official birthplace of AI, with John McCarthy's work (Russell and Norvig, 2010). Concomitantly, Allen Newell, Cliff Shaw, and Herbert Simon presented the Logic Theorist. It was one of the first programs to simulate human behavior when solving mathematical problems (Newell and Simon, 1956).

Intelligence as a concept has been studied by several disciplines and for various purposes. For instance, Philosophy concerned itself with the metaphysical questions associated with the existence of intelligent beings, while Neuroscience is more preoccupied with the biological aspects of intelligence. Thus, there are multiple working definitions of intelligence.

In the field of Neuroeconomics, Lee (2020) defines intelligence as the ability to solve complex problems or make decisions with outcomes that benefit the actor. In respect to the computer sciences, this definition closely resembles Russell and Norvig (2010) definition of intelligence, which is *"Intelligence is concerned mainly with rational action. Ideally, an intelligent agent takes the best possible action in a situation"*. In a general sense, AI refers to the ability of a machine to mimic capabilities of the human mind. Capabilities such as learning from examples and experiences, recognizing objects, understanding, responding, making decisions, and solving problems.

The field of AI aims at not only understanding intelligence but building intelligent agents as well. To that end, AI assimilated knowledge from numerous fields, including but not limited to Philosophy, Mathematics, Economics, Neuroscience, Psychology, and Statistics. Statistics was paramount to the development of AI, providing the elementary concepts to the field.

Figure 2.2 presents a Venn diagram of AI and its most closely related fields of study. Statistics encompasses AI and its sub-fields, defining the basis on which AI is build. As the field of AI matured, its methods began to branch out and new sub-disciplines emerged. ML and Expert Systems are modes of AI used for different tasks. The former is applied for tasks in which the computer must develop is own rules of operation, and the latter is employed with human made rules. Deep Learning is the most recent established form of ML, in which neural networks inspired by biological brains are constructed for various tasks.

Since AI's inception in the 1950s up to its ubiquitous presence nowadays, this field has seen multiple moments of great recognition, and consequently, financial funding, and also times of little attention. These periods are respectively known as AI Booms and AI Winters.

AI has grown into the mainstream in recent decades, defining a period of AI Boom. Terms like *"bots"* and *"neural networks"* are part of the common lexicon. A great variety of tasks are routinely executed by AI agents, so much that AI is part of the daily life of many people, even though the person using is not aware of it.

The dissemination of AI happened in association with both technological and theoretical progress. The former includes the ever-increasing processing power of computers, in addition to

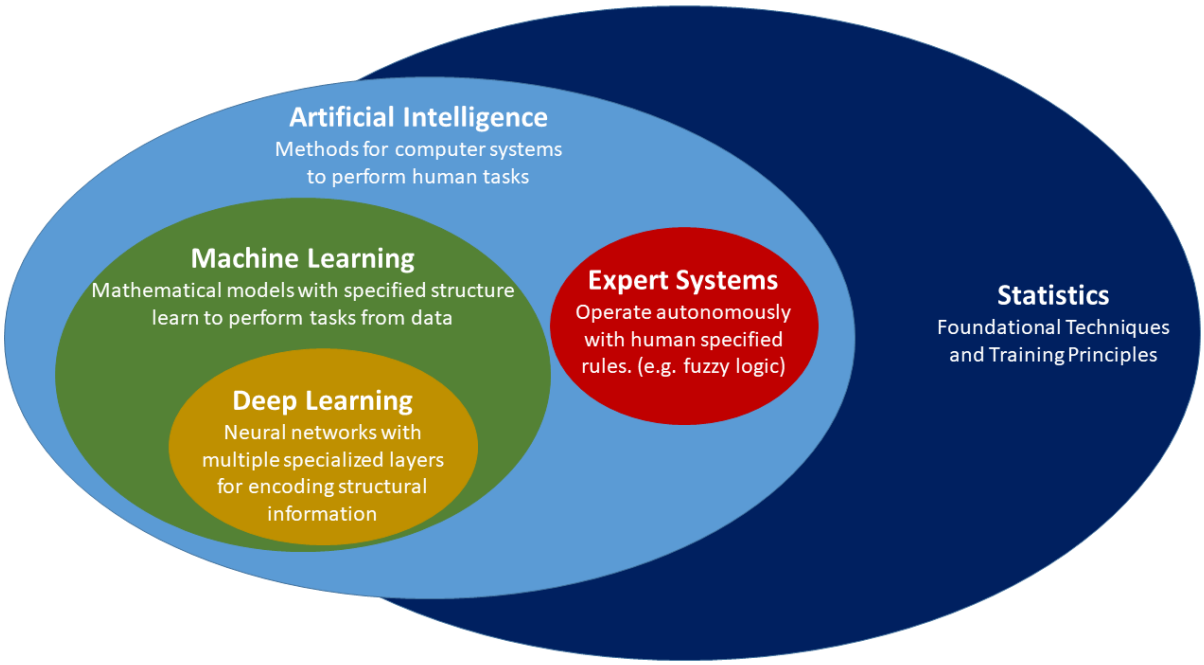

Figure 2.2: Relationship between AI, ML, Deep Learning, Expert Systems, and Statistics (Haupt et al., 2021).

the availability of large datasets, i.e., Big Data (BD), accompanied by the computational capacity to manipulate such vast amounts of data. AI demands huge collections of data to learn about the multitude of factors involved in the performance of a given task. The latter relates to the fact that AI has also evolved as an academic discipline, making more use of the scientific method and benefiting from greater integration between its sub-fields and related disciplines (Russell and Norvig, 2010).

2.4 MACHINE LEARNING

The term *"Machine Learning"* was coined by Samuel (1959) in his seminal paper *"Some Studies in Machine Learning Using the Game of Checkers"*. Samuel wrote a program for a computer to play checkers and save each move made in the game, such that the computer could study the moves and their results – and there lies his innovation. By analyzing the previously executed moves and their results, the computer could improve its performance in the game by examining the actions and their consequences, in fact learning from the data. Thus, the computer was able to better its performance based on its own experience of playing the game.

ML uses algorithms that iteratively learn from collections of data in order to improve their performance, acquire information, and make predictions. The structure of an ML algorithm differs from more traditional algorithms. Conventional algorithms are based on creating a set of rules and thresholds aiming to generate a result from data processing, while an ML algorithm must define its own rules of operation to achieve the established goal.

ML is devoted to building systems that learn, meaning systems that automatically improve their performance through experience. A computer program is said to have learned from

experience E concerning a class of tasks T and performance measured P, if its performance on tasks T, measured by P, improves with experience E (Langley and Simon, 1995). Learning involves the act of searching – one searches a hypothesis space to identify the hypothesis that best fits the available training examples and other constraints or prior knowledge (Mitchell, 1997).

An ML model requires a coherently established learning problem. For that, it is necessary a well-defined task, an evaluation metric to be improved, and a reference of training experience (Mitchell, 1997). ML performance depends on the quality of the source of experience provided, i.e., the training dataset.

The training dataset is composed of feature vectors and may also include target vectors, in which case is denominated a labeled dataset. A feature is used to represent numeric or symbolic attributes of a phenomenon. Features should be informative, discriminating, and independent from each other. Targets are the solutions to the tasks worked upon by the model.

Thus, developing the training dataset is fundamental for building an ML model. Based on the attributes of the training dataset, there are different manners in which the ML algorithm will learn from the data. ML is categorized as per the learning method used, with the two main modes of learning denominated **Supervised Learning** and **Unsupervised Learning**.

Supervised Learning is used when the training set is labeled. In this approach, the ML model will statistically analyze the features vector, identify patterns and correlations, and produce an output. The model analyzes its output with the correct solutions, i.e., the target vector, and hence the model learns from its successes and mistakes (Bishop, 2006).

Regarding **Unsupervised Learning**, the program performs an iterative process to identify a structure in the unlabelled data. After the model's training, both learning methods use a test dataset to accurately depict the algorithm's performance by applying the ML model to data that was not initially trained upon.

Beyond the more common forms of learning, there is also **Semi-supervised Learning**, used for datasets containing labeled and unlabeled data. In the **Reinforcement Learning** method, the system observes the environment, selecting and performing actions, obtaining rewards or penalties according to its performance, differing from other methods by not using a training set (Géron, 2017).

AI and ML are reckoned as universal disciplines. Their methods are applied to fields ranging from biomedical to linguistics and, evidently, meteorology. ML programs can be classified according to the class of task performed, being divided into algorithms for **Classification**, **Regression**, and **Clustering**.

In **Classification**, the goal is to categorize an element of a set into a class among a set of classes, defining a discrete-valued answer. Classification algorithms can be used for a profusion of tasks, among them perhaps the most emblematic one being computer vision. In this task, the computer must identify objects in static or moving pictures and assign them to classes.

A task whose outcome is a continuous variable defines a **Regression** problem. Regression methods pertains to tasks associated with probabilistic outcomes instead of more deterministic

ones. Regression is fit for tasks such as forecasting weather events and predicting the chance of a event occurring in a specific period. Certain classification tasks can also be performed with regression methods. Some regression algorithms originate from adapted classification algorithms so that not only the class to which an element belongs is presented but also an associated probabilistic value describing the element's probability of belonging to the defined class or any other class.

Clustering algorithms are used to join elements of a set into groups. The elements in the same group are ideally very similar to each other and dissimilar to elements in another group. Clustering is especially useful when the number of groups and their characteristics are not known a priori – such as in anomaly detection.

The classification of ML methods also considers their form of generalization as a criterion. Given a set of examples, the system should be able to generalize to examples it has never encountered before. An ML model is classified as **Instance-based** if the model compares new data to known data, or **Model-based** when the system detects patterns in the training set and builds a predictive model. The system can also be categorized according to its ability to learn incrementally from a data stream, denominated **Online Learning**, or not, known as **Batch Learning**.

ML is of practical value for innumerable domains. It is particularly useful for data mining in sets that may contain valuable implicit regularities that can be discovered automatically, poorly understood fields where humans may not have enough knowledge to develop effective algorithms or cases where the program must dynamically adapt to changing conditions (Mitchell, 1997). One of the significant challenges of AI is solving tasks that are easy for people to perform but difficult to describe in formal terms – problems that are solved intuitively, such as speech or facial recognition (Goodfellow et al., 2016).

Even with the nearly ubiquitous adoption of ML in meteorology, this technique is often criticized for its inability, in some instances, to present the rules and thresholds used by the system to make its predictions. Because of this handicap, many users focus on system interpretability, using techniques such as Model Interpretation and Visualization (MIV). In addition to aiding in understanding the model, MIV can identify new hypotheses for further investigation, such as new features for the training of the model.

With the dissemination of ML in the weather sciences, the trade-offs between predictability and interpretability must be understood. A model can make better forecasts at the cost of becoming less understandable by human forecasters (McGovern et al., 2019). This entails a choice, for models that are not understood by the user are typically less used, given that the user tends to distrust the model.

2.5 SUMMARY

This Chapter presented the hypotheses and concepts that form the theoretical basis of this study in the nowcasting of CG lightning with an ML model. The Chapter begins by declaring the four main concepts involved in this work: nowcasting; cloud electrification and the subsequent occurrence of lightning; AI; and ML.

The first section is concerned with the presentation of nowcasting, a form of prediction made in the scale of a few hours in the future and most related to the forecast of high-impact, rapid evolution weather events. It is followed by the section elucidating the process of cloud electrification and atmospheric conditions conducive to lightning occurrence. The section emphasizes the complex and dynamic nature of lightning, highlighting the challenges associated with accurate prediction.

Subsequently, the Chapter brings the concept of AI, presenting a brief text about its history and popularization to multiple fields of study. To conclude, the Chapter outlines the basic principles of the ML method. In the next Chapter, the literature reviewed on lightning nowcasting methods is set forth.

3 LITERATURE REVIEW

The field of atmospheric science has long been captivated by the unpredictable nature of lightning. With its dazzling displays of light and sound, lightning remains both a captivating natural phenomenon and a formidable meteorological hazard. The ability to accurately forecast lightning not only piques scientific curiosity but also holds significant implications for public safety, aviation, agriculture, and a wide range of industries. However, lightning remains a complex and challenging phenomenon to forecast with precision.

The present Chapter comprehensively examines existing research and studies related to lightning prediction, focusing on applying Machine Learning (ML) techniques. Lightning poses significant challenges in forecasting due to its complex and dynamic nature, which has prompted researchers to explore innovative approaches to improve prediction accuracy and lead time. This Chapter aims to critically analyze and synthesize the literature to identify the current state-of-the-art, gaps, and areas for further exploration in lightning nowcasting using ML.

3.1 LIGHTNING NOWCASTING: PAST, PRESENT AND FUTURE

Atmospheric events, specially severe weather, impact our daily lives. Knowledge about their occurrence and evolution proves to be invaluable since immemorial times. Hence, before the existence of sensors capable of describing the atmosphere, arose a forecast method termed Persistence Forecast.

Forecasting by persistence is based on the idea that current weather condition will persist and that future weather will be the same as the present (e.g., if it is raining now, a forecast predicting rain for the next hour) (Wilks, 2006). This method can be employed based solely on short-time observations. For its simplicity, this method proves to be useful for various situations, considering the forecast is made for a specific location and for a short lead time.

Storms are high-impact weather phenomena and can be predicted with more straightforward methods and less data in comparison with lightning, given their broad influence on local atmospheric conditions. In past times, due to limited resources and a lack of data sources, weather forecasters focused on predicting the storms, with few efforts made concerning the forecasting of lightning. Lightning was predicted as a consequence of the storms, not the focus of the forecast per se.

As technology progressed and the world became more interconnected and interdependent, lightning forecasting began to gather the attention of meteorological services. As the first systems for lightning prediction became operational, they became a basic necessity for various sectors – although they lacked accuracy and had lead-times in the order of minutes, their operational value was quickly recognized. Different methods of lightning nowcasting have been implemented throughout the years, varying in the focus of the prediction to the methods used.

As long term data became available, statistical methods based on Numerical Weather Prediction (NWP) were employed for the forecast of lightning. NWP is a computational intensive forecast method in which a series of processes to predict future atmospheric conditions are made by solving dynamics and physics equations that describe the movements and changes of the atmosphere (Schön and Dittrich, 2019). Considering the innate complexity of thermodynamic and fluid mechanics equations, NWP models are executed on supercomputers. NWP models process current weather observations gathered from multiple sensors to forecast future weather for a Region of Interest (ROI) or even the entire world.

In relation to lightning forecasting, NWP is used to determine whether the environment is conducive to the formation of thunderclouds and subsequent generation of lightning. Based on the sources of data used, a formula expressing the relationship between the measured atmospheric conditions and the occurrence of lightning is derived. This formula may be obtained using statistical processing such as multivariate analysis or linear regression (Sato et al., 2008).

The main fault of this method is that the rules on which the forecasts are made are static, meaning that the same thresholds are considered for storms that happen at different times of the year and under different conditions. Moreover, NWP demand substantial computational resources, which restricts their processing speed, particularly for nowcasting applications.

In contrast, the advantage presented by this method is that it can be tuned to apply to other regions, provided that the local climatology is known. Furthermore, the rules implemented are entirely explicit for the forecasters, allowing for a more straightforward interpretation of the predictions and facilitating human-based input.

The ML method is proposed as an alternative to statistical techniques. The computational demands of simulating lightning within numerical models hinder the efficiency of lightning nowcasting, where timeliness is vital. In comparison, observation-based data-driven lightning models have surfaced as efficient methods for producing lightning forecasts, utilizing ground-truth data with lower computational costs (Mansouri et al., 2023).

ML is used to build forecast models based on local data, enabling the system to develop its own rules and thresholds. With the advent of ML in the weather sciences, the prospect of better forecasts of atmospheric phenomena based on vast amounts of multi-sourced data was established. Data from satellite, radar, and other mesoscale observing networks are becoming increasingly available at an astonishing speed, which makes it impossible to extract useful information for lightning prediction manually (Zhou et al., 2020).

The ML method benefits from this abundance of data, but the acquisition and preprocessing of these large datasets remains challenging. Data from different sources describe different physical parameters and are captured at different spatiotemporal scales. Even in the highly automated method of ML, experienced meteorologists are still required. Different observations differ significantly in their physical meanings, and experts in the subject are needed to explain and extract the relevant data.

The increased accessibility of ML tools and frameworks has empowered meteorologists to innovate and refine predictive models, further advancing the efficiency of weather forecasting. In our work, we decided to use ML for lightning forecasting for the following:

1. ML algorithms can process vast amounts of data, enabling the identification of complex patterns and relationships that traditional methods might overlook;

2. Computational efficiency – low computational resource requirements in comparison to NWP methods;

3. Short time to generate forecasts after the model is trained – an important feature given that our model updates in 5 min. timesets.

3.2 RELATED WORK AND THE STATE-OF-THE-ART

In this Section, the literature on lightning forecasting is discussed. Our search for this literature was done on the Google Scholar and the American Meteorological Society (AMS) Journals databases. The terms used for this search were: lightning; atmospheric electrical discharge; prediction; forecast; nowcast; ML; and Artificial Intelligence (AI). The articles found with this search were filtered as to select the papers describing the development of a model (be it statistical or based on AI) for lightning prediction in the very-short term. This resulted in a selection of papers defining the state-of-the-art in lightning prediction.

Weather forecasting systems are mathematical models and thus can be analyzed by several metrics. The decision of which metric to consider depends on the model's application. However, besides the simple consideration of a given model's accuracy or Critical Success Index (CSI), several other characteristics are relevant for assessing a model's quality.

The preeminent traits of a weather forecasting model are its spatial and temporal resolution and lead time. A high spatial resolution allows the human decision-maker to issue warnings more explicitly considering the affected area. The first global weather models had a resolution in the hundreds of kilometers – nowadays, this resolution has increased to the tens of kilometers.

A rapidly update model with a high temporal resolution can track the rapid changes in the weather. Nowcasting models have a temporal resolution ranging from a few minutes to a couple of hours, while models for longer forecast times are updated every few hours. A satisfactory lead time enables actors to take measures in anticipation of high-impact phenomena.

Table 3.1 presents the characteristics of the selected works reviewed for this study. NWP based techniques and the more modern method of ML are presented. The texts presented concern both the nowcasting of Total Lightning (TL) as of Cloud-to-Ground Lightning (CG) lightning. Models currently in operational use, i.e., applied in real-life scenarios, and models developed for research are presented.

Table 3.1: Correlated work in lightning nowcasting.

Paper	Data Source(s)				Forecast		Purpose		Spatial Resolution (km)	Temporal Resolution (min.)	Lead Time (min.)	Method	Region of Interest	Results
	LDLN	Satellite	Radar	Radiosonde	Total Lightning	Cloud-to-Ground Lightning	Research	Operation						
Sato et al. (2008)			●		●			●	1	10	30	Statistics	Japan	Forecast of TL with a CSI of 0.3-0.4
Leite et al. (2011)	●				●			●	2	1	60	Neural Network (SOM)	Sao Paulo and Parana, Brazil	Forecast of TL density and record of events
Schön and Dittrich (2019)	●	●			●		●		Undefined	Undefined	15	Deep Learning	Germany	Forecast of TL with the best performance at 15 min.
Zhou et al. (2020)	●	●	●			●	●		5	Undefined	60	Deep Learning	Central-eastern and Southern China	Successful integration of multisource data. Forecast of CG with a CSI of 0.36
Mansouri et al. (2023)		●			●		●		10	15	60	Deep Learning	Northern South America	Forecast of the probability of TL occurrence
Song et al. (2023)		●			●		●		Undefined	Undefined	60	Gradient Boosting	Continental USA	Forecast of TL with innovative features (aerosol data) with a CSI of 0.53
Our Proposal	●					●	●	●	10	5	60	Gradient Boosting	South-Southeastern Brazil	Forecast of CG for the next 60 min. in 5 min. intervals

TL - Total Lightning; CG - Cloud-to-Ground Lightning; LDLN - Lightning Detection and Location Network; CSI - Critical Sucess Index; SOM - Self Organizing Maps; SVM - Support Vector Machine; USA - United States of America.

Traditional lightning forecasting methods are based on statistical models and NWP. Their quality depends on the physical parameterizations and approximations considered, given that it is only possible to know the current state of the atmosphere worldwide partially.

Natural disasters beset Japan. Tsunamis and earthquakes have ravaged the country, and thus great efforts have been made to develop early warning systems. Sato et al. (2008) developed a TL forecasting model using radar data based on the previous success of early warning systems for earthquakes, even considering their short lead time. In this work, the forecast of lightning is made with two modules: the first is used to predict the formation, transition, and decline of thunderclouds by verifying the degree of reflectivity found in radar images, and the second predicts the potential for lightning strike from the state of thunderclouds using statistical methods. The equation is derived from past events of TL associated with radar data.

The authors expect that the information provided by this system may be used empirically for reference purposes. The preeminent feature of the model is it's high resolution (1 km). However, the model's outputs define a binary prediction, and so it does not provided data about the severity of lightning events, in contrast to our regressor. By predicting not only the occurrence of lightning, but also its quantity, we can supply data indicative of the severity of the atmospheric phenomenon.

The research problem in Leite et al. (2011) proposed improving risk management associated with lightning disturbances in power transmission systems. The determining factors for the occurrence of power transmission faults are the characteristics of the lightning event, such as its peak current, and the properties of the power line itself, e.g., its insulation level and

transmission capacity. Self Organizing Maps (SOM) were used to build a neural network that generates maps with a 4 km^2 resolution updated minute-by-minutely.

The developed SOM model did not discriminate between Intra-Cloud Lightning (IC) and CG discharges, which is a interesting feature for the energy sector, given that their main interest is in the prediction of CG strikes. The resulting network was able to predict TL based on lightning data in a ROI similar to our defined ROI. Hence, we defined this model as one of the baseline models to be compared with our nowcasting model. The results of this comparison are presented in Section 5.3.

With the popularization of ML, weather scientists began to look for more specific and complex tools from this field to assist in their work. Convolutional Neural Networks (CNNs) are a mode of ML inspired by how a biological brain operates, and this technique gave origin to deep learning. Deep learning is an ML method based on neural networks with several layers. The works of Schön and Dittrich (2019), Zhou et al. (2020), Cintineo et al. (2022), Leinonen et al. (2022) describe the development of neural networks for lightning prediction. We chose to consider Schön and Dittrich (2019) and Zhou et al. (2020) in the state-of-the-art for the innovative characteristics of the work, which are subsequently presented.

Schön and Dittrich (2019) implemented a CNN to nowcast TL using a architecture inspired by UNet++ and ResNet based on satellite images and data from a Lightning Detection and Location Network (LDLN). The authors previously conducted a similar work using Random Forest (RF) as the algorithm (Schön et al., 2018). Thus, this article presents an updated and more complete version of this work. The CNN performs a binary classification task, viz., the presence or absence of lightning for the next 15 min. The algorithm was trained on a dataset of satellite images and lightning observations taken between 01/06/2017 and 04/07/2017. Considering the forecast is based on images, the authors note that a CNN was a natural choice for the task presented.

In their work, the authors try to distinguish areas affected by thunderstorms and areas of fair weather, tackling a task closely related to image segmentation. Their results indicate that the CNN achieves a similar performance compared to RF. However, CNNs offers the advantage of directly processing image slices instead of single pixel values compared to RF, eliminating the need for additional preprocessing steps. The faults of this model is the lack of discrimination between CG an IC discharges and the short lead time, defined at 15 min. Moreover, the period considered for training is approximately one month, differing from our work that leverages multiple years of data.

Zhou et al. (2020) developed a deep learning network trained on satellite, radar, and LDLN data. The authors claim that this is the first time a deep learning algorithm has been used to integrate multi-source observation data in lightning nowcasting and extract lightning initiation features. This work used multi-source data but also tested different combinations of sources to verify which resulted in the best performance (e.g., only satellite; satellite and radar; radar and LDLN; satellite, radar, and LDLN). The network was based on SegNet and named

"LightningNet", with 29,128,577 trainable parameters, which required many training samples to prevent overfitting. It was verified that the more predictors LightningNet used, the better the performance.

The authors attribute the capability of multi-source data analysis of the LightningNet to the robust feature learning capability of the network. Nonetheless, the authors also consider that additional details about the mechanisms need to be further explored in the future. The developed network outputs predictions for the next hour only, which prevents the monitoring of the evolution of a storm event. This feature is present in our work, for the predictions are generated for every five min. interval for up to one hour.

In Mansouri et al. (2023), the authors propose to treat lightning nowcasting as a video processing problem, using a residual U-Net structure for this task. The dataset used in this work is composed of one year of TL events sourced from satellite. The data was processed as to generate zero-one lightning maps, i.e., the presence or absence of lightning strikes. The resulting network outputs continuous values between zero and one indicating the probability of lightning for the next 15, 30, 45, and 60 min.

The main finding of this study is that an increase in the model's size does not entails in better prediction capability, as this capability reaches a plateau. Thus, this work motivates the use of smaller models to reduce memory usage and computation costs. For the next steps, Mansouri et al. (2023) declares their intention to extend the period of data considered, from one year to several years. Our model stands apart from this work for the predictions are generated in 5 min. increments and for the use of six years of data.

Song et al. (2023) based their work on the novel use of aerosol as a feature for lightning prediction. The authors state that a key limitation of current ML models for lightning prediction is their exclusive reliance on meteorological data, overlooking the substantial impact of aerosols on lightning patterns. The lightning data was obtained from satellite, while the meteorological variables and aerosol data (aerosol optical depth and composition) were sourced from earth monitoring models. The selected model was a LightGBM gradient boosting framework, with its output being a binary classification result, where 0 represents no lightning occurrence and 1 represents lightning occurrence within the next hour.

The foremost contribution of this study is the assessment of the viability of aerosol data as a feature for lightning prediction. Aerosols influence lightning occurrence by stimulating convection, promoting particle collisions and enhancing charge dissipation. Conversely, aerosols also have significant radiative properties that inhibit particle activation. This consistency indicates that aerosol data is useful as a temporal predictor for lightning. Although the authors innovated in the use of aerosol variables as input, the model was developed only for research purposes. The model is a binary classifier, which indicates the presence/absence of lightning for the next hour. Thus, it does not provide important features for operational use present in our work, such as the lightning conditions for the next hour in 5 min. increments and the severity of the thunderstorm, by presenting the number of lightning events.

Most of the reviewed works presented in this Chapter concerns the forecast of TL. The prediction of TL inhibits the discrimination between IC and CG discharges, i.e., the differentiation between purely atmospheric discharges and events capable of affecting society. Focusing the prediction on CG lightning allows for alerts to be sent when a hazard is possible, not for when simply there is electric activity in the atmosphere. The emission of alerts imply in the interruption of lightning-sensitive activities, which entails a cost. Additionally, IC lightning typically precedes CG lightning, and so alerts regarding TL events may be sent earlier than necessary, entailing a longer stoppage time for lightning-sensitive activities.

Models developed for research purposes may not be fit for operational use. These works focus on the identification of relevant variables for lightning prediction or the interpretation of how the model generates results. Factors such as the time needed to generate the predictions and if the data will be available in the foreseeable future may not be considered. The process of Research to Operations (R2O) is not simply done. A multitude of factors must be considered for the operationalization of the model to be successful, including the definition of the required computational resources, continuous availability of data, contingency measures to deal with lack of data, and assigned personnel to provide support for it.

Our study aims to build a ML model for the nowcasting of CG lightning fit for operational use. The inner workings of the model are also discussed, intent on providing clarity to the end user and facilitate the interpretation of the model's forecasts. The foremost contribution of our work lies in the development of a new methodology for lightning prediction, resulting in a tailor-made model for a region that lacks a reliable CG lightning prediction. In Chapter 4, a complete description of our methodology is introduced.

3.3 SUMMARY

The chapter commences with an overview of traditional lightning forecasting methods, including statistical models and the physical parametrizations involved. It highlights the limitations of these approaches in capturing the intricate relationships between atmospheric variables and lightning activity.

Subsequently, the Chapter delves into the emergence of ML as a promising alternative for lightning prediction. It examines machine learning algorithms employed in previous studies and assesses their effectiveness in capturing lightning patterns. In the next chapter, we expound our proposition for the task of CG lightning nowcasting.

4 STUDY METHODS

This chapter presents the methods of our study, which focuses on the use of Machine Learning (ML) for Cloud-to-Ground Lightning (CG) nowcasting for the next hour in 5 min increments. This segment provides an overview of the datasets used, a cardinal step for using ML techniques, and the methods that will guide this work.

The data collection process is presented, highlighting the significance of processing the datasets. Historical lightning data was collected from a Lightning Detection and Location Network (LDLN) present in the Region of Interest (ROI). Data processing includes removing untrustworthy data, normalization, and feature engineering to ensure the quality and relevance of the input data for the ML model.

Section 4.2 presents the exploratory analysis conducted on the CG lightning data at our disposal. This analysis was done to find the CG occurrence patterns on the ROI, which in turn allows for the curation of the training data selection for the model. Subsequently, The methods employed are presented, denoting the steps in developing an ML model for CG lightning prediction.

The computational aspects of our study were all performed using Python 3.9.13 (Van Rossum and Drake, 2009) and libraries for data analysis. The reading and processing of this data were done with the Pandas 1.4.4 library (Wes McKinney, 2010), useful for the manipulation of datasets, and Numpy 1.22.4 (Harris et al., 2020), applicable for numerical operations of data. The ML algorithms were sourced from Scikit-learn (Pedregosa et al., 2011). The visual data presented on the maps in this document made by the authors were developed with Matplotlib 3.6.2 (Hunter, 2007). The computational system used is a MacBook Pro released in 2017, and it's specifications are described in Table 4.1.

Table 4.1: Description of the computer system used in our study.

MacBook Pro 2017	
Operational System	macOS Ventura 13.0.1
Processor	2,3 GHz Dual-Core Intel Core i5
Memory	16 GB
Storage	256 GB

4.1 LIGHTNING DATA COLLECTION AND PROCESSING

The analysis, monitoring, and prediction of lightning is possible through various types of weather sensors. All forms of lightning are associated with charge displacement and, therefore, can be detected by measuring the electric and magnetic fields related to the movement of charges (Rakov, 2016). The primary tool for detecting and characterizing lightning events is a LDLN.

LDLNs are arrays of ground-based sensors that monitor different radio wave bands to detect, locate, and characterize lightning. The quality of the data sourced from a LDLN depends on the frequency of the waves surveyed by the sensors and the density of receiving sensors (Rudlosky et al., 2019). The number of sensors that comprise the LDLN defines the spatial coverage and accuracy of the network. While a single satellite or radar can cover vast sways of a region, LDLN demands the installation of multiple sensors to accurately capture the lightning occurrences in a region.

LDLN collect data in real-time, providing lightning information with a time precision of milliseconds, in contrast to other systems like satellites and radars. Given that CG strikes are easier to detect and most prone to impacting society, initially, most LDLN were capable of only detecting CG lightning. Modern LDLN have progressed to detect and discriminate Total Lightning (TL), i.e., both Intra-Cloud Lightning (IC) and CG lightning. Besides discriminating between IC and CG discharges, LDLNs also define the localization of the point of impact of a CG event on the Earth's surface.

Alas, the entirety of Brazil is not covered by a LDLN. Given that most socio-economic activities occur in the country's southeastern region, the National Integrated Network for Detection of Atmospheric Discharges (RINDAT) was developed with a focus on this area. RINDAT is comprised of 19 sensors distributed in the states of Sao Paulo (SP), Minas Gerais (MG), Parana (PR), Goias (GO), Mato Grosso do Sul (MS), Espirito Santo (ES), Rio de Janeiro (RJ), and the Federal District (DF). These sensors are maintained by the Minas Gerais Energy Company (CEMIG), Parana Technology and Environmental Monitoring Service (SIMEPAR), Furnas and National Institute of Space Research (INPE). Table 4.2 presents the city that houses each sensor and responsible agency. The location of each sensor is presented in the map on Figure 4.1.

RINDAT's sensors boasts a detection range of 625 km with a mean error of 0.5 km to 2 km in the localization of the electric discharge. Exceedingly intense lightning can be detected by sensors from thousands of kilometers away, but these events are not reliably described and thus disregarded. RINDAT displays efficiency in lightning detection of 70% to 90%, hence capturing at least 70% of all lightning events in its observation zone. It can discriminate between IC and CG lightning with an accuracy of 80% to 90% (Beneti et al., 2000). The performance of RINDAT varies according to the region considered depending on factors such as the number of sensors observing the region and local climatological conditions.

From 2000 until 2018, RINDAT was not a TL network, capturing with reliability only CG events. In 2018, the network's processing central was updated, and new sensors were integrated into the network, thus creating a more homogeneous and reliable LDLN. From 2018 onwards, RINDAT could detect and discriminate TL. Because of this, this work considers RINDAT's data from 2018 until 2023.

The Brazilian south-southeastern region is observed by RINDAT. This LDLN supplies data about the date and time of the event with a precision of milliseconds, localization, an estimate

Table 4.2: Localization of RINDAT sensors and responsible agency.

City	**Agency**
Tres Marias - MG	CEMIG
Camargos - MG	CEMIG
Serra da Canastra - MG	CEMIG
Pissarao - MG	CEMIG
Irape - MG	CEMIG
Castro - PR	SIMEPAR
Paranagua - PR	SIMEPAR
Andira - PR	SIMEPAR
Uniao da Vitoria - PR	SIMEPAR
Guaira - PR	Furnas
Novo Horizonte - PR	Furnas
Brasilia - DF	Furnas
Rio Verde - GO	Furnas
Serra da Mesa - GO	Furnas
Itajobi - SP	INPE
Araraquara - SP	INPE
Tres Lagoas - MS	INPE
Serra - ES	INPE
Rio de Janeiro - RJ	INPE

of the intensity of lightning discharges, location uncertainty measurements, type of lightning, and the number of sensors that reported the lightning.

The following data sourced from RINDAT is considered in this study: date (year, month, and day); time (hour, minute, and second); spatial coordinates (latitude and longitude); peak current value (Kiloamper (kA)); polarity (positive or negative); lightning type (CG = 0 and IC = 1); and the number of sensors that declared the event. Figure 4.2 displays a sample of the data collected from RINDAT about TL events that happened on the ROI in 2018.

The following data sourced from RINDAT is considered in this study: date, as defined by the year, month, and day (int64); time, obtained from hour, minute, and second (int64); spatial coordinates as declared by the latitude and longitude (float64); peak current value in kA (int64); lightning type, with CG being zero and IC being one (int64); and the number of sensors that declared the event (int64). Figure 4.2 displays a sample of the data collected from RINDAT about TL events that happened on the ROI in 2018.

RINDAT's lightning data was made available by SIMEPAR for our research. Yearly files in Comma Separated Values (CSV) format containing TL data were provided from 2000 until 2023. The complete dataset of TL in the ROI from 2000 to 2023 contains 150,043,364 lightning events. Considering the period from 2018 to 2023, this dataset is reduced to 65,873,198 electric discharges, with 50,057,460 of those events being IC and 15,815,738 being CG events. As presented on Table 4.3, the majority of events are IC discharges.

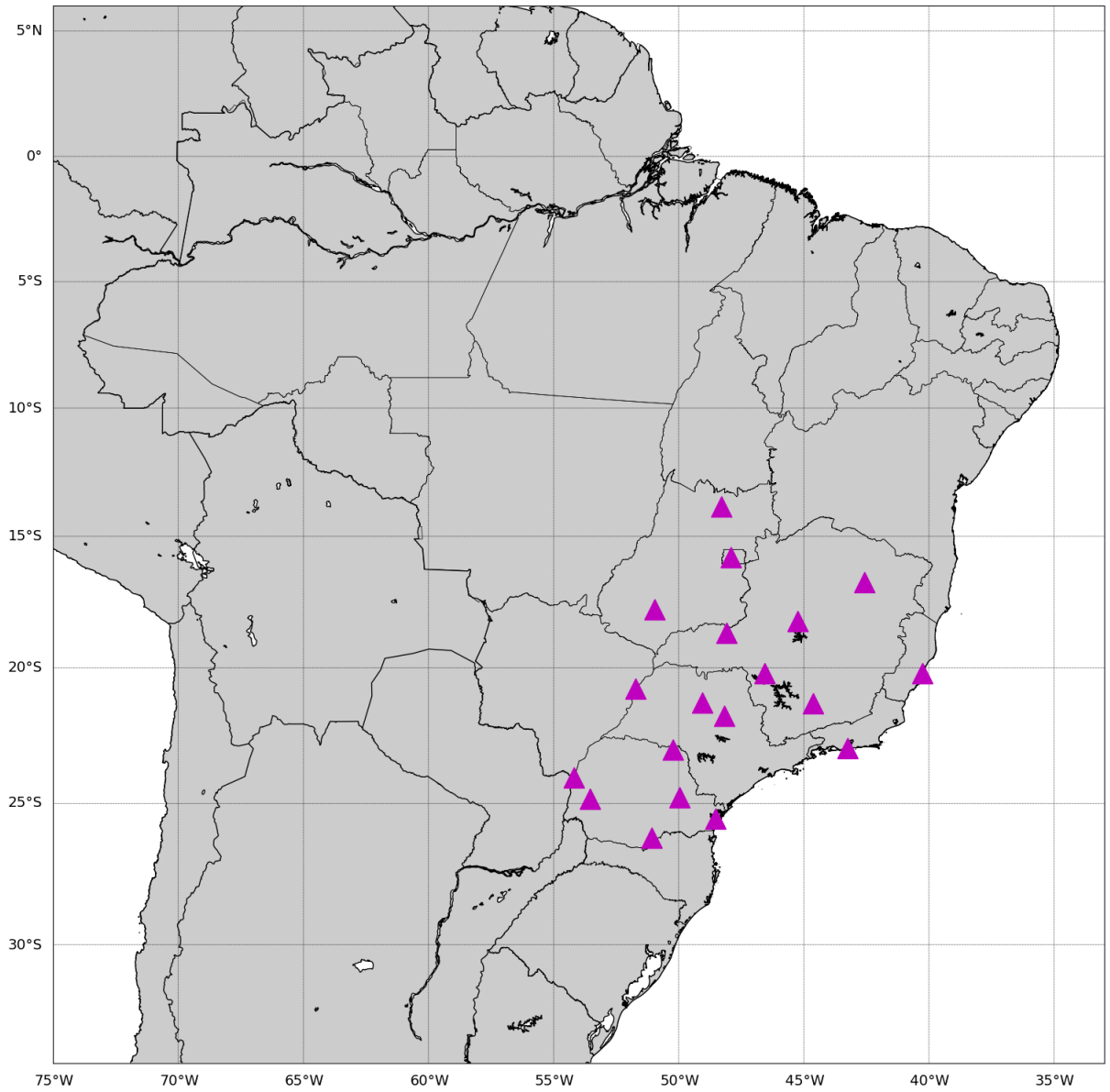

Figure 4.1: Network of RINDAT sensors (magenta triangles) in Brazil.

year	month	day	hour	minute	second	nanosecond	latitude	longitude	currentpeak	lightningtype	numsensors
2018	1	1	0	0	7	770365440	-25.1366	-49.0414	3	1	4
2018	1	1	0	0	8	351351040	-24.8817	-48.9767	-8	1	3
2018	1	1	0	0	10	345615360	-25.0889	-49.2192	5	1	3
2018	1	1	0	0	11	539907840	-25.2046	-49.1562	-7	1	3

Figure 4.2: Sample of the dataset of TL events on the ROI declared by RINDAT.

Given that RINDAT may not detect every lightning occurrence and assign non-lightning electric events as lightning, the deletion of spurious events was the first procedure in processing the data. This deletion process was based on two pieces of information: (i) the peak current value; and (ii) the number of sensors engaged in detecting lightning.

Firstly, events with a declared peak current equal to 0 kA were deleted. These events may have been lightning occurrences of low intensity, but given that the sensor could not characterize

them, they are deemed unreliable data. This work considers only the absolute value of the peak current, given that the polarity of lightning is a convention related to whether the lightning was going upwards or downwards (Rakov, 2016). The movement of charges always follows from the negative center of charges to the positive center.

Any single sensor can detect lightning, but the reliability of the event's characterization comes from multiple sensors capturing and describing the event. To visually present the importance of defining a minimum number of sensors, two maps were developed, presented on Figure 4.3. The left map is of TL density considering lightning events captured by one or more sensors; and the right map is of TL density considering lightning events captured by at least three sensors (i.e., employing a minimum number of sensors).

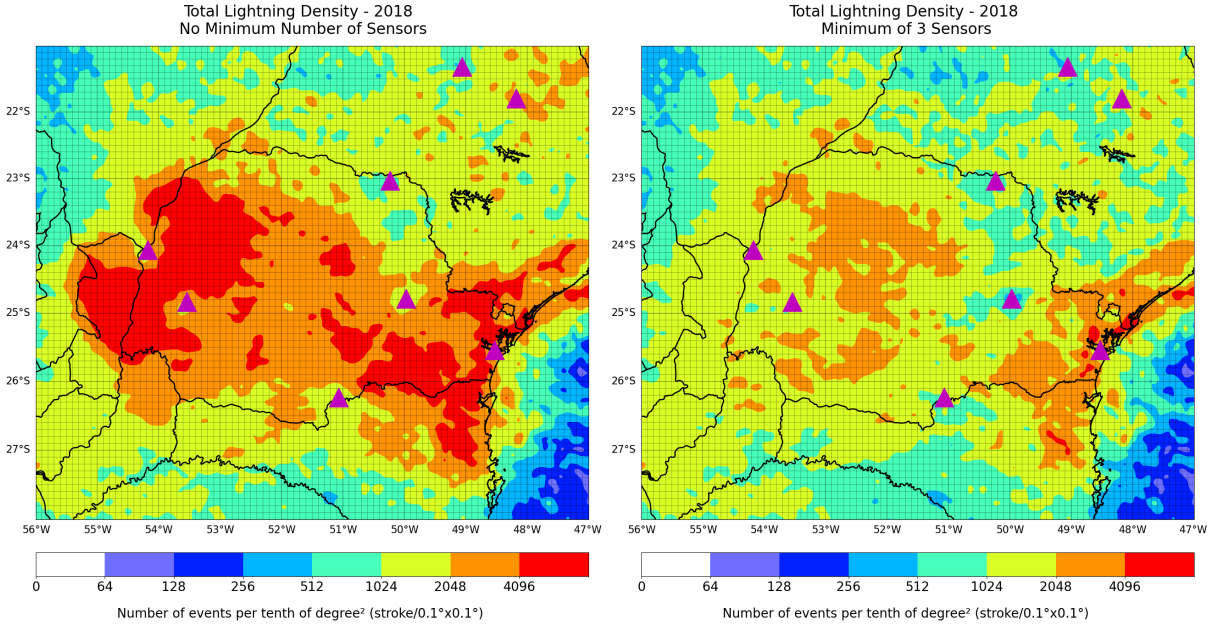

Figure 4.3: The left map presents the TL density in a one-year period considering events detected by any number of sensors. The right map presents the TL density with a minimum of three sensors detecting the event. RINDAT sensors are presented as red triangles.

The left map on Figure 4.3 shows that the area just around a sensor is deprived of lightning, showing lesser density values than the surrounding region. Lightning strikes detected by one sensor that occur in proximity to said sensor are disregarded by the central processing unit, for these events overloads the detectors and cannot be characterized.

We opted to dismiss lightning strikes detected by less than three sensors, for their location has a significant margin of error, usually in the range of kilometers. The position of a lightning detected by one sensor will have a margin of error so great that it effectively renders this information useless. In regard to lightning detected by two sensors, the location of the event may effectually be at any point in a straight line between the two sensors.

Therefore, given that the lightning location is a foremost attribute for this study, the minimal number of sensors engaged in detecting lightning is defined as three. Three sensors can define the localization of an event given their triangulation capacity, hence defining a point

of occurrence for the lightning event with an error typically in the order of tens to hundreds of meters.

The map on the right of Figure 4.3 presents the TL density on the ROI but considers only lightning detected by three or more sensors. By establishing a minimum number of sensors, the dataset of TL is reduced by 41.6%. However, the patterns of lightning occurrence in the ROI become more in accordance with the climatological norm established in DiGangi et al. (2022). Additionally, regional variations in the LDLN capacity due to sensor density are reduced.

After the deletion of the spurious events declared by RINDAT, the dataset of TL from 2018 until 2023 was comprised of 39,227,152 discharges, with 25,153,966 IC and 14,073,186 CG events. Given that CG strikes tends to be events of greater impact and thus detect by more sensors, the difference in the number of events is reduced, with almost 40% of the dataset being CG discharges (Table 4.3). Thus, given that this work is concerned with the nowcast of CG lightning, this study's fundamental dataset comprises 14,073,186 events.

Table 4.3: Number of lightning events on the dataset.

Lightning Dataset (2018-2023)	
Intra-Cloud	50,057,460 (76%)
Cloud-to-Ground	15,815,738 (24%)
Total	65,873,198 (100%)
Cleaned Lightning Dataset (2018-2023)	
Intra-Cloud	25,153,966 (64%)
Cloud-to-Ground	14,073,186 (36%)
Total	39,227,152 (100%)

4.2 DATA CHARACTERIZATION

The labor of CG lightning nowcasting is a daily duty for meteorological services. Alerts must be issued for the beginning of CG occurrences and for the cessation of the lightning activity so interested parties can resume their lightning-sensitive activities. Although CG events displays a seasonal pattern, a CG strike may effectively happen anywhere and anytime. Table 4.4 presents the number of days with CG lightning activity on the ROI in the years 2018 until 2023.

Table 4.4: Number of days with CG lightning and respective percentage from the total number of days in the year.

Year	**Days with CG**
2018	321 (88%)
2019	327 (90%)
2020	292 (80%)
2021	305 (84%)
2022	311 (85%)
2023	310 (85%)

The year with the most days with CG lightning is 2019, in which out of the 365 days of the year, 327 had CG events. 2020 had the fewest days with CG strikes, totaling 292 days out of 366. The average percentage of days with CG events is 85% from 2018 until 2022; in other words, the vast majority of days has the presence of CG discharges as declared by RINDAT.

Given that any single CG event can be the source of disruptions such as power outages or crop fires, the endeavor of CG lightning nowcasting must be done continuously throughout the year. Thus, automatic systems such as ML models are ideally suited for this forecasting task, given their capacity to operate autonomously uninterruptedly, freeing the human forecaster for further decision-making.

Although lightning occurrences may arise from other atmospheric events, they are most associated with thunderstorms. Therefore, this atmospheric electric phenomenon is most common during the rainy season, in the summer period of our ROI (DiGangi et al., 2022). Figure 4.4 presents a heat map of the mean CG lightning density per month from 2018 until 2023. It is noticeable that January and October are the months with the most intense CG lightning activity trough the years. July, being in the middle of the dry season, is the month with the least incidence of CG events.

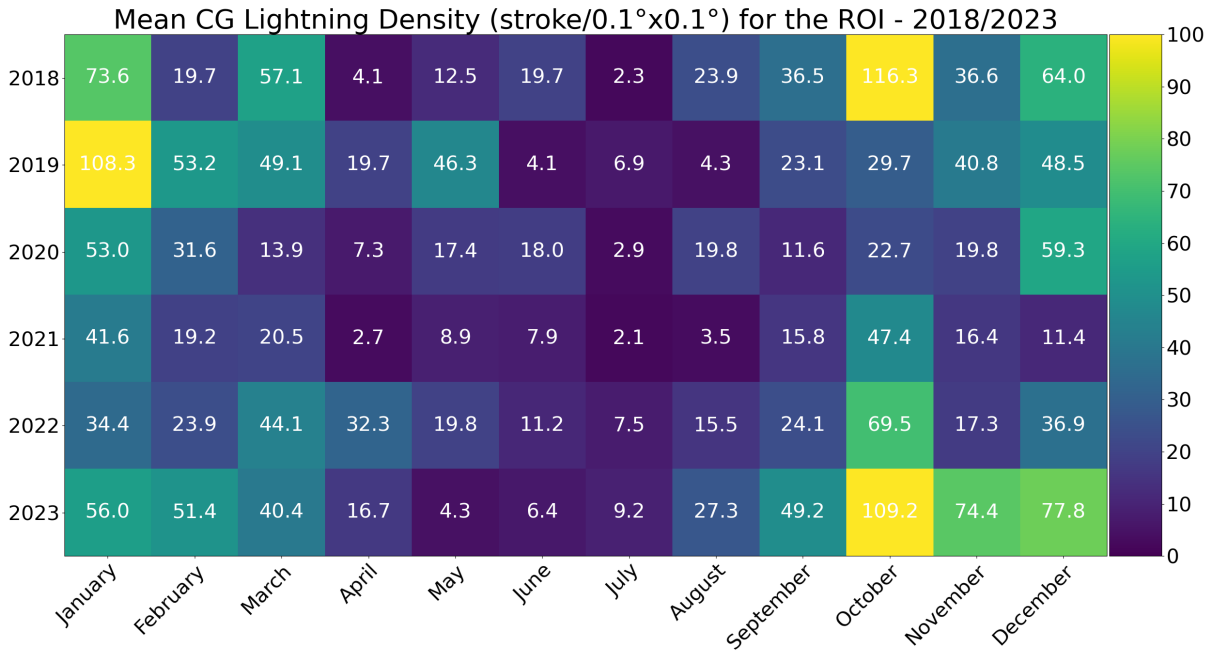

Figure 4.4: Heat map of the mean density of CG lightning per month from 2018 until 2023.

The training, validation, and test datasets are composed of CG events that took place at different times of the year, given that the occurrence of CG lightning depends on different atmospheric conditions throughout the seasons. The seasons with the most CG strikes are the summer (December, January, and February) and spring (September, October, and November) months. Winter (June, July, and August) and autumn (March, April, and May) are characterized

as the seasons with fewer numbers of CG lightning. Besides the absolute number of lightning events, the colder seasons are also characterized by lightning events of less intensity, considering their peak current.

Figure 4.5 presents the accumulated CG lightning density maps from the years of 2018 until 2023 of the summer (left map) and spring (right map) seasons. In the summer, most CG events happens on the coast of the ROI because those lightning strikes arise from a thermodynamic origin. When the air parcel from the ocean meets the mountainous region on the coast, the parcel rises to surpass the mountain in a convective manner, boosting the development of electrically charged zones in the clouds and the subsequent occurrence of CG lightning.

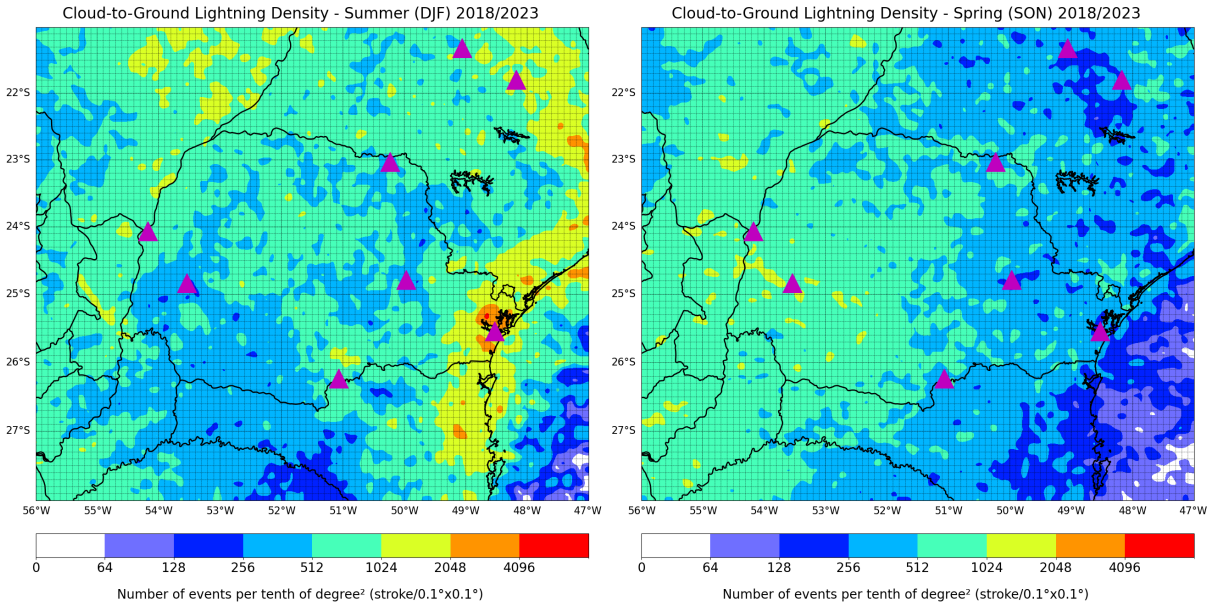

Figure 4.5: The left map displays the CG lightning density on the ROI in the summer (DJF) and the right map displays the CG lightning density during the spring months (SON).

In the spring, CG events occur predominantly in the central, southwestern region of the South American continent, associated with synoptic systems. This pattern occurs because the presence of Mesoscale Convective System (MCS) is more common in the spring. MCS define vast atmospheric phenomena comprised of a cluster of thunderstorms that usually persists for several hours, generating large precipitation values and lightning (Beneti, 2012). Given this variation on the structure of CG activity, it is of essence to consider events from throughout the year, so the ML algorithm can learn from different patterns of CG occurrence on the ROI. Thus, data from different times of the year from 2018 until 2023 composes the datasets.

Fewer numbers of lightning activity characterize the autumn and winter seasons. Figure 4.6 presents the accumulated CG lightning density in the years 2018 until 2023 on the autumn (left map) and winter (right map) periods. In contrast to the warmer seasons, CG activity is more scattered during these colder months. As a transition season from summer to winter, autumn displays greater values of CG lightning than the winter period.

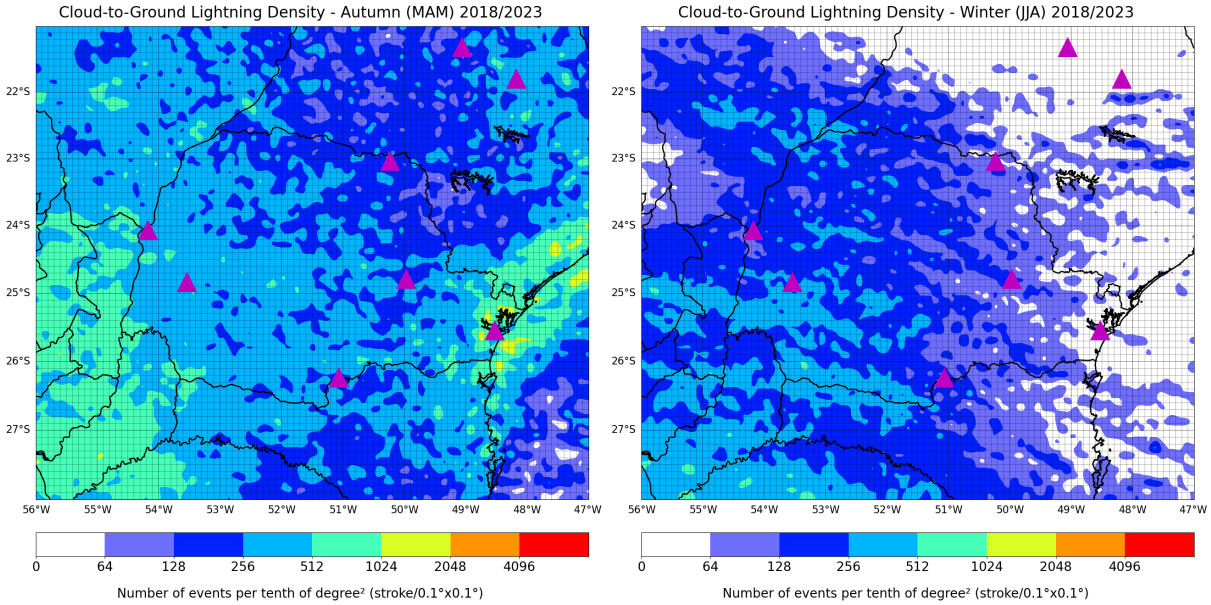

Figure 4.6: The left map displays the CG lightning density on the ROI in the autumn (MAM) and the right map displays the CG lightning density during the winter months (JJA).

Winter has the least number of CG strikes from all seasons, with most of the atmospheric electric activity concentrated on the western zone of the ROI. This pattern arises due to the lower temperatures of this season, which inhibits the thermodynamic forcing found on the coast that fosters lightning on the country's seaside. The greater values of CG discharges found in the western region arise from the onset of the occurrences of MCS that give origin to the large numbers of lightning found in the central, southwestern region of the continent during spring.

The method of CG lightning nowcasting proposed in this study is based on past lightning occurrences to predict lightning. Hence, we analysed the correlation of lightning events through time. For this analysis, we assessed the auto-correlation between events.

The auto-correlation method is a mathematical tool used to analyze how similar a series is to itself at different time lags. Auto-correlation can reveal if values tend to be similar across consecutive hours, or if they fluctuate randomly with no relation to previous readings (Wilks, 2006). The auto-correlation value ranges between -1 and 1. A value of 1 indicates perfect auto-correlation, meaning the signal is identical to itself at all lags. A value of -1 indicates perfect anti-correlation, where the signal's values at different lags are complete opposites. A value of 0 means there's no auto-correlation, implying the signal's values at different lags are unrelated.

We calculated the auto-correlation between the CG lightning density in 05 min. intervals in our ROI from 2018 until 2021 – period of the training dataset. However, only times with at least one lightning occurrence were considered, since the auto-correlation between intervals absent of lightning is irrelevant for this study. We verified a auto-correlation of 77% between events in the selected years. This indicates that future lightning occurrence is related to past lightning occurrences in a relevant manner, thus useful as a predictor element.

4.3 FROM DATA TO FORECAST: METHODS FOR A MACHINE LEARNING PREDICTION

Developing an ML model requires several design choices, including defining the type of training, the target function to be learned, a representation of that function, and an algorithm to learn the target function from the training examples (Mitchell, 1997). Specifically to the present work, the development of the CG lightning nowcasting ML model required three basic steps: (i) the selection of features and targets about lightning occurrences in the ROI and construction of datasets comprised of these data; (ii) the training and validation of ML algorithms; (iii) testing of the optimal algorithm found.

We begin by selecting data pertinent to our problem, i.e., variables useful for lightning prediction. Still, lightning appears to strike at random. Its occurrence is chaotic, governed by a myriad of factors. This highly dynamic phenomena can occur virtually anywhere in the world and during all seasons (as presented on Table 4.4). Hence, the prediction of individual events is deemed implausible (Lopez, 2016). This in turn begets the question, how can lightning be predicted?

Although the prediction of singular lightning discharges is unfeasible, forecasts for average or accumulated lightning activity are useful for a range of applications. Hence, the present work concerns the forecast of CG lightning density for a specific area in the very-short term horizon.

Density is the number of events in a circumscribed area during a designated time interval (Equation 4.1). Being so, our ROI is divided in a grid of 0.1° by 0.1° of geographical coordinates. This amounts to a resolution of a tenth of a degree squared, roughly equivalent to an area of 100 km^2. Because our ROI spans 7° of latitude and 9° of longitude, this resolution defines a grid of 6,300 elements.

$$Density = Number\ of\ Events\ per\ Time\ Interval/Area \tag{4.1}$$

For the calculation of CG lightning density for each element, the time interval was fixed at 5 min. This interval was selected considering a operational necessity. Greater accumulation times (e.g., 15 min.) would result in a slower updated model, being unable to capture the quick changes in the storm evolution. Reduced accumulation times, such as 1 min., entails a model that would not be possible monitored by a human forecaster, for it would generate too many forecasts.

And so, CG lightning density in 5 min. intervals for all 6,300 elements was obtained from 2018 until 2023. This data comprises the dataset from which was obtained the features and targets for the ML model. Figure 4.7 contains an instance of the CG lightning density on the ROI.

For the graphical representations of the model's input and output in this document, the density values were grouped in classes. Although we work with a continuous variable – the density of lighting events, which range from zero to the scale of hundreds – both as input and output of the regressor model, we decided to group these values for the observed and predicted maps to facilitate human interpretability. For operational purposes, the difference between one or

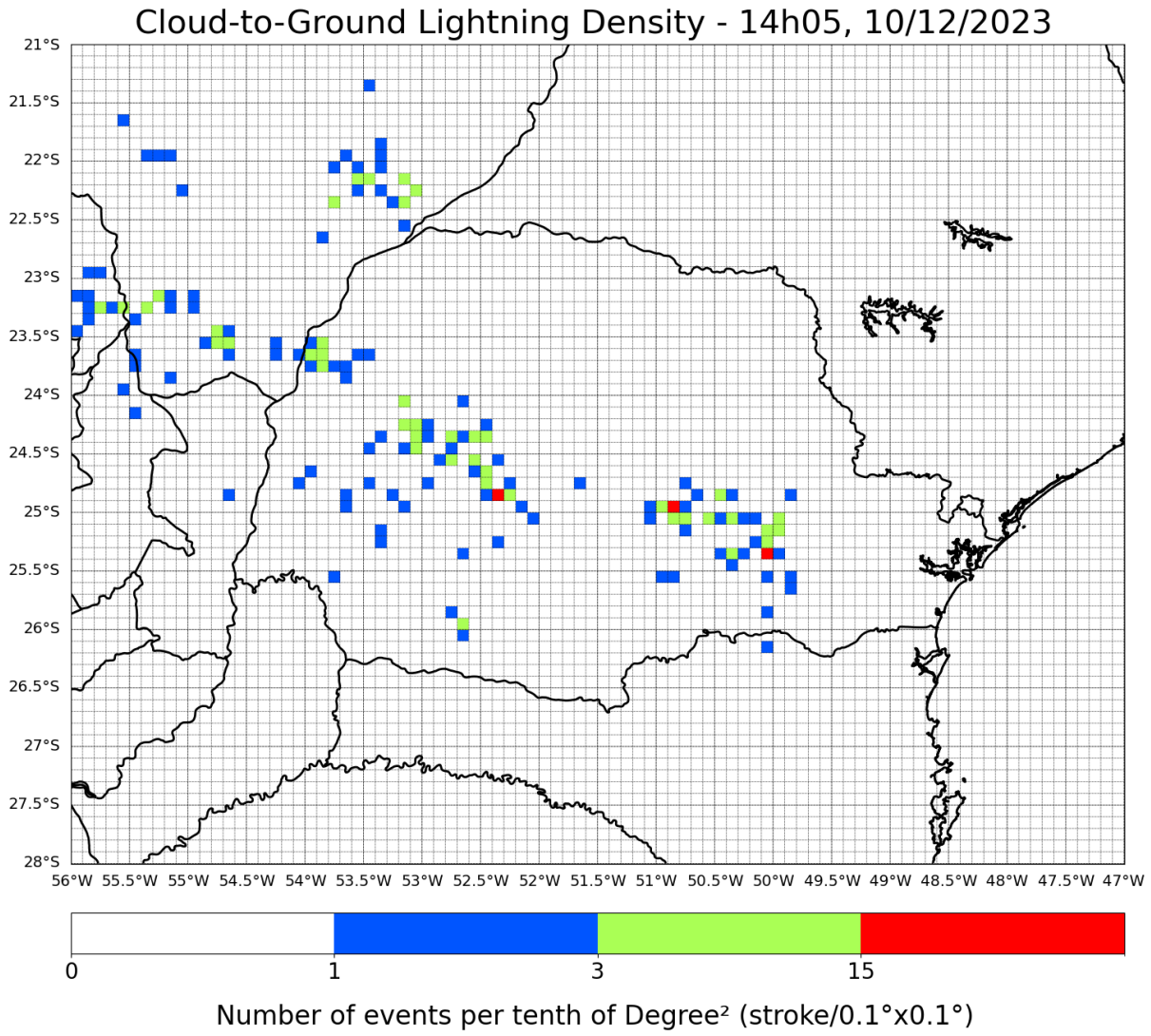

Figure 4.7: CG lightning density in our ROI in a five minute interval.

two CG strikes is negligible. It is of greater interest to present a representation of the severity of lightning occurrences in a area.

Bearing that, we decided to group the CG lightning density values in four classes:

- Class 0: 0 CG lightning (No lightning);
- Class 1: 1 to 3 CG lightning (Scarce lightning);
- Class 2: 3 to 15 CG lightning (Moderate lightning);
- Class 3: 15 or more CG lightning (Severe lightning);

These four classes are used to indicate the severity of lightning events in a area, ranging from no lightning activity (Class 0) to severe numbers of lightning (Class 3). Namely, Our developed model is a regressor, providing a continuous value as a result. For example, for element [i=54, j=67] the CG lightning density predicted by the model will be six events at 17h05 UTC. In

this case, on the graphical maps, the element in the grid will be colored green. Thus, indicating that in said area there is moderate CG lightning occurrences.

The scale on the maps in this study (i.e., the four classes) is based on the distribution of the maximum number of CG strikes in a element in a 5 min. interval from 2018 until 2022. The classes were defined based on the percentiles of this distribution, with those being: the minimum (one CG lightning); 25% (three CG lightning); and the 75% (14 CG lightning). And so, the lightning classes thresholds were defined at 0, 1, 3 and 15 CG lightning events.

The goal of the ML model is to predict the CG lightning density for each element in the grid in 5 min. intervals for up to one hour. Therefore, the target for the model is the CG lightning density value for a element at a given time, with this element hereafter denominated as **Target**. The set of targets comprises the target array.

The features for the training of the model were the past CG occurrences in the region surrounding a Target. The premise of this work is to nowcast CG lightning as an event that happens not in isolation since lightning events interact with each other and the environment but as an event that naturally evolves associated with the atmospheric conditions. The past lightning conditions surrounding an element are determinant to the future lightning occurrences in said element.

Thunderstorms display a life-cycle, and so do lightning occurrences. Scarce lightning activity is found at the onset of a storm, with greater numbers found during its peak. Few lightning events are present on the edges of a thunderstorm, and they are majorly IC since the electrified zones in the clouds are not yet strong enough to produce the more intense events of CG. Therefore, the lighting conditions (position and quantity) are a significant feature for forecasting the density of CG discharges. The ML model is expected to learn about the location and severity of past lightning occurrences and thus be able to predict the location and intensity of CG activity.

The region that surrounds a Target is defined in the the two elements next to said Target in all directions. This amounts to an array of five by five elements (25 elements in total), in which the elements presents the CG lightning density during a specific time interval. The element at the middle of the array (position i=2, j=2) is in the same location as the Target. These arrays are denominated **Neighbourhood Arrays (NAs)**. Each Target is accompanied by three NAs: i) one of the past 0 to 4 min. (t-5 min.); ii) one of the past 5 to 9 min. (t-10 min.); and iii) one of the past 10 to 14 min. (t-15 min.).

Further sizes of array and time intervals were experimented upon, but the optimal size found was arrays of 25 elements representing the past 15 min. of CG lightning activity. The results of these experiments are presented on Section 5.1. Although it may seem little past time is considered, lightning is a fast evolving phenomena. Its occurrence is dictated by the atmospheric conditions in a short-time range.

The use of both features and targets as input to train the model defines the employment of a supervised ML method in our study. The features plus the targets arrays comprise the training and validation dataset, while the testing dataset comprises just the features array.

Figure 4.8 presents the creation of one example for the training and validation dataset. The magenta boxes on the CG lightning density maps represent the region surrounding the Target (array of five by five elements). The NAs are developed from this region in various time intervals, and the Target is comprised of a single element. These NAs are then reshaped into a one-dimensional array and concatenated, with the Target appended at the end. Each example in the training and validation datasets is composed of 76 columns: the first 75 columns are the features (three NA each with 25 elements), and the last column is the Target.

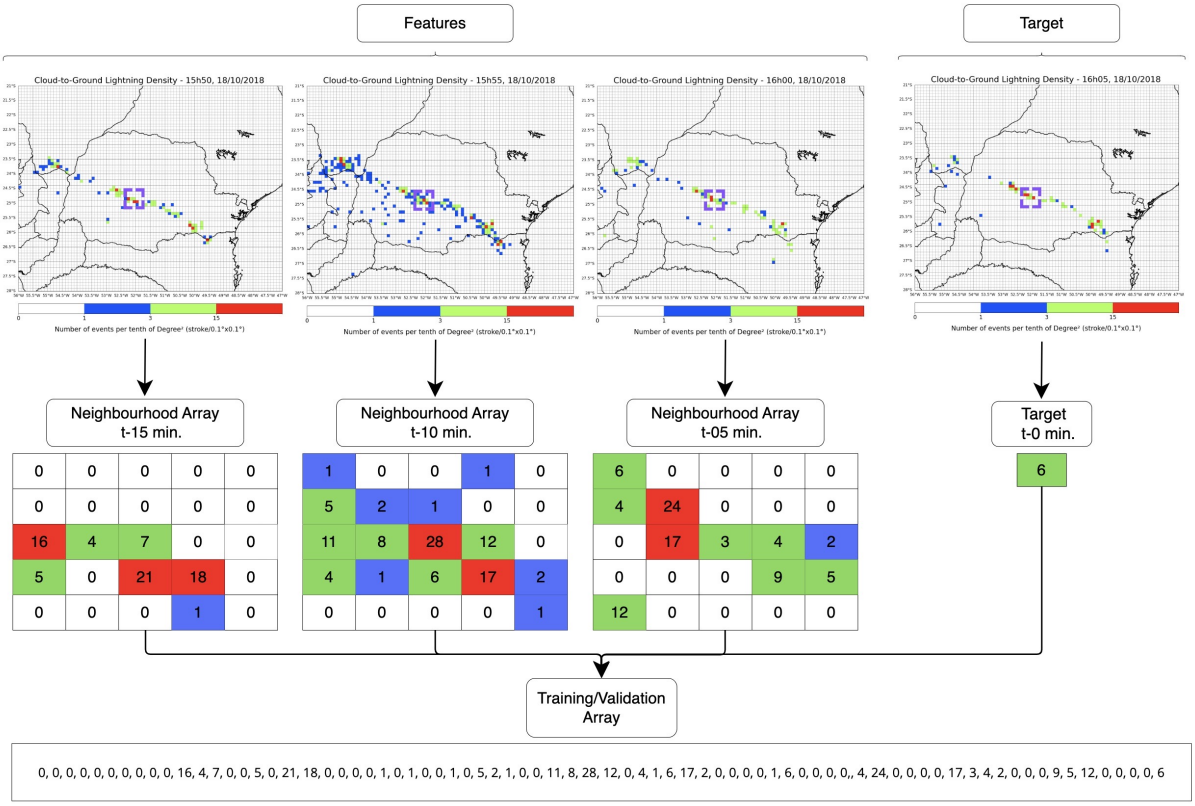

Figure 4.8: Diagram of the development of one example for the training and validation datasets.

Given that the NAs have a length of five elements, two for each side of the Target, the elements at the border of the ROI can not be analyzed by the ML algorithm for their neighborhood is non-existent. Thus, we defined that the ML model considers only the 4800 central elements of the ROI, excluding a frame of five elements.

The selection of NAs for the datasets was made based on a minimum number of elements with lightning in the array. We calculated the number of elements with CG lightning for every NA from 2018 until 2022. Thus, this distribution ranges from one up to 25, which would be if every element in the NA had a CG strike. From this distribution, we found that the median of elements with at least one CG event in a NA were two.

The developed training/validation dataset contains three types of data: i) only features with lightning, i.e., the NAs are comprised of arrays containing at least two elements with CG events and the Target is zero CG events; ii) only Target with lightning, meaning that the NAs are comprised of zeros and the Target has at minimum one CG event; and iii) both Features and

Target with lightning, expressly the NAs contains at least two elements with lightning and the Target has at least one CG strike. By doing that, we ensure that the ML model can learn about the presence and absence of CG lightning and thus can be capable of predicting both the occurrence and non-occurrence of CG discharges. Table 4.5 presents the distribution of the data.

Table 4.5: Data distribution according to the presence of lightning in the NAs or as the Target.

	Data Amount
Only Features with Lightning	20,000 (33%)
Only Target with Lightning	20,000 (33%)
Both Features and Target with Lightning	20,000 (33%)
Total	60.000 (100%)

The combined training and validation datasets contain 60,000 examples, divided into 60% for training (36,000 examples) and 40% for validation (24,000 examples). Datasets of 10,000, 20,000, 40,000 and 60,000 examples were evaluated, based on the metrics Median Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2 Score (R2).

As presented on Table 4.6, a dataset size of 10,000 examples display fine metric values, but this arises from the small validation dataset size of only 4,000 examples. Optimal values were found for a dataset of 60,000 examples, which presented performance values similar to the smaller datasets and had a significant validation dataset of 24,000 examples. Further dataset sizes did not resulted in a improvement of the metrics, as the model reaches a plateau, not benefiting from the additional training data.

Table 4.6: Variation of selected performance metrics by dataset size.

Performance Metric	**Dataset Size**			
	10,000	**20,000**	**40,000**	**60,000**
MAE	0.0037	0.0039	0.0038	0.0038
MSE	5.5e-05	6.6e-5	5.8e-05	6e-05
RMSE	0.0074	0.0081	0.0076	0.0077
R2	0.32	0.29	0.41	0.39

A test dataset was developed to accurately determine the ML model capacity for CG lightning nowcasting. The test dataset is a collection of data used to objectively evaluate a final model build based on the training and validation datasets. The test dataset used is composed of 25 events from 2023 containing only the features data. The description of the test dataset and analysis of the ML model performance is presented in Chapter 5.

The CG lightning density values were normalized to ensure the uniformity of the data, guaranteeing that the features are on a similar scale. Equation 4.2 presents the normalization method used.

$$x_{normalized} = (x - x_{minimun})/(x_{maximun} - x_{minimun}) \tag{4.2}$$

The minimum value considered is zero, i.e., no lightning in the element. For the maximum value, we obtained the distribution of the maximum number of CG strikes accumulated in a 5 min. interval per element from 2018 until 2021. The maximum value verified was 339 CG strikes. However, we decided to define the maximum value at 255, which approximately corresponds to the 99.9% percentile (252 CG strikes) and match the gray-scale values, because very high values are considered extreme events and not relevant for the training of the model.

And so, the data was prepared in a format readable as input for ML algorithms. Given that our work is a regression task, three ML regressor algorithms were selected: Linear Regression (LR), Gradient Boosting Regressor (GB), and Random Forest (RF). These algorithms were chosen based on the satisfactory performance they displayed for meteorological forecasts in the reviewed literature in McGovern et al. (2019). These algorithms were employed with their default parameters on the training dataset. The selected algorithm for application on the test set underwent fine-tuning, with its parameters presented on Section 5.1.

After the training of the models, they were applied on the validation dataset. This allowed for the verification of the model's capacity by developing performance metrics relevant to ML regressors. The selected regressor metrics were as follows: MAE; MSE; RMSE; and the R2.

The MAE (Equation 4.3) defines the mean of the absolute difference between the observed and predicted values. This metric is appropriate for our study due to its straightforward interpretation and scale compatible with the target of the estimate. Equation 4.4 presents the MSE, which describe the mean of the squared differences between the predicted and actual values, providing a measure of the discrepancy in the residual data. The RMSE is obtained by the square root of the MSE (Equation 4.5), used to synthesize forecast errors into a single measure. R^2 (Equation 4.6) provides a measure of how well observed outcomes are replicated by the model, given the proportion of total variation of outcomes explained by the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{4.3}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \tag{4.4}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \tag{4.5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2} \tag{4.6}$$

The predictions made from the test dataset were grouped into classes. This was done to enable the verification of the True Positives (TPs), True Negatives (TNs), False Positives (FPs),

and the False Negatives (FNs) values of the predictions. This allowed for the construction of a contingency table, as exemplified on Table 4.7.

Table 4.7: Contingency table.

	Prediciton Positive	**Prediciton Negative**
Observation Positive	TP	FN
Observation Negative	FP	TN

A TP indicates that the model correctly predicted that there were CG events in a element from a given class. A TN denotes that the model rightly predicted that there was not a given class in a element. Both the TP and TN indicate a correct response from the model. A FP is found when the model accused a element of having a particular class, but such a class is absent on the observed data. Lastly, a FN indicates that the model failed to predict the non-occurrence of class for a element. Therefore, the FPs and FNs denote the model's mistakes.

The verification of the aforementioned metrics enabled the development of classification metrics, allowing for a deeper analysis of the model's results. The selected subsequent metrics are: **Classification Error**, **Precision**, **Recall**, **Specificity**, **f1-Score** and the **Critical Success Index (CSI)**. These classification metrics and the other previously mentioned metrics values are presented and discussed in Chapter 5.

The **Classification Error** indicates the fraction of wrong predictions from the total. To verify the number of positive predictions that were factually correct, the **Precision** of each class was obtained. **Recall** presents the number of TP that were successfully identified. **Specificity** defines the fraction of the TN that were correctly predicted as negative and is also known as the True Negative Rate. The harmonic mean between the precision and recall is defined as the **f1-Score**. The **CSI** is used to measure the ratio of correctly predicted events to the total number of events, including both hits and false positives, providing a comprehensive assessment of the model's predictive capability.

While these classification metrics provide valuable insights into the model's performance by assessing individual grid points, they fail to account for the spatial or contextual relationships between neighboring data points. This limitation can be problematic in scenarios where the context or proximity of points carries significant importance, such as in our study. To address this issue, we elected a further metric to evaluate the ML model, namely the **Fraction Skill Score (FSS)**.

The FSS offers a more holistic assessment by considering the validity of predictions within a defined neighborhood around each point, accounting for spatial uncertainty and providing a more robust evaluation (Skok and Roberts, 2016). By using the FSS, we ensure a more comprehensive and context-aware evaluation of the model's predictive capabilities.

Equation 4.7 presents the method used to calculate this metric, where P_i represents the predicted fraction of the $i-th$ land cover class and O_i represents the observed fraction of the $i-th$ land cover class. This score ranges from zero to one, with one indicating a perfect agreement between the forecast and the observations and zero indicates no skill. An FSS value above 0.5 is considered a threshold for indicating useful skill for our task of lightning prediction (Ebert et al., 2013).

$$FSS = 1 - \frac{\sum_i |P_i - O_i|}{2 \sum_i max(P_i, O_i)} \tag{4.7}$$

To calculate the FSS, the data used as input is the predicted and observed fields. Both fields are thresholded to create binary fields where grid points are marked as one if they exceed a specified threshold and zero otherwise. In our study, we defined the threshold at one lightning event. The neighbourhood is defined at two elements, maintaining a refined evaluation of the model. Larger neighbourhoods would entail a coarser assessment of the model.

4.4 SUMMARY

This chapter presented the data and methods used to develop an ML model for CG lightning nowcasting. The chapter begins by presenting the principal tool for obtaining lightning data in a ROI, the LDLN.

The LDLN used in this study is RINDAT, Brazil's national LDLN. Historical lightning data from 2000 until 2022 was collected and processed to ensure the veracity of the data and its readability by ML algorithms. This Chapter elucidates the need for comprehensive and reliable data to capture the spatiotemporal dynamics of atmospheric conditions and lightning activity. It also discusses the challenges associated with data integration and processing, such as data cleaning and normalization.

Section 4.3 delves into the feature engineering process, presenting the selection of data based on which the ML model learns about the patterns of occurrence of CG lightning in the ROI. It describes splitting the dataset into training, validation, and test datasets. Following the data setup, the chapter discusses the ML methods employed. It provides an overview of the training and evaluation process of the ML models.

Chapter 4 highlights the importance of high-quality data collection, processing, and feature engineering in capturing the complexity of lightning occurrences. The next Chapter presents and discusses the ML models results.

5 ANALYSES OF THE RESULTS

This Chapter presents the results of a applying Machine Learning (ML) method for Cloud-to-Ground Lightning (CG) nowcasting. It introduces the findings and analyses derived from the experiments conducted, examining the effectiveness of ML models in predicting lightning occurrences in the very-short term.

Section 5.1 discusses the experiments with the selected ML algorithms in the task of CG lightning nowcasting. The models are analyzed with the use of typical metrics for the study of ML regressors, namely: Median Absolute Error (MAE); Mean Squared Error (MSE); Root Mean Squared Error (RMSE); and R2 Score (R2). Based on these metrics, an algorithm was chosen for further analysis with the secondary metrics classification error, precision, recall, specificity, f1-Score, Critical Success Index (CSI) and the Fraction Skill Score (FSS).

The Chapter then follows by presenting specific case studies in Section 5.2. Considering specific cases allows a human forecaster to interpret the model's capacity easily. Interpreting these results enables more in-depth scrutiny of the model's success and mistakes.

The Chapter concludes by presenting an analysis of the baseline models. These models were developed intent on providing a benchmark for the ML method developed in our study. In the Chapter's summary, the key findings are resumed.

5.1 EVALUATION OF THE CLOUD-TO-GROUND LIGHTNING NOWCASTING MODEL

Primarily, the Linear Regression (LR), Gradient Boosting Regressor (GB), and Random Forest (RF) algorithms were trained with the training dataset (36,000 examples). Subsequently, the trained models were applied on the validation dataset (24,000 examples). The performance of the models on the validation dataset was measured through the MAE, MSE, RMSE, and R2 metrics.

Table 5.1 presents the value of these metrics pertaining to each model, with a measurement unit of CG lightning per element. We chose to present the results for the prediction of the next five min., for this lead time presented the foremost results. Highlighting these results allows for a focused discussion on the model's performance, ensuring that readers can quickly grasp its practical implications.

Table 5.1: ML models performance metrics on the validation dataset for a lead time of five min.

	Linear Regression	**Gradient Boosting Regressor**	**Random Forest**
MAE	0.0038	0.0038	0.0037
MSE	6.1e-05	6e-05	6.4e-05
RMSE	0.0078	0.0077	0.008
R2	0.38	0.39	0.35

From Table 5.1, it is noticeable that the models displayed a similar performance for the defined regression task. Being so, we selected GB as the model for further analysis based on the algorithm characteristics.

GB is an ensemble algorithm based on the combination of multiple weak learners, namely decision trees, to build a strong predictive model (Friedman, 2001). Decision trees work by recursively splitting the data into subsets based on feature values, creating a tree-like structure of decisions. Each internal node represents a decision on a feature, and each leaf node represents an outcome.

Figure 5.1 denotes a diagram of the GB model's process. The process works iteratively, where each new tree aims to correct the errors made by the previous tress. Initially, a simple model is trained, and the residuals (the differences between the actual and predicted values) are calculated. A new model is then trained to predict these residuals. This process continues, with each subsequent model trying to correct the errors of the combined ensemble of previous models.

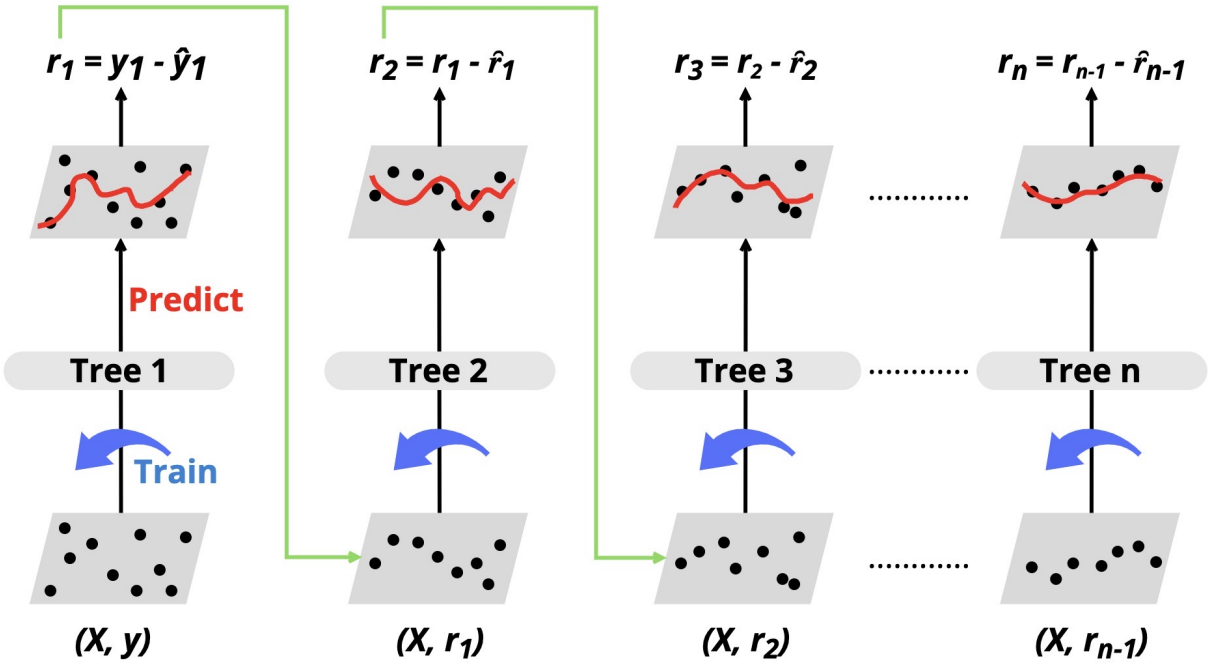

Figure 5.1: GB model construction process.

The GB algorithm excels in capturing complex, non-linear relationships in the data due to its ensemble of weak learners. By incorporating regularization techniques (like learning rate and subsampling) the algorithm iteratively corrects errors of the ensemble. GB also provides multiple parameters for selection, allowing for greater fine-tuning. Furthermore, a key feature of the GB method is the capacity to provide feature importance for model interpretability, which is later presented in this document (Natekin and Knoll, 2013).

After a fine-tuning process, the GB model used in this study comprised of an ensemble of 500 trees, being the number of boosting stages to be made. The minimum number of samples

to create a leaf was defined at five. Each tree had a maximum of four nodes, which defines the maximum depth of the trees. If this parameter is set to none, the nodes are expanded until all leaves are pure or until all leaves contain less than the minimum number samples. The learning rate defines the reducing of the contribution of each tree, set at 0.01 for our study. The loss function selected was the squared error.

For the definition of the features, two parameters were considered: the size of the Neighbourhood Array (NA), defined at 3x3 (9 elements), 5x5 (25 elements) and 9x9 (81 elements); and the past time interval considered, with those being the past 15 min. (three NAs), the past 30 min. (six NAs), and the past 60 min. (twelve NAs). Table 5.2 contains the MAE for each combination of NA size and time interval considered.

Table 5.2: Variation of the MAE by type of feature.

MAE (number of CG lightning per element)		**Size of the NA**		
		3x3	**5x5**	**9x9**
Past time considered	**15 min.**	0.0038	0.0038	0.0035
	30 min.	0.0037	0.0035	0.0035
	60 min.	0.0036	0.0035	Unable to generate

Given the defined restriction that every NA must contain at least two elements with CG events, we were unable to develop a dataset of 60,000 examples with a NA of 81 elements of the past 60 min. Although the MAE is reduced as the the size and number of NAs increased, this difference is minute, at the fourth decimal place.

Considering that no meaningful difference was verified in regards to the MAE metric, a statistics test was applied to assess the models. The chosen test was a Wilcoxon signed-rank Test. This test compares two sets of performance metrics obtained from a paired scenario. This could involve, for example, evaluating the performance difference between two machine learning models on the same dataset before and after applying a specific tuning technique. The Wilcoxon signed-rank test provides a valuable non-parametric alternative for ML practitioners when assessing the significance of differences in model performance.

It's established that for values greater than 0.5 for the Wilcoxon test defines a significant difference between the model's performance. For every size of NA and past times considered, the Wilcoxon test returned a value lower than 0.5. Hence, no discernible difference was verified between the models considering the variation in the area and past time considered.

Due to the models equivalent performance, some hypotheses were drawn up to describe this behavior. The first hypotheses considered that the training data contained few examples of severe thunderstorms, and thus limited data of fast moving lightning occurrences. Hence, the increase in the area and past time considered would not affect the model's performance. This hypotheses was refuted for the training dataset contained data from the months of January and October, which are the months with most of the lightning activity in the year, as presented on Section 4.2.

A second hypotheses was that the training dataset was mostly comprised of lightning events arising from local phenomenons, and not severe thunderstorms that span hundreds of kilometers. This hypotheses was also rebutted for the reason cited on the previous paragraph. We verified that the training dataset contained data about severe thunderstorms that occurred throughout the year, especially in the month of October. In October, it is verified the presence of Mesoscale Convective System (MCS) – which are cluster of thunderstorms spanning vast swaths of territory, as mentioned in Section 4.2.

Our third developed hypotheses was that fast moving thunderstorms do not generate lightning. We verified that this is not in accordance to the contemporary literature. A thunderstorm develops rapidly due to the intense updrafts within cumulonimbus clouds, which are driven by the unstable atmospheric conditions. The strong updrafts and downdrafts within these clouds facilitate the separation of electrical charges, leading to the generation of lightning (Hayward et al., 2020).

Conclusively, the elected hypotheses was that the model attributed importance to the recent lightning activity spatially near the Target. To verify this, we examined the models' feature importance. Our selected model is a GB, which is a ensemble of 500 trees. We counted the number of times each feature (i.e., a element of each NA) appeared in the trees. Figure 5.2 presents the distribution of the number of occurrences for each feature's of the GB model trained on NA of 81 elements and for the past 30 min.

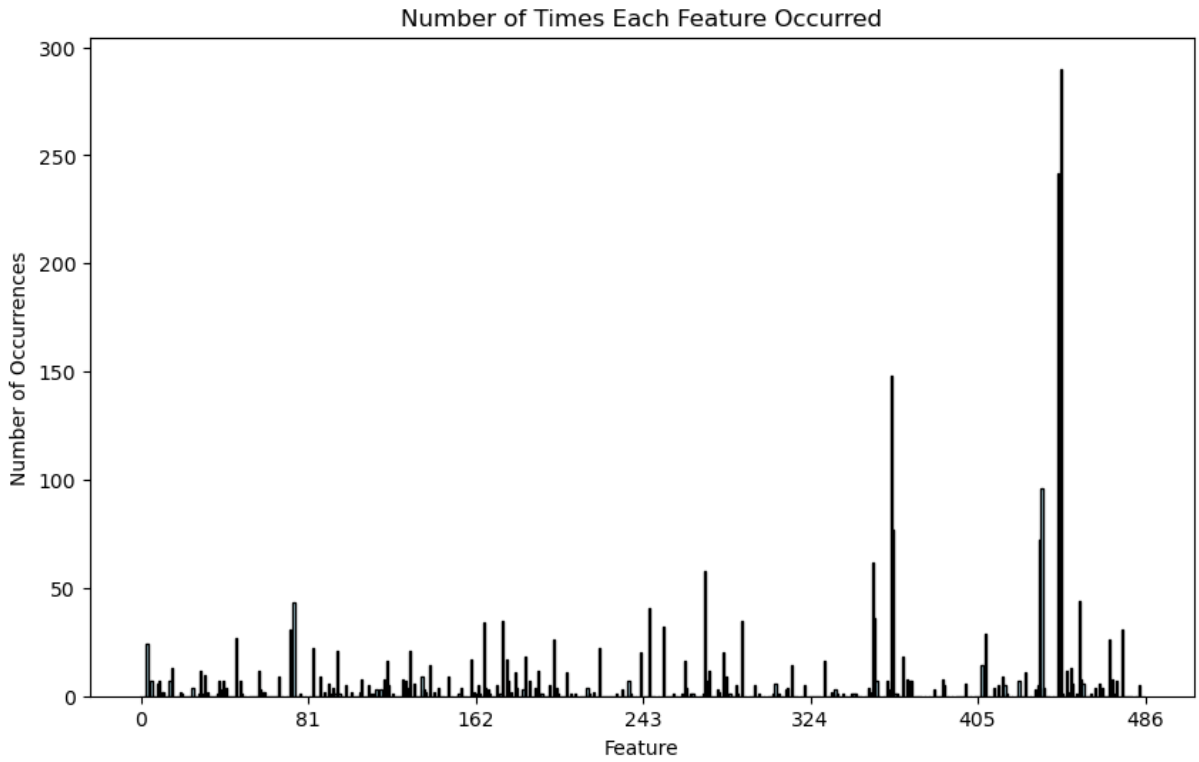

Figure 5.2: Count of occurrences of each feature for the GB model trained with NA of 81 elements of the past 30 min.

The predominant features are present on the NAs of the last 5 min. (elements 405 until 486) and of the past 6 to 10 min. (elements 324 until 405) in an area of 5x5 elements. The model attributed the most importance to the element which is in the same position as the Target. The second most important feature is the element east adjacent to the Target, and the third most important is the element north adjacent to the Target. Thus, the ML algorithm in fact considers the elements closest to the Target as the significant features to develop its prediction. As a consequence of this characteristic, the increase in area and past time does not lead to a increase in the model's performance.

This behaviour is in accordance to the atmospheric physics involved in lightning occurrence. Characterized by rapid and significant changes, lightning is a phenomenon whose occurrence is heavily influenced by the prevailing atmospheric conditions within a localized area.

Being so, the features were selected based on physical reasons. We elected to use NA of the past 15 min. for thunderstorms are fast-moving weather phenomena. The atmospheric conditions of the past 60 min. may not be a relevant input for the model, for they may be associated with a different thunderstorm cell.

Given that each element defines an area of approximately 100 km^2, we selected as feature a NA of 25 elements. A bigger NA of 81 elements would have most elements with a density value of zero CG events most of the time, except for extreme thunderstorms. A smaller NA of nine elements considers only the area immediately adjacent to the Target, and from the verified feature importance, the atmospheric electric conditions surrounding a greater area around the Target are relevant.

Therefore, the selected algorithm for further evaluation was the GB model trained on NAs of 25 elements of the last 15 min. Figure 5.3 presents the variation of the selected regressor metrics per lead time. The values of the metrics were normalized in the range of zero to one using the min-max method so they could all be presented in a single chart. It is noticeable that as the lead time increases, the model's performance decreases.

And so, the GB model was applied on the test set. The test dataset contains 25 events from 2023, amounting to 120,000 examples to comprise the dataset for each lead time. These 25 events are from the three days with the most CG lightning activity in 2023, and are distributed as follows: 13 events from 10/04/2023, six events from 12/10/2023, and six events from 12/23/2023.

These are events which the model had no prior contact, and so can be used to provide a realistic measurement of the ML model capacity. The test dataset is the biggest of all three datasets (training, validation, and test) to guarantee a trustworthy measure of the model's performance. Figure 5.8 at the end of this Chapter presents a diagram of the generation of a prediction for one Target from the test dataset.

In view of the fact that the predicted values from the regressor algorithm were subsequently grouped in four classes, the development of a confusion matrix is possible. The confusion matrix on Table 5.3 presents the model's output associated with the observed values for a lead time of five min. The classes on the columns refer to the predictions, and the rows refer to the

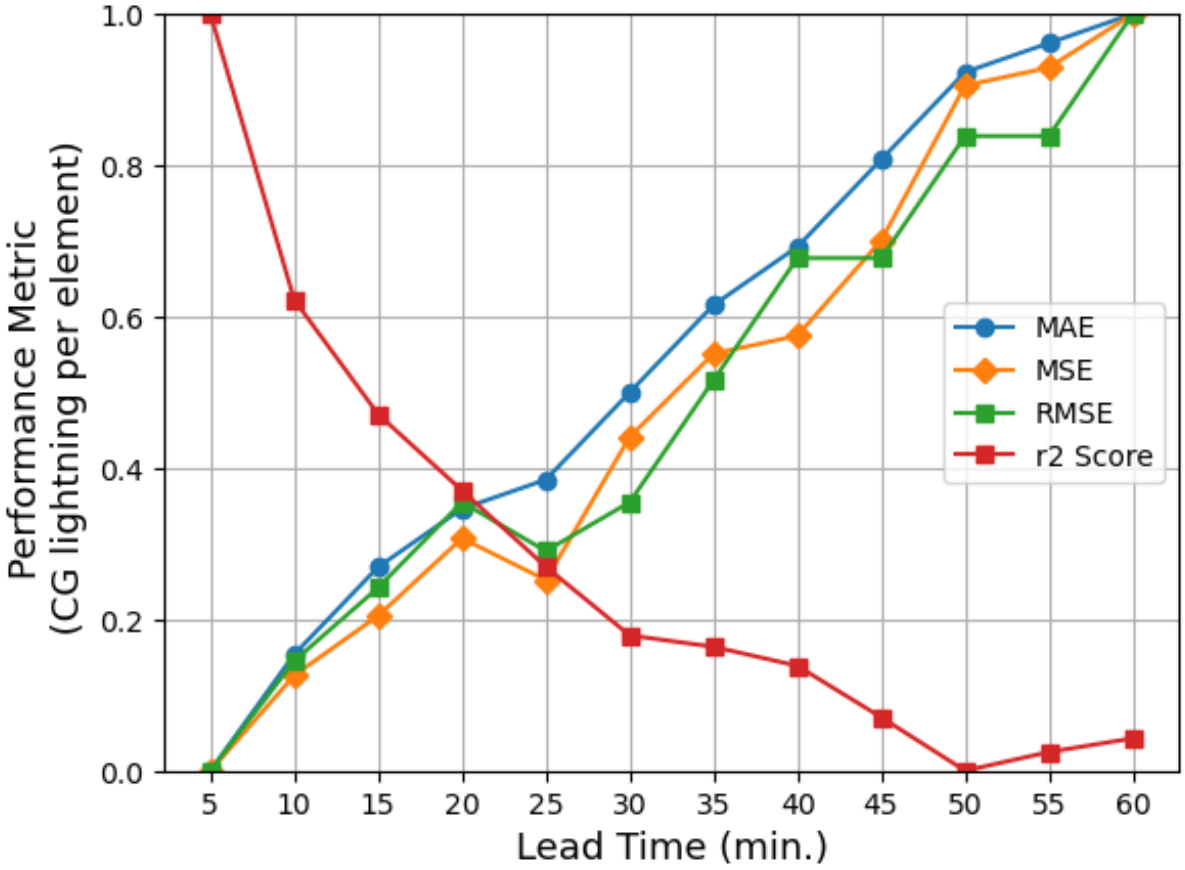

Figure 5.3: Variation of the GB model's performance according to regressor metrics per lead time.

observed classes. Ideally, there would only be values on the diagonal line of this table, indicating a perfect performance of the model.

On the test dataset, the model correctly predicted the CG lightning density value on 88,172 Targets, wrongly predicting the values of the remaining 31,828 Targets. For example, out of the total 86,511 Targets observed as Class 0, the model correctly predicted 86,183 cases and wrongly predicted 315 Targets as Class 1 and 13 Targets as Class 2 but did not predict any Target as Class 3.

Table 5.3: Confusion matrix of the test dataset for a lead time of five min.

Class	**0**	**1**	**2**	**3**
0	86,183	315	13	0
1	6,289	1,490	272	24
2	575	397	327	43
3	7	10	18	37

By analyzing Table 5.3, we note that the majority of Targets are from Class 0, and each subsequent class has fewer events on the test dataset. This is because the test dataset comprises 25 cases containing every Target of the Region of Interest (ROI) for a specific time interval. As mentioned on Section 4.3, our ROI encompasses 6,300 elements. Excluding the five elements

frame, there are 4,800 elements left in the ROI. Thus, given that 25 cases were selected, the test dataset contains 120,000 Targets. This procedure was done to ensure that the case studies could be obtained from the test set. The case studies, introduced in Section 5.2, present the model's forecast for the entirety of the ROI instead of individual Targets.

Although confusion matrix and contingency table are usually interchangeable terms, in our work, the confusion matrix refers to Table 5.3, which presents the observed and predicted values, and Table 5.4 presents the contingency table, a derived measure of the confusion matrix. A contingency table summarizes the relationship between several categorical variables, denoting the success and mistakes of the model concerning each class. The metrics on a contingency table are the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values for each class.

Table 5.4: Contingency table of the test dataset for a lead time of five min.

Class	**TP**	**TN**	**FP**	**FN**
0	86,183	2,618	328	6,871
1	1,490	87,203	6,585	722
2	327	94,355	1,015	303
3	37	95,861	35	67

By examining Table 5.4, we found that the values of the TN column indicate the model's capacity for not overestimating the values on a given Target. In contrast to that, the high FP values, particularly concerning Class 1, indicate that the model predicted more Targets with CG lightning than in the observed data.

Subsequently, classification metrics were derived from the TP, TN, FP, and FN for each of the twelve lead times. The selected classification metrics were the CSI, Classification Error, Precision, Recall, Specificity, and f1-Score, and the charts denoting these values are presented at the end of this Chapter.

The first metric analysed is the **CSI**, presented on Figure 5.9. In regards to the CSI metric, the model shows a significant disparity in performance across the four classes. The model achieves an CSI of 92% for the first class for the first lead time, indicating a high number of successful predictions for this category. However, the performance drops sharply for the other classes and for further lead times.

Classes 0 and 1 presented the greater **Classification Error** values, associated with the fact that most data in the test dataset pertains to these two classes. Classes 2 and 3 had similar low values for this metric, denoting that the model made a correct prediction in most cases. This metric is denoted on Figure 5.10.

Considering the **Precision** (Figure 5.11), the model excels at predicting Class 0 Targets, but struggles with the other classes, displaying low precision values. This indicates that the model can satisfactorily predict the non-occurrence of CG lightning, a fundamental feature

for nowcasting the cessation of lightning activity in an area. Interestingly, the second highest precision is found for the Class 3, which is related to severe lightning occurrence.

The highest **Recall** values were found for Class 0 (Figure 5.12). Classes 1 and 2 presented a similar recall value, identifying more than half of the events pertaining to those classes for almost every lead time. The recall and precision metrics should be jointly analyzed, given that the improvement of one measure may reduce the other.

The model had a satisfactory performance in regard to the measured **Specificity** (Figure 5.13), with specificity values above 80% for every class except Class 1, reinforcing the model's capacity to predict the non-occurrence of a given class. Class 0 had the highest **f1-Score**, given the model's high capacity of predicting this class. Due to the low recall values associated with the other classes, the f1-Score is significantly lower for Classes 1, 2, and 3, as presented on Figure 5.14.

The metrics presented so far assess the model's performance on a point-by-point basis, which is limiting when dealing with the spatially distributed CG lightning events. By considering the **FSS**, we are able to measure the agreement between predicted and observed events within a specified neighborhood, thus accounting for spatial proximity and distribution. This approach allows for a more realistic assessment of the model's performance, capturing not only the quality of individual predictions but also the overall pattern and coherence of the forecasts.

To asses the model by the FSS, we defined a neighbourhood size of two elements. In this way, the model can be evaluated accounting for the spatial distribution of elements while still maintaining a fine-grained view of the model's performance. Figure 5.4 contains this score for each lead time. The model displays an FSS score above 0.5 for a lead time of up to 25 min, with a maximum verified at five min.

Despite the results achieved with our ML model in CG lightning forecasting, a limitation remains: the model's inability to predict the initiation of the first lightning strike in a thunderstorm. The model relies on data from the initial lightning event to begin its forecasting process, meaning it is able to predict lightning after the occurrence of the first lightning event. Addressing this shortfall will require further research and the integration of additional atmospheric parameters and predictive techniques that can identify the precursors of lightning initiation.

5.2 CASE STUDIES: LIGHTNING NOWCASTING FOR THE REGION OF INTEREST

This Section presents three case studies on the nowcasting of CG lightning for the defined ROI by the ML GB Regressor model. Three cases were selected among the 25 that comprise the test dataset to provide an overview of the model's capacity.

Although the model outputs are made for specific grid points, we chose to interpolate the data using the quadratic interpolation method for the generation of the observed and predicted maps. Interpolation provides a more continuous and comprehensive spatial representation of the lightning activity, which is particularly valuable for visualization and analysis. This continuous

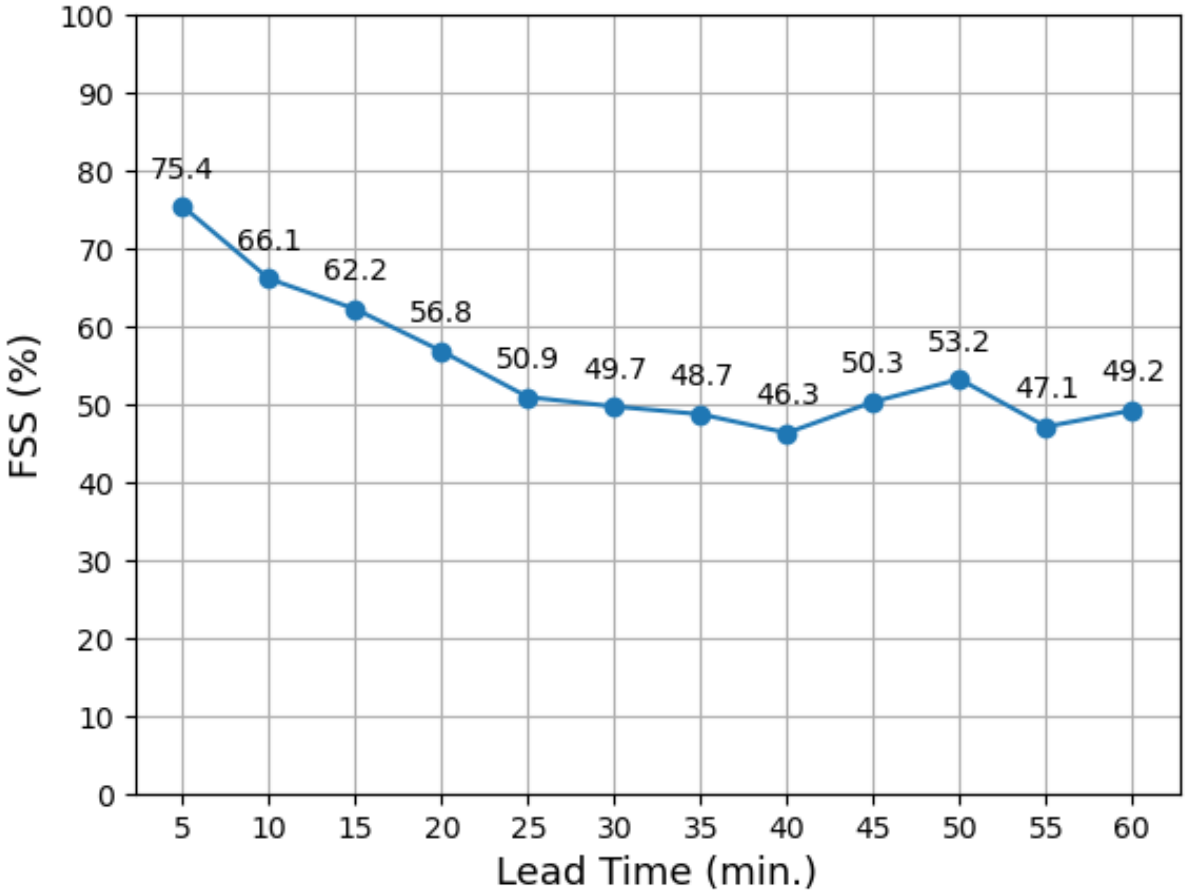

Figure 5.4: Variation of the model's performance according to the FSS per lead time.

representation helps to better understand the spatial patterns and trends of lightning events, making it easier to identify areas of high risk.

By filling in the gaps between grid points, interpolation ensures that the predictions are more aligned with the real-world spatial distribution of lightning events, thus improving the reliability of the forecast. This approach can help in reducing the potential errors associated with discrete grid-based predictions by smoothing out anomalies and providing a more comprehensive view of the forecasted CG lightning activity.

The quadratic interpolation method, which fits a parabolic curve through data points, provides a smooth and continuous surface that captures the underlying trends (Dodgson, 1997). This proves to be useful for lightning data, which can exhibits non-linear spatial patterns due to varying atmospheric conditions.

The cases selected to comprise the test dataset focused on days with significant lightning activity. These high-activity days present challenging and diverse scenarios that test the model's ability to correctly predict lightning events under varying conditions. Additionally, these case studies provide insights into the model's practical applications, as they reflect real-world situations where accurate and timely lightning forecasts are crucial for public safety, disaster preparedness, and operational decision-making. This targeted selection of case studies not only highlights the

model's effectiveness but also underscores its potential impact in mitigating lightning-related risks.

Figure 5.5 presents the **First Case**, a 5 min. interval of the density of CG strikes from a thunderstorm that passed through the ROI on the 4th of October of 2023. By examining Figure 5.5, we notice that the model accurately calculated the storm's cells shape. However, the model expanded on the size of the cells, even joining some cells together. This indicates that the model tends to predict a more homogeneous shape of CG occurrences than in the observed data.

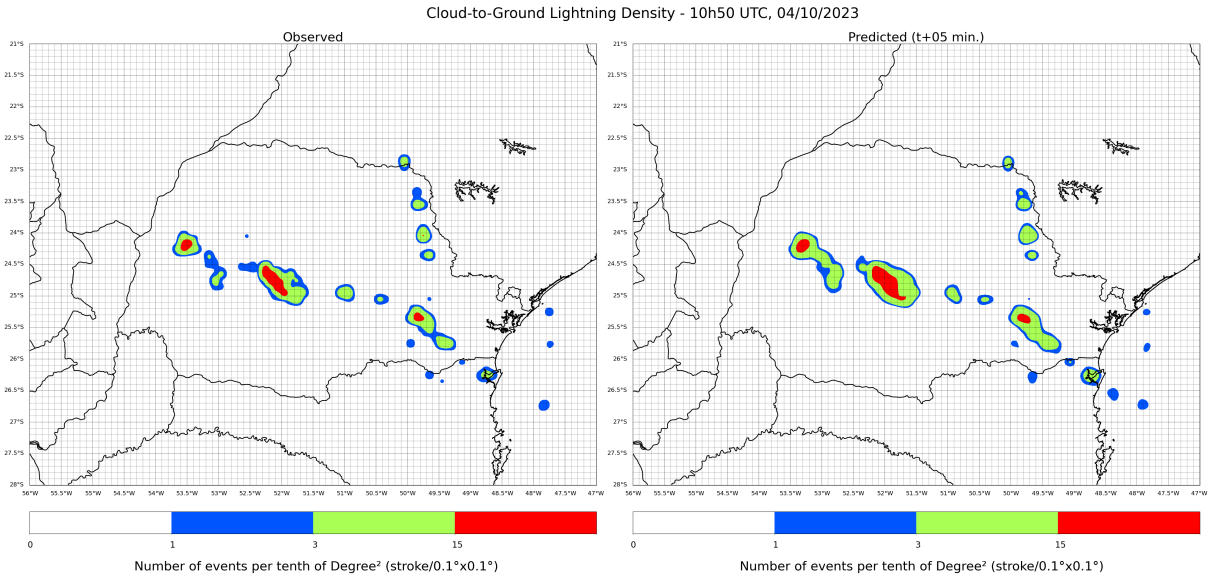

Figure 5.5: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023.

Therefore, we understand that the underlying assumption of the model is that the neighboring Target of a Target of Class 1, 2, or 3 probably also defines a element with CG strikes. This is in accordance to the expected, given that this was the assumption we made when building the training dataset. The model presented a difficulty to predict the expected gaps in lightning occurrences in the region of a storm.

The assumption that elements with lightning usually have elements with lightning next to it is reasonable for the classes with greater CG lightning values, but for Class 1 this assumption can be an issue and results in the wrong prediction of Targets of Class 0 as Class 1, resulting in a shape excessively uniform.

The overestimation of the number of elements with CG events can lead to the emission of false alarms and thus disrupt activities. This may result in financial losses and, for the alarm-issuing institution, loss of credibility. Nonetheless, this sort of overestimation that leads to the prediction of uniform shapes is not preposterous. A Lightning Detection and Location Network (LDLN) is incapable of detecting every single lightning event, and those that are captured, present a margin of error in their location. Hence, this overestimation may be seen as a built-in error margin in the model's prediction due to the inherent error in the observed data.

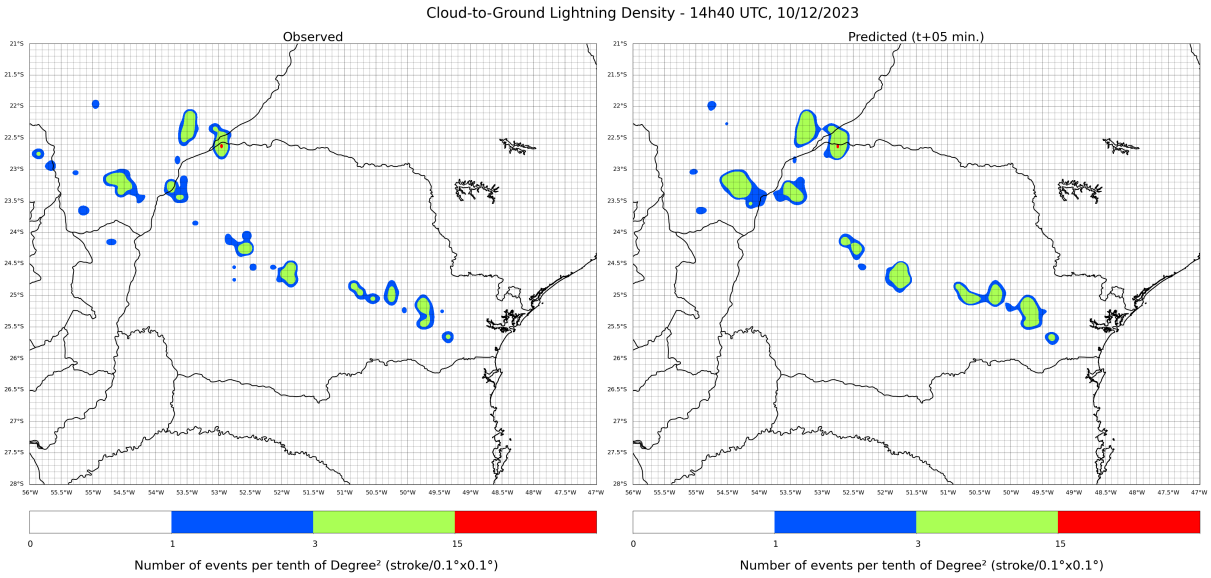

Figure 5.6: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 14h40 UTC on 12/10/2023.

The **Second Case** on Figure 5.6 pertains to the nowcast of a storm on the 10th of December of 2023 at 14h40 UTC. The model correctly predicted the location of the larger areas with CG strikes. However, some of the more isolated lightning events, spanning a single grid element, were not forecasted by the model. This fact can also be noted on the third selected case.

The **Third Case** considered is of a event with CG lightning activity spread throughout our ROI on the 23th of December of 2023 (Figure 5.7). The model successfully predicted the severity and position of the spread cases of CG lightning. However, it overestimated the severity of events in one cell, predicting events of Class 2 as Class 3.

The underestimation of a Target's class is an issue that may impede that the required attention is given to an area. Even though warnings must be issued regardless of the number of

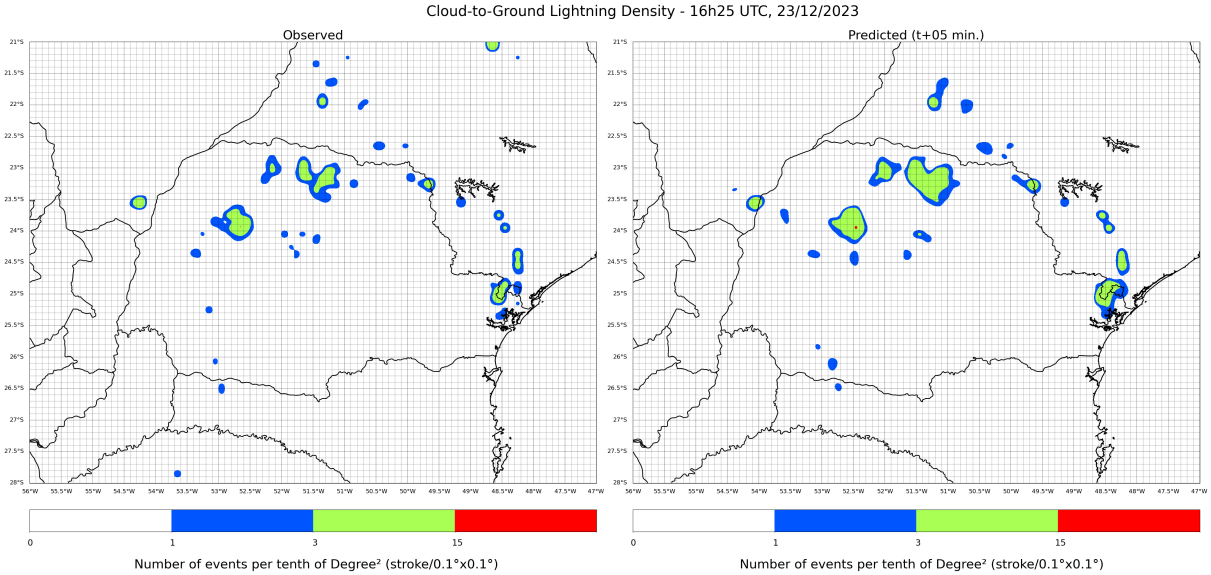

Figure 5.7: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 13h10 UTC on 28/12/2022.

lightning, severe events may require additional steps. Additionally, lightning is an indicator of the atmospheric conditions. It can provide information about the intensity and trajectory of a storm, for lightning events occur at the storm front and the number of lightning events is directly associated to the severity of the storm (Jr. and Domingues, 2002). Thus, the underestimation of the intensity of lightning occurrences entails the underestimation of the intensity of the storm.

Besides these case studies, the complete forecast of a event in presented on Chapter 7. This forecast is comprised of 12 figures, all related to the same five min. interval, denoting the difference in the model's performance as the lead time is increased.

5.3 BENCHMARKING METHODS FOR LIGHTNING NOWCASTING

ML methods requires large amounts of curated data for the machine to accurate learn and represent the problem. The construction of this dataset requires knowledge about the problem and about ML itself, to be able to translate the problem to a computer-interpretable format. Furthermore, depending on the selected method, it can be computationally expensive to employ the ML model for operational scenarios, perhaps even impeding it's use. Thus, the use of a ML method implies in advantages over more traditional methods.

To verify if our developed ML CG lightning nowcasting model surpass other established forecasting methods, we developed baseline lightning nowcasting models. These models were implemented according to conventional techniques, such as Numerical Weather Prediction (NWP), and modern methods found in the literature. In the present Section, we benchmark these models, analysing their performance in the task of lightning nowcasting.

The first method considered is termed **Persistence Forecast**. Persistence forecast is based on the idea that the current weather condition will persist and that future weather will be the same as the present (Bansal, 2022). This forecast mode is useful for nowcasting weather phenomena of medium to large extent. This method was applied for the prediction of CG lightning in our ROI for the months of January and October – the months with most lightning occurrences – of 2023.

The first iteration of this method was employed considering if there was any lightning in the NA in past times. Different past times were considered, with the best result found verifying if there was lightning in the past 5 min. If there was any lightning in any element of the NA, a prediction is made that there will be lightning in the target element. If no lightning is found, a prediction of lightning absence is made.

This persistence model displayed a high FSS of 0.67 for a lead time of 5 min. which is in accordance to the expected. This high accuracy arises from the fact that most of the elements are absent of lightning most of the time, and the persistence model correctly predicted this. However, the model has a poor performance for the prediction of lightning occurrence, presenting an F1-Score of 26%. Hence, the ML method surpass the use of persistence forecast, given its capacity for the prediction of lightning presence.

Furthermore, as the lead time of the prediction is increased, the performance of the persistence model drops sharply in regard to the FSS. Expressly, for a lead time of 10 min, the FSS is 0.53, and for a lead time of 15 min, the FSS drops below 0.5, reaching 0.42. Given these further lead times resulted in FSS below 0.5, they are defined as unuseful for CG lightning nowcasting.

The second iteration of the persistence forecast model considered the number of elements with lightning in the NA to make a lightning prediction. Three models were developed given the number of elements that had to have lightning in the past to declare that there will be lightning in the element in the next five min. The minimum number of elements with lightning in the NA selected were 25%, 50% and 75% of the total number of elements in the NA. These models displayed a equivalent performance to the first iteration model, without presenting significant gains.

Based on the method proposed in Leite et al. (2011), we developed a model for lightning prediction using **Self Organizing Maps (SOM)** (Vettigli, 2018). For the weather sciences, SOMs are useful for viewing the distribution of synoptic weather patterns over a region, inspect severe weather and rainfall events, and classify meteorological events (Skific and Francis, 2012). Our

implementation of this neural network was for a classification task, with the task being a binary CG lightning prediction (i.e., presence or absence of lightning) for the next 5 min. in our ROI.

The model was unable to capture the association between the past atmospheric electric conditions and future CG lightning occurrence with the same quality as the ML regressor algorithms employed in this study. The application of the SOM with its default hyperparameters resulted in a FSS of 0.66. Because of that, we chose to fine-tune the model, experimenting with the following parameters: number of neurons; learning rate; and neighbourhood function. The optimal parameters found were eight neurons, a learning rate of 0.5 and bubble as the neighbourhood function, which resulted in a FSS of 0.68. This performance may have arisen from the SOM's characteristic of focusing on synoptic weather events, which are large scale phenomena.

5.4 SUMMARY

In this chapter, we present a comprehensive evaluation of our ML model developed for CG lightning nowcasting. The model's aim was to predict the location of CG lightning events and the number of events per element, indicating the severity of events, as denoted by the four Classes (null, scarce, moderate, and severe). Our evaluation methodology included traditional regressor and classification metrics and advanced spatial metrics to ensure a robust assessment of the model's performance.

Firstly, the model as a regressor was evaluated. The metrics analysed denoted the proficiency of the selected algorithms for our selected task. Due to the similar performance of the models, the GB algorithm was selected for more in-depth experimentation due to its characteristics. The model was trained with various sets of features. The selected set of features comprise data about CG events of the past 15 min. in a area of 5x5 elements. This set was chosen by assessment of the feature importance and alignment with the atmospheric physics involved in lightning occurrence.

We highlight the model's capacity for predicting not only the presence or absence of CG strikes, but the number of events, thus being able to declare the severity of the atmospheric conditions in an area. By grouping the regressor values in classes, we facilitate the interpretation of the forecasts, and thus reduce the time needed for the emission of warnings. To provide a more continuous spatial representation, we applied interpolation to the grid-based predictions, resulting in smoother and more visually coherent maps that better reflect real-world lightning patterns.

Our analysis included case studies on days with significant lightning activity, selected to test the model under diverse and challenging conditions. These case studies illustrated the model's robustness and practical utility in forecasting CG lightning strikes, emphasizing its potential applications in enhancing public safety and disaster preparedness. Furthermore, baseline models were build based on established lightning forecasting methods. These baseline models provided

satisfactory results for a lead time of five min., but they lack capacity for extended lead times, in contrast to our model, which is able to predict lightning for up to one hour.

In this Chapter, the efficacy of our ML model in predicting both the location and intensity of lightning events is verified. The combination of traditional and advanced evaluation metrics, along with detailed case studies, provides a comprehensive validation of the model's performance, underscoring its value in operational lightning forecasting. The conclusion of this study is given in the next Chapter of this dissertation.

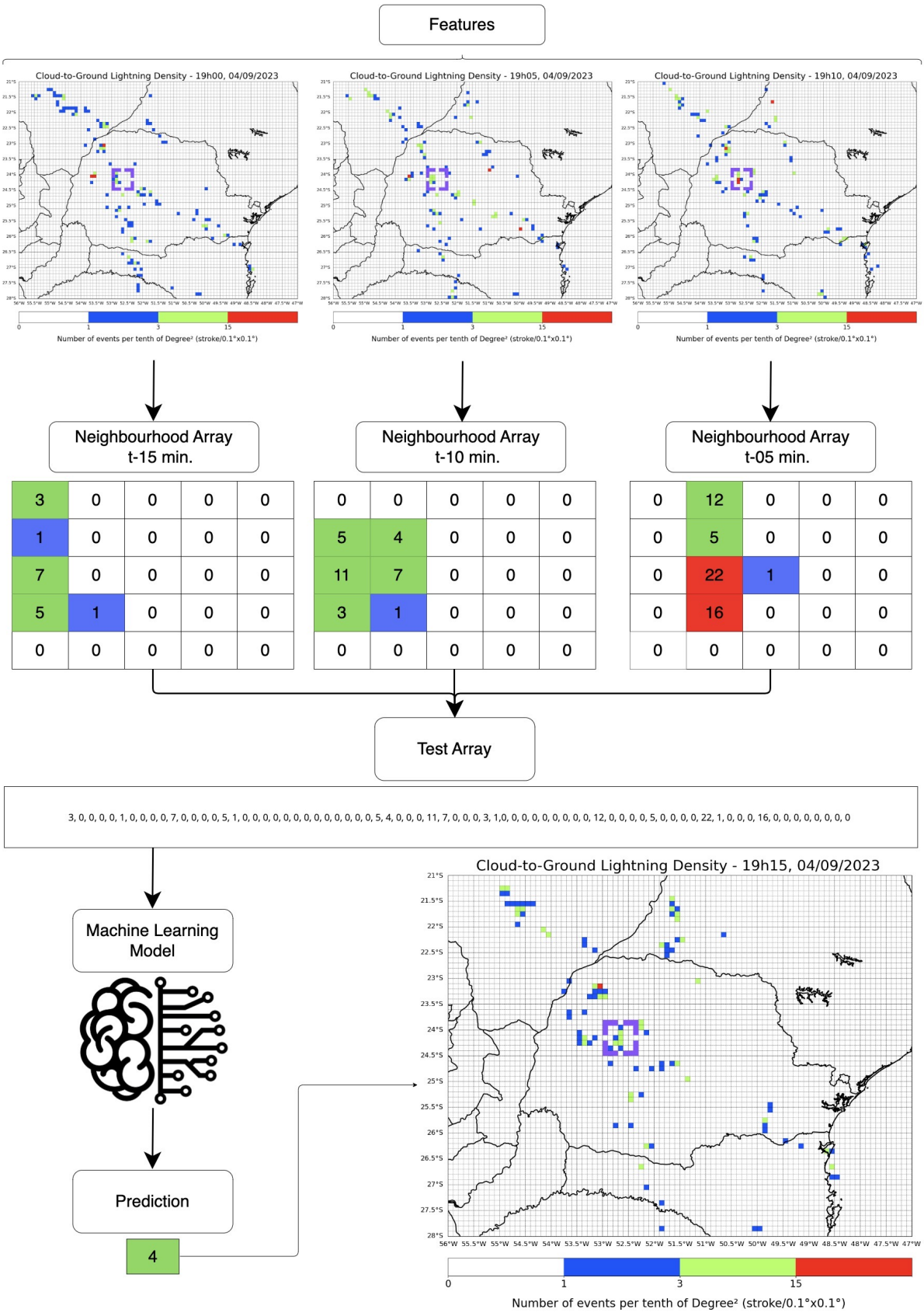

Figure 5.8: Diagram of the development of the forecast for one Target from the test dataset.

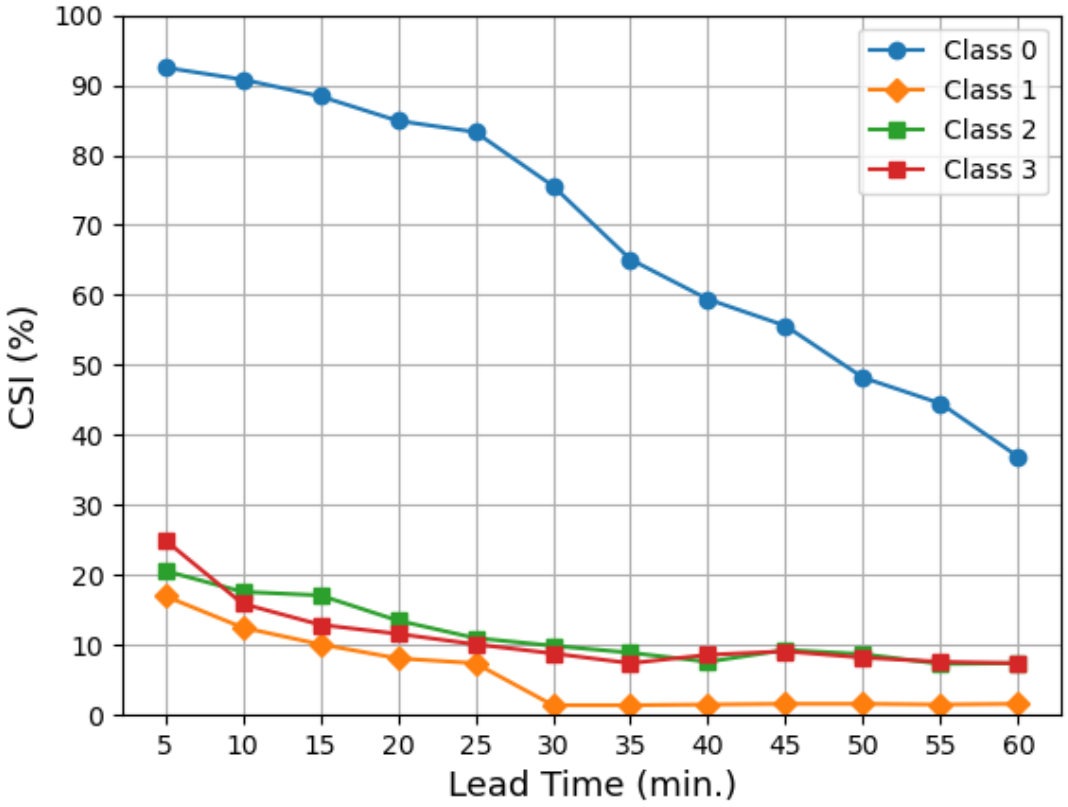

Figure 5.9: Variation of the model's performance according to the CSI per lead time.

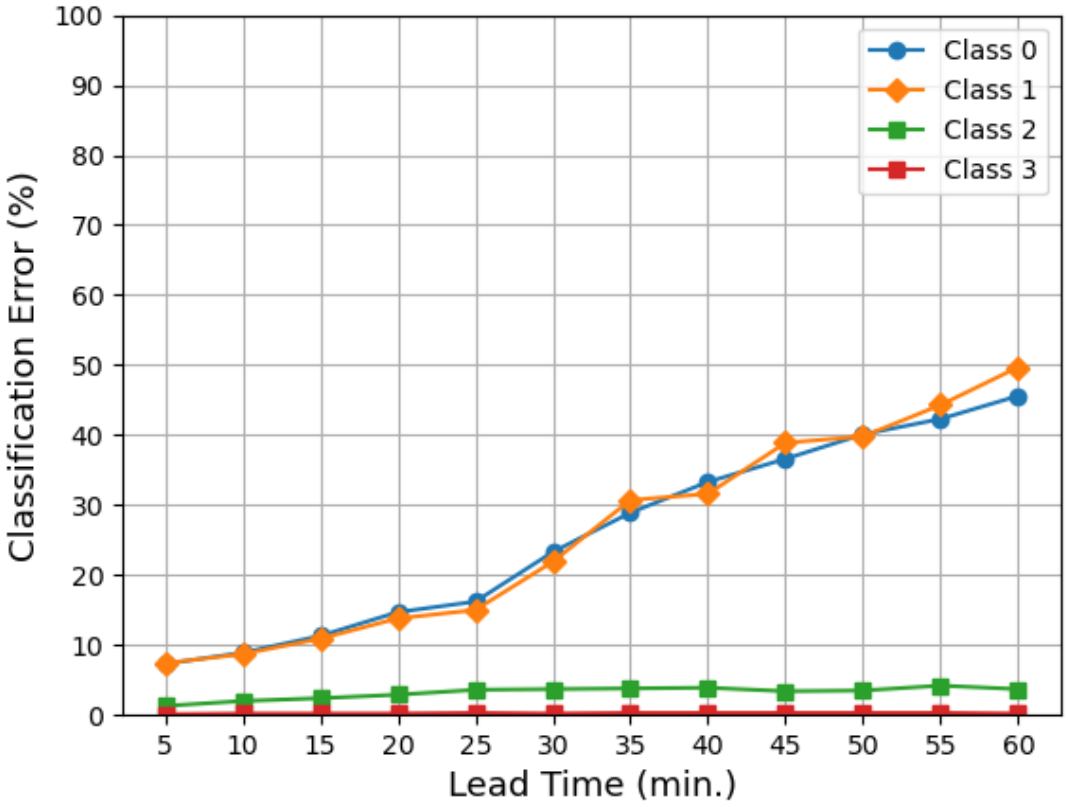

Figure 5.10: Variation of the model's performance according to the classification error per lead time.

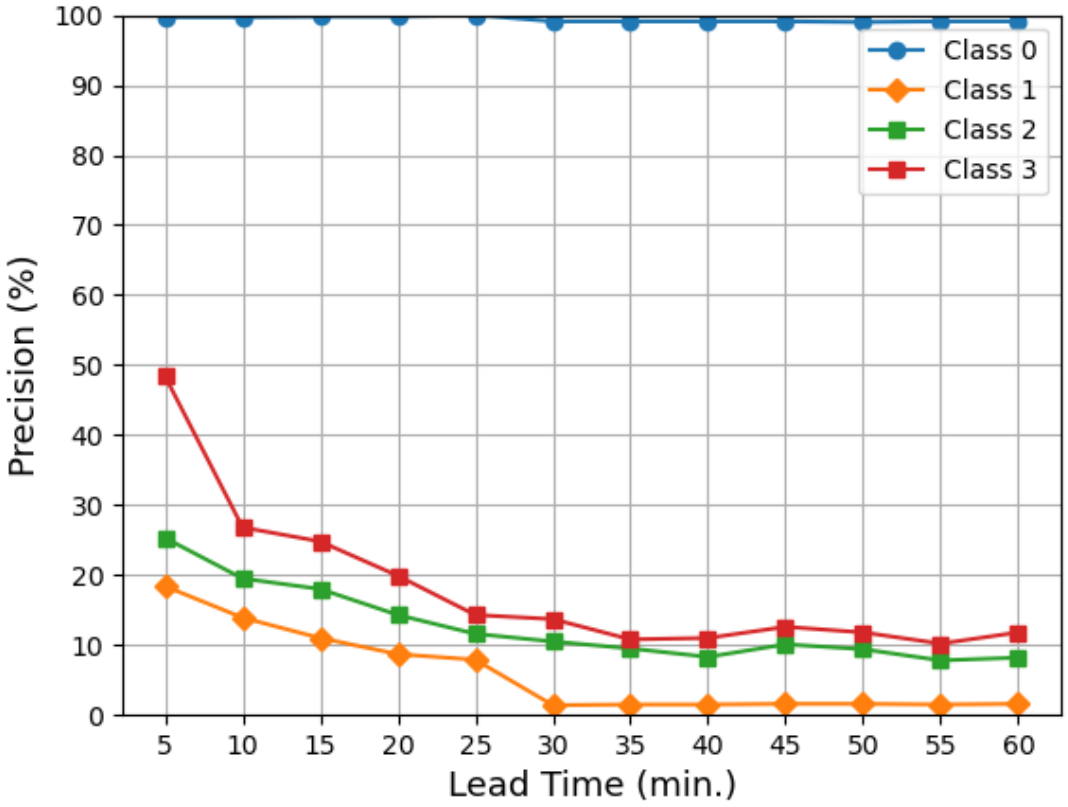

Figure 5.11: Variation of the model's performance according to the precision per lead time.

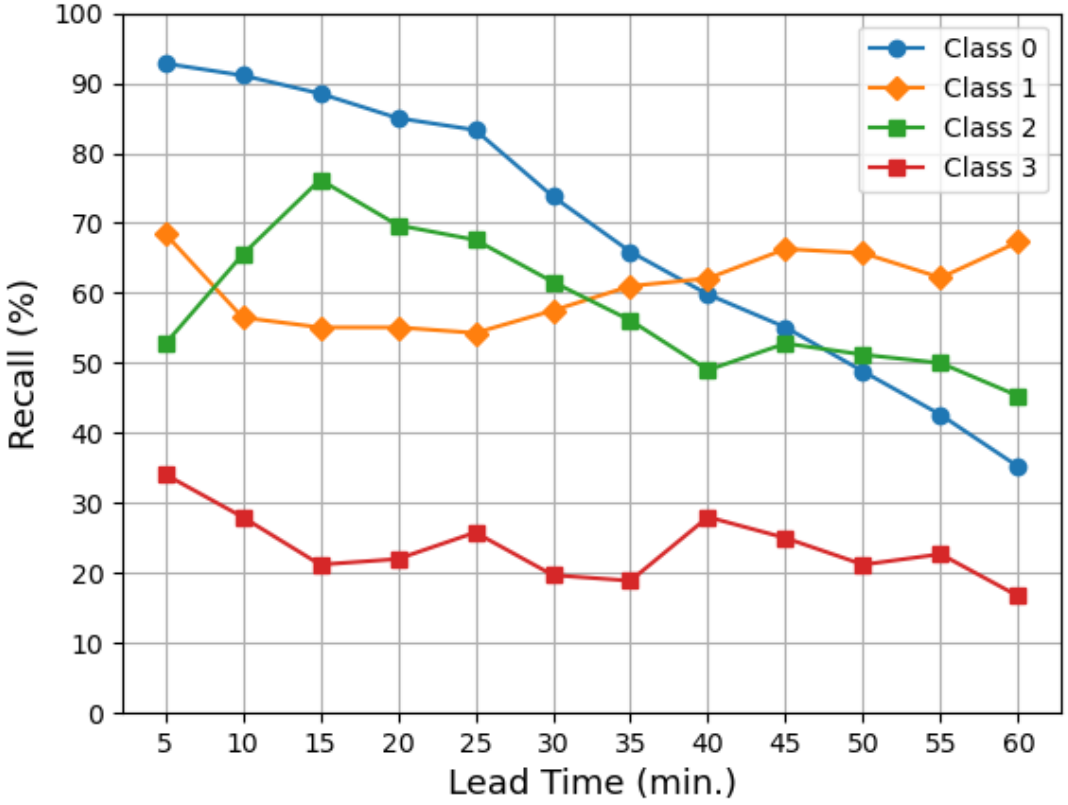

Figure 5.12: Variation of the model's performance according to the recall per lead time.

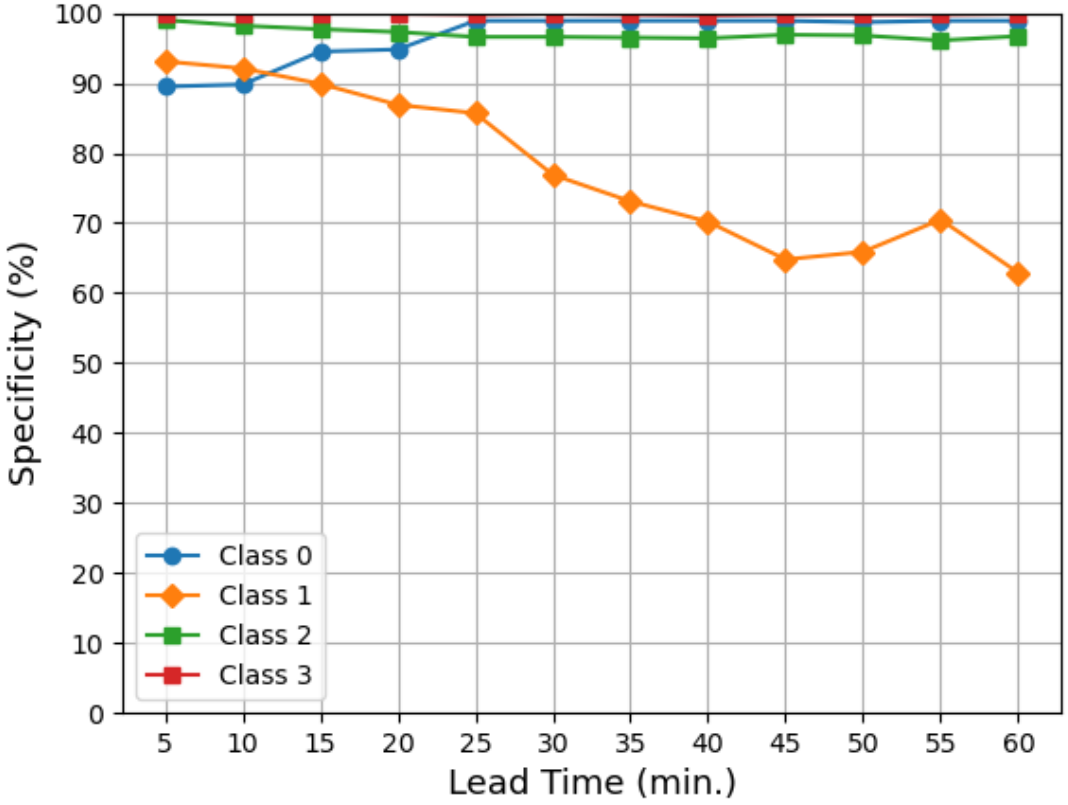

Figure 5.13: Variation of the model's performance according to the specificity per lead time.

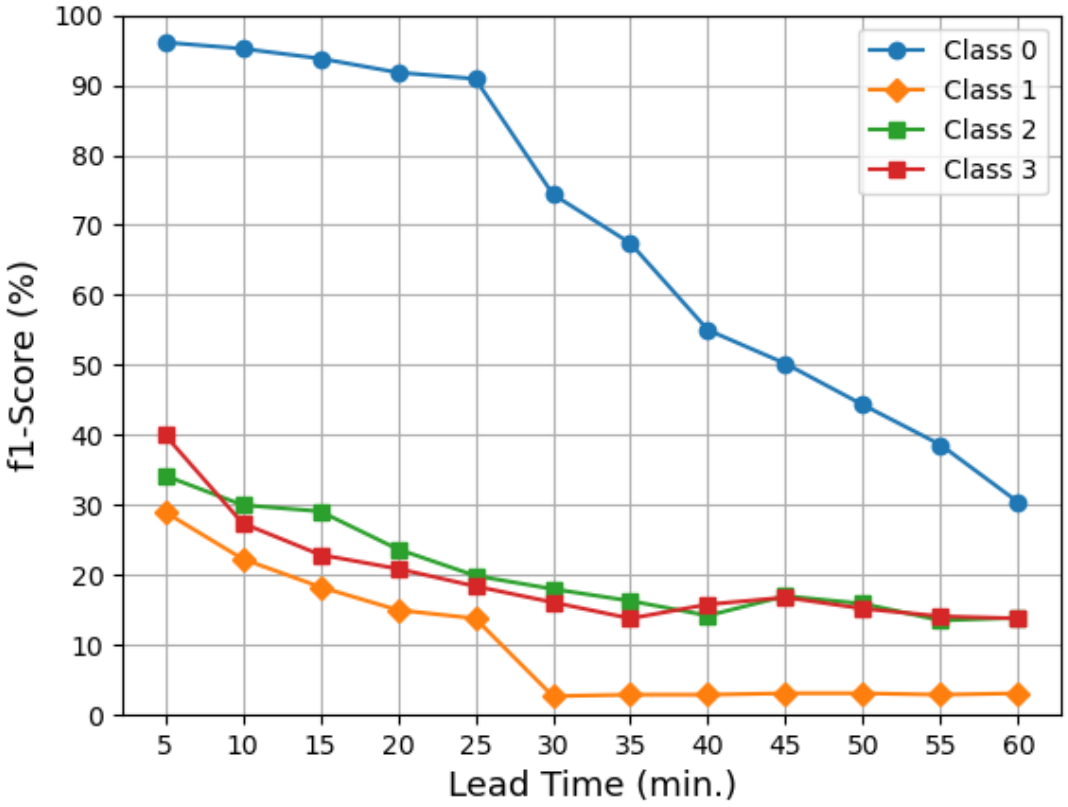

Figure 5.14: Variation of the model's performance according to the f1-Score per lead time.

6 CONCLUSION OF THE STUDY

Our study addresses the problem of Cloud-to-Ground Lightning (CG) forecasting for the next hour in five min. increments for the Brazilian south-southeastern region. Lightning is an electric discharge arising from the difference in the magnitude of the electric field in different levels of the atmosphere. Like many other atmospheric events, lightning occurs because of an imbalance in the atmosphere. These events acts to disperse the high electrified zones found in the upper levels of the atmosphere to the lower levels, effectively behaving to balance the electric charges on the atmosphere.

The influence of these atmospheric electric discharges is not limited to the atmosphere. It is estimated that lightning causes damage of billions of dollars worldwide, besides the immeasurable loss of lives (Cooper and Holle, 2018). South America's largest country, Brazil, has an average of 120 deaths per year due to CG strikes (ELAT/INPE, 2016). Considering its vast territorial extension and geographical location, Brazil provides the conditions of a natural laboratory for studying lightning (Jr. and Domingues, 2002).

As mentioned in Chapter 1, forecasting weather phenomenons involves modeling an extraordinarily complex and chaotic structure, the atmosphere (Rasp, 2018). Vast networks of sensors spread across the globe operate uninterruptedly to acquire information about many atmospheric variables, providing data for monitoring, predicting, and studying the weather and climate.

Massive amounts of data are generated daily from various detectors, ranging from elementary on-site sensors to sophisticated million-dollar remote systems like satellites. Given the colossal quantity and multifarious types of atmospheric data, automatic methods are needed to sift through and extract value from these datasets. Thus, Machine Learning (ML) techniques are highly applicable to meteorological studies for their ability to derive information and make patterns explicit in large datasets.

Hence, we decided to implement an ML method for the task of CG lightning nowcasting. Bearing that, the selected algorithm was the Gradient Boosting Regressor (GB) trained and validated on past CG occurrences declared by National Integrated Network for Detection of Atmospheric Discharges (RINDAT) from 2018 until 2022. The model was tested on the most recent data available, the year 2023.

The patterns of CG events on the Region of Interest (ROI) were analyzed to construct the datasets and the necessary feature engineering. Given that our study concerns an atmospheric event, the CG lightning seasonality was verified. Most events happen during the summer months, predominantly on the coast, due to the intense thermodynamic activity in the region. Spring ranks as the second season considering CG occurrences, given the incidence of Mesoscale Convective System (MCS), fostering large numbers of lightning in the central, southwestern zone of the ROI.

Autumn and winter, periods of colder temperatures and thus stabler weather, define the months with lesser CG strikes numbers.

The training and validation datasets were built with data from different periods of every year from 2018 until 2022 to present the seasonality of lightning to the model. Three algorithms were trained and had their performance qualified on the validation dataset, namely, a Linear Regression (LR), GB, and Random Forest (RF) algorithms. In view of the similar capacity of the algorithms, the GB method was selected for its ability to model non-linear data and provide a interpretation of the model inner-workings, by analysing the decision trees constructed.

The evaluation of the model's performance was thorough and multifaceted. By employing traditional classification metrics such as precision, recall, and specificity, we ensured a solid foundation of performance assessment. However, recognizing the limitations of point-based metrics, we also incorporated the Fraction Skill Score (FSS) to evaluate the spatial coherence of the predictions. This dual approach provided a more holistic understanding of the model's effectiveness.

The inclusion of case studies, selected for their significant lightning activity, further validated the model's robustness and practical utility. These case studies demonstrated the model's ability to handle diverse and challenging conditions, underscoring its potential impact in operational settings.

This study has successfully developed and validated a ML model that improves on the forecasting of CG lightning for our ROI in the very-short term. The advancements presented in this research contribute to the broader field of weather forecasting and machine learning, paving the way for more accurate and reliable predictions of natural phenomena. As we continue to face increasing challenges from extreme weather events, the development of sophisticated predictive models such as the one presented here will be essential in safeguarding lives and property.

6.1 NEXT STEPS

Building on the success of the constructed model presented in this dissertation, several key next steps are proposed to further enhance the research and its applications in lightning nowcasting.

To improve the model's quality, future research should consider integrating additional features besides lightning activity. Given that lightning presents a seasonality, the inclusion of the date and time of events may prove to be a useful predictor element. Moreover, the integration of a wider range of atmospheric variables, such as precipitation data and topography, may provide additional context and improve the model's ability to capture the complex interactions that lead to lightning events.

Exploring more modern and sophisticated ML techniques, such as deep learning, could uncover new patterns and improve the model's predictive performance. A possible new avenue is to treat the problem of lightning forecasting as a image segmentation problem, which is a task prone for neural networks.

Integrating real-time data streams, such as radar and satellite imagery, may enhance the model's responsiveness. The use of a stream ML method can result in a system that continuously updates the model with the latest data, improving its forecasting capabilities and provide up-to-date predictions.

By pursuing these next steps, the research can continue to advance the field of lightning forecasting, ultimately leading to more accurate, reliable, and actionable predictions. These efforts will contribute to improved safety measures, better preparedness for lightning-related hazards, and a deeper understanding of the atmospheric processes that drive lightning activity.

REFERENCES

ANEEL (2021). Agência nacional de energia elétrica - geração. Acesso em: 23 fev. 2022.

Bansal, A. K. (2022). Sizing and forecasting techniques in photovoltaic-wind based hybrid renewable energy system: A review. *Journal of Cleaner Production*, 369:133376.

Beneti, C. A. A. (2012). *Caracterização Hidrodinâmica e Elétrica de Sistemas Convectivos de Mesoescala*. PhD thesis, Universidade de São Paulo - Instituto de Astronomia, Geofísica e Ciências Atmosféricas, São Paulo - SP.

Beneti, C. A. A., Leite, E. A., Garcia, S. A. M., Assunção, L. A. R., Filho, A. C., and Reis, R. J. (2000). RIDAT – rede integrada de detecção de descargas atmosféricas no Brasil: situação atual, aplicações e perspectivas. *XI Congresso Brasileiro de Meteorologia*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Browning, K. A. and Monk, G. A. (1982). A simple model for the synoptic analysis of cold fronts. *Quarterly Journal of the Royal Meteorological Society*, 108(456):435–452.

Cintineo, J. L., Pavolonis, M. J., and Sieglaff, J. M. (2022). Probsevere lightningcast: A deep-learning model for satellite-based lightning nowcasting. *Weather and Forecasting*, 37(7):1239 – 1257.

CNN (2014a). Lightning strike kills 11 in colombia. https://us.cnn.com/2014/10/06/world/americas/colombia-lightning-strike/index.html. Accessed 02/06/2023.

CNN (2014b). Lightning strikes in colorado park kill 2 people in 2 days. https://edition.cnn.com/2014/07/12/us/colorado-lightning-strike/index.html. Accessed 02/06/2023.

CNN (2014c). One dead, 13 injured after lightning strikes at southern california beach. https://edition.cnn.com/2014/07/27/us/lightning-strike-venice-beach/index.html. Accessed 02/06/2023.

Consultor Jurídico (2023). Agropecuária deve indenizar mãe de vaqueiro morto ao ser atingido por raio. https://www.conjur.com.br/2023-dez-05/agropecuaria-deve-indenizar-mae-de-vaqueiro-morto-ao-ser-atingido-por-raio/. Accessed 06/12/2023.

Cooper, M. A. and Holle, R. (2018). *Reducing Lightning Injuries Worldwide, Springer Natural Hazards Series*. Springer International Publishing.

DiGangi, E., Lapierre, J., Stock, M., Hoekzema, M., and Cunha, B. (2022). Analyzing lightning characteristics in central and southern South America. *Electric Power Systems Research*, 213.

Dodgson, N. (1997). Quadratic interpolation for image resampling. *IEEE Transactions on Image Processing*, 6(9):1322–1326.

Ebert, E., Brown, B., Fowler, T., Gill, P., Göber, M., Joslyn, S., Mittermaier, M., Nurmi, P., Watkins, A., and Weigel, A. (2013). Progress and challenges in forecast verification. *Meteorological Applications*, 20.

ELAT/INPE (2016). Infográfico - morte por raios. `http://www.inpe.br/webelat/homepage/menu/el.atm/mortes.por.raios.-.infografico.php`. Accessed 11/11/2021.

Elsenheimer, C. B. and Gravelle, C. M. (2019). Introducing lightning threat messaging using the GOES-16 day cloud phase distinction RGB composite. *Forecaster's Forum*, 34:1587–1600.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media, Inc.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Haupt, S. E., Gagne, D. J., Hsieh, W. W., Krasnopolsky, V., McGovern, A., Marzban, C., Moninger, W., Tissot, V. L. P., and Williams, J. K. (2021). The history and practice of AI in the environmental sciences. *Bulletin of the American Meteorological Society*.

Hayward, L., Whitworth, M., Pepin, N., and Dorling, S. (2020). Review article: A comprehensive review of datasets and methodologies employed to produce thunderstorm climatologies. *Natural Hazards and Earth System Sciences*, 20(9):2463–2482.

Holle, R. L. (2014). Some aspects of global lightning impacts. In *2014 International Conference on Lightning Protection (ICLP)*, pages 1390–1395.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Instituto Brasileiro de Geografia e Estatística, I. (2023). Censo brasileiro de 2022.

Itaipu Binacional (2023). Itaipu binacional: Perguntas frequentes. https://www.itaipu.gov.br/sala-de-imprensa/perguntas-frequentes. Accessed 02/06/2023.

Jr., O. M. and Domingues, M. O. (2002). Introdução à eletrodinâmica atmosférica. *Revista Brasileira de Ensino de Física*, 24(1).

Krehbiel, P. R. (1986). *The Earth's Electrical Environment*, pages 90–113. National Academy Press.

Langley, P. and Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the Association for Computing Machinery*.

Lee, D. (2020). *Birth of Intelligence*. Oxford University Press.

Leinonen, J., Hamann, U., and Germann, U. (2022). Seamless lightning nowcasting with recurrent-convolutional deep learning. *Artificial Intelligence for the Earth Systems*, 1(4):e220043.

Leite, E. A., Igarashi, A. Y., and Jusevicius, M. A. (2011). Sistema de previsão probabilística espacial de eventos de descargas atmosféricas e sua aplicação na vigilância meteorológica do sistema elétrico. *XIX Seminário Nacional de Produção e Transmissão de Energia Elétrica*.

Lopez, P. (2016). A lightning parameterization for the ECMWF integrated forecasting system. *Monthly Weather Review*, 144(9):3057–3075.

Mansouri, E., Mostajabi, A., Tong, C., Rubinstein, M., and Rachidi, F. (2023). Lightning nowcasting using solely lightning data. *Atmosphere*, 14(12).

McGovern, A., Lagerquist, R., II, D. J. G., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T. (2019). Making the black box more transparent - understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, 100(11):2175–2199.

METED (2014). GOES-R GLM: Introduction to the geostationary lightning mapper. https://www.meted.ucar.edu/goes_r/glm. Accessed 28/10/2021.

Michie, D., Spiegelhalter, D., and Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood.

Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill.

Miyazaki, T. and Okabe, S. (2010). Experimental investigation to calculate the lightning outage rate of a distribution system. *IEEE Transactions on Power Delivery*, 25(4):2913–2922.

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7.

Newell, A. and Simon, H. (1956). The logic theory machine: A complex information processing system. *IRE Transactions on Information Theory*, 2:61–79.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peterson, M. (2021). Where are the most extraordinary lightning megaflashes in the Americas? *Journal of American Meteorological Society*.

Rakov, V. A. (2016). *Fundamentals of Lightning*. Cambridge University Press.

Rasp, S. (2018). *Statistical methods and machine learning in weather and climate modeling*. PhD thesis, Ludwig Maximilian University of Munich, Munich - Germany. 143 pgs.

Rudlosky, S., Goodman, S., and Virts, K. (2019). GOES-R lightning detection fact sheet. Technical report, National Oceanic and Atmospheric Administration (NOAA).

Russell, S. J. and Norvig, P. (2010). *Artificial Intelligence - A Modern Approach*. Pearson Education, Inc.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal*, 3(3).

Sato, H., Sugiyama, A., and Ban, H. (2008). Online lightning prediction system. *NTT Technical Review*, 6(2).

Saunders, C. (2008). Charge separation mechanisms in clouds. *Space Science Review*, 137:335–353.

Schön, C., Dittrich, J., and Müller, R. (2018). The error is the feature: how to forecast lightning using a model prediction error. *Proceedings of 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Schön, C. and Dittrich, J. (2019). Make thunderbolts less frightening — predicting extreme weather using deep learning. In *33rd Conference on Neural Information Processing Systems*, Vancouver - Canada.

Skific, N. and Francis, J. (2012). Self-organizing maps: A powerful tool for the atmospheric sciences. In Johnsson, M., editor, *Applications of Self-Organizing Maps*, chapter 13. IntechOpen, Rijeka.

Skok, G. and Roberts, N. (2016). Analysis of fractions skill score properties for random precipitation fields and ecmwf forecasts: Fss properties for random precipitation fields and ecmwf forecasts. *Quarterly Journal of the Royal Meteorological Society*, 142.

Sollaci, L. B. and Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association*, 92:364–367.

Song, G., Li, S., and Xing, J. (2023). Lightning nowcasting with aerosol-informed machine learning and satellite-enriched dataset. *Climate and Atmospheric Science*, 6(126).

Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

V. A. Rakov, M. A. U. (2003). *Lightning: Physics and Effects*. Cambridge University Press.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Vettigli, G. (2018). Minisom: minimalistic and numpy-based implementation of the self organizing map.

Wallace, J. M. and Hobbs, P. V. (2006). *Atmospheric Science: An Introductory Survey*. Academic Press.

Warner, T. A. (2020). Electrification of clouds. `https://ztresearch.blog/education/electrification-of-clouds`. Accessed 30/10/2021.

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences*. Academic Press.

Williams, E. R. (1985). Large-scale charge separation in thunderclouds. *Journal of Geophysical Research: Atmospheres*, 90(4):6013–6025.

World Meteorological Organization (2017). *Guidelines for Nowcasting*. Chairperson, Publications Board.

Zhou, K., Zheng, Y., Dong, W., and Wang, T. (2020). A deep learning network for cloud-to-ground lightning nowcasting with multisource data. *Journal of American Meteorological Society*, 37:927–942.

7 APPENDIX: FORECAST OF AN EVENT FOR ONE HOUR

This Appendix presents a case study on Cloud-to-Ground Lightning (CG) lightning forecasting for our Region of Interest (ROI). The forecasts are made for up to one hour, in five-minute increments, providing a granular view of how the model predicts lightning occurrence over short time intervals.

As verified by the performance metrics, the model display a up to standard capacity for a lead time of up to 30 min. It is noticeable that as the lead time increases, the model tends to underestimate the number of elements with CG lightning and the severity of events. The areas with the greater numbers of lightning are predicted, but the more isolated events are not, particularly for lead times longer than 45 min. Nonetheless, the areas which the model predicted that were under lightning activity were correct.

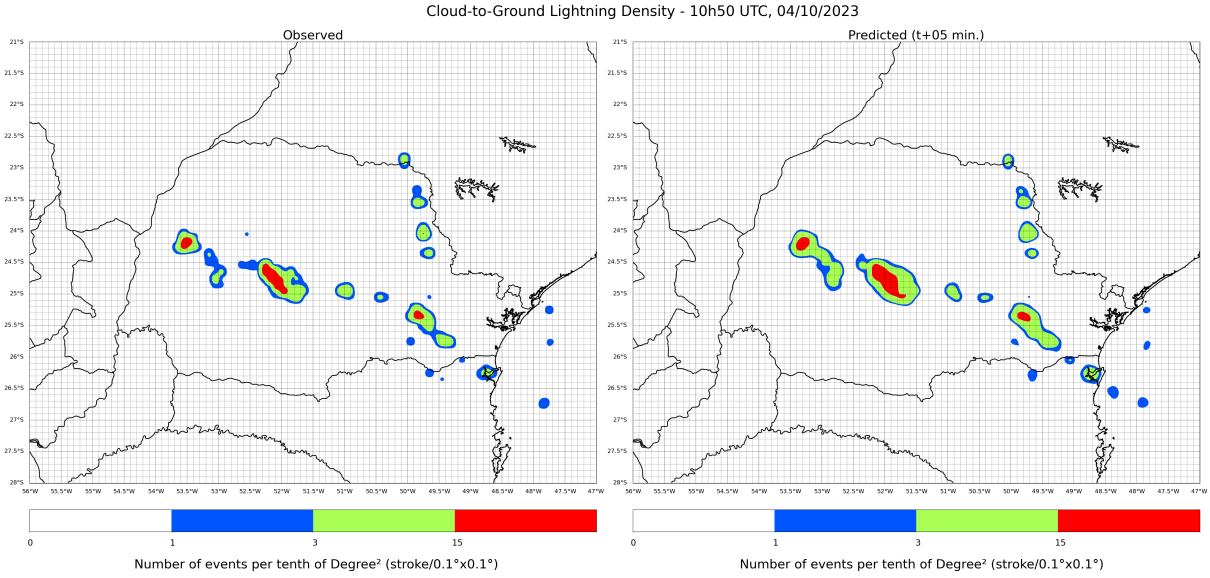

Figure 7.1: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of five min.

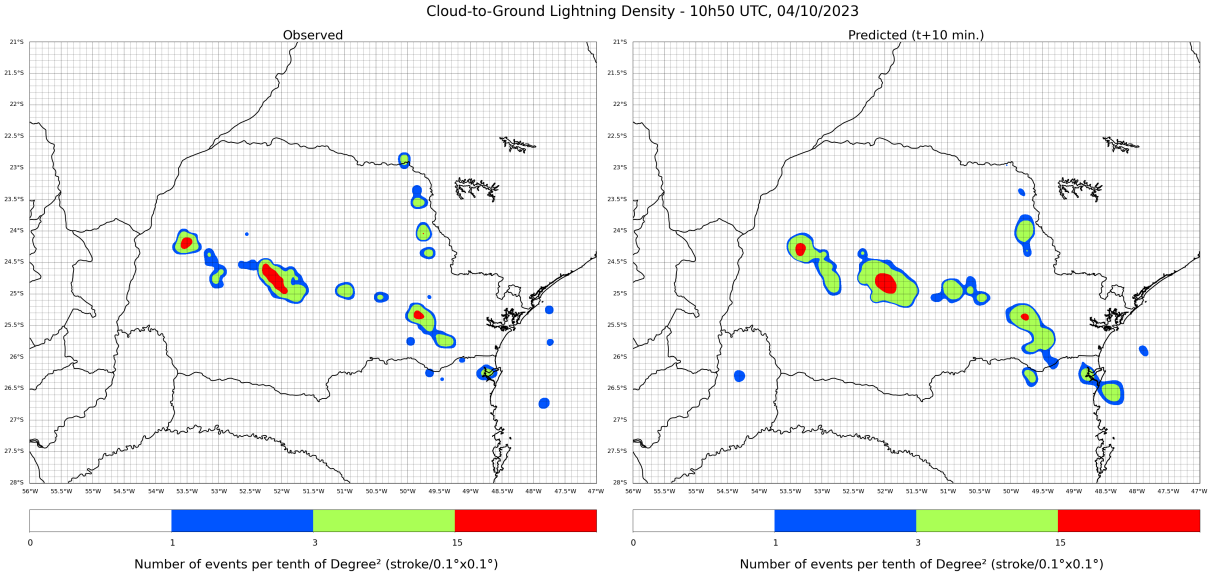

Figure 7.2: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 10 min.

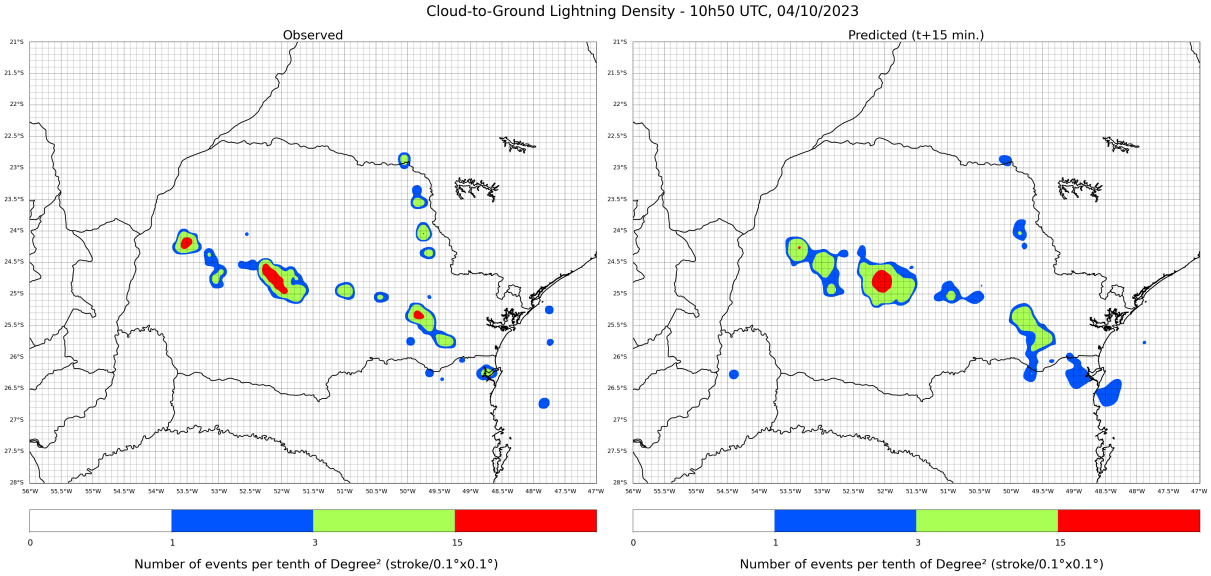

Figure 7.3: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 15 min.

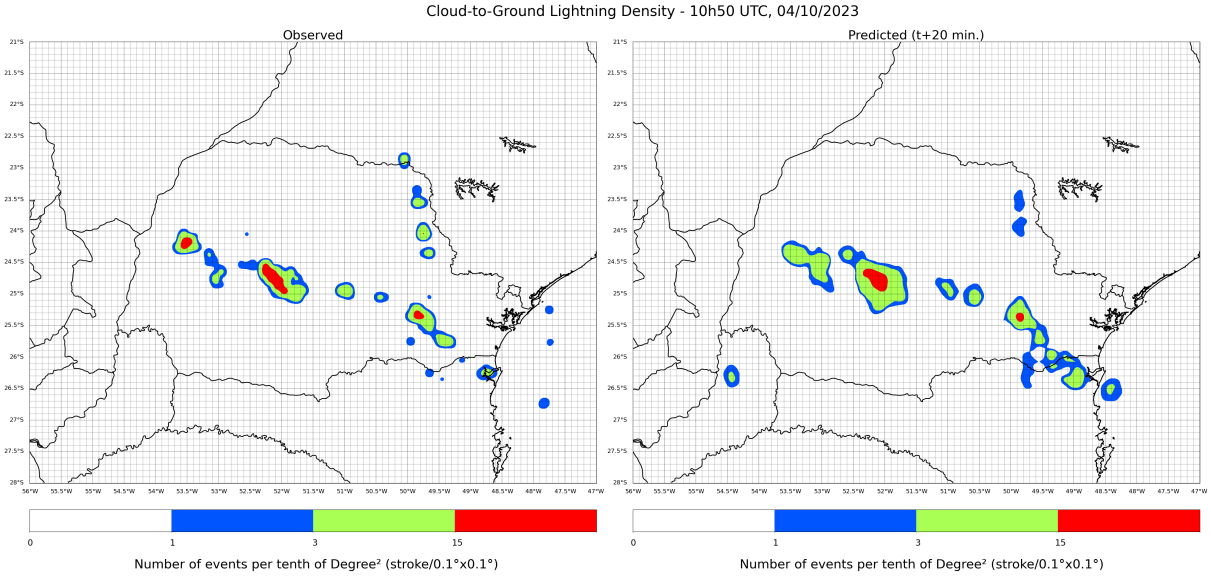

Figure 7.4: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 20 min.

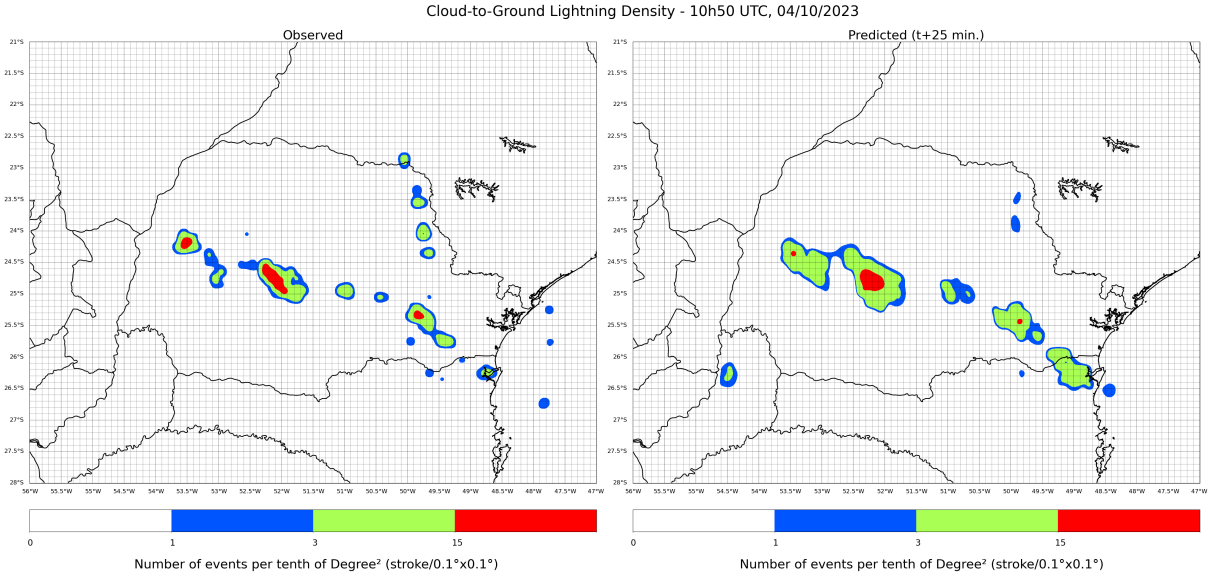

Figure 7.5: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 25 min.

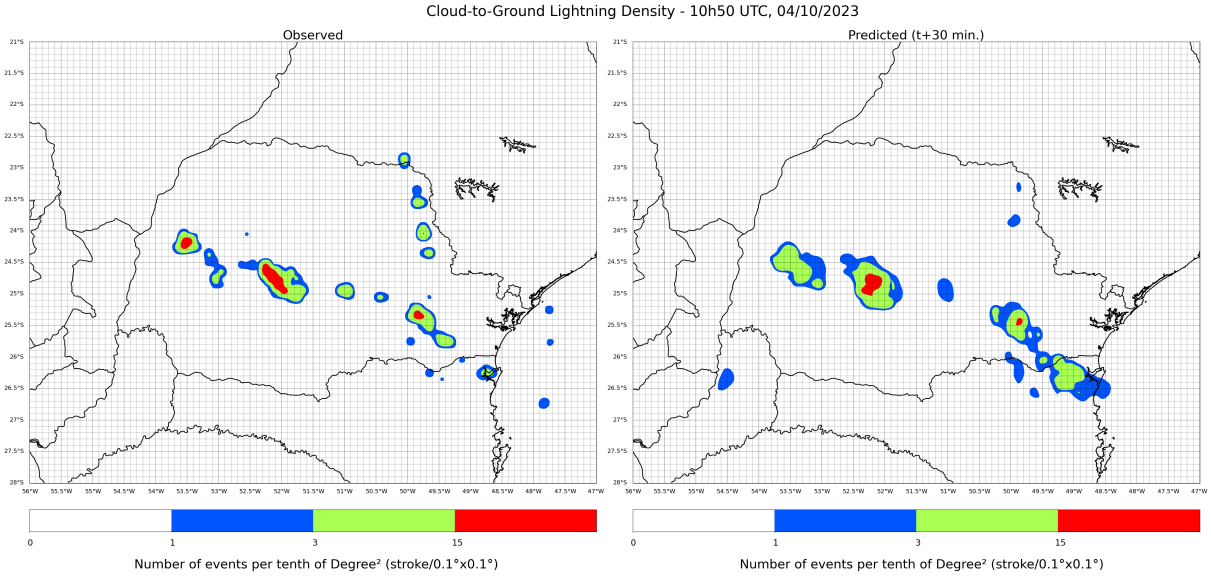

Figure 7.6: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 30 min.

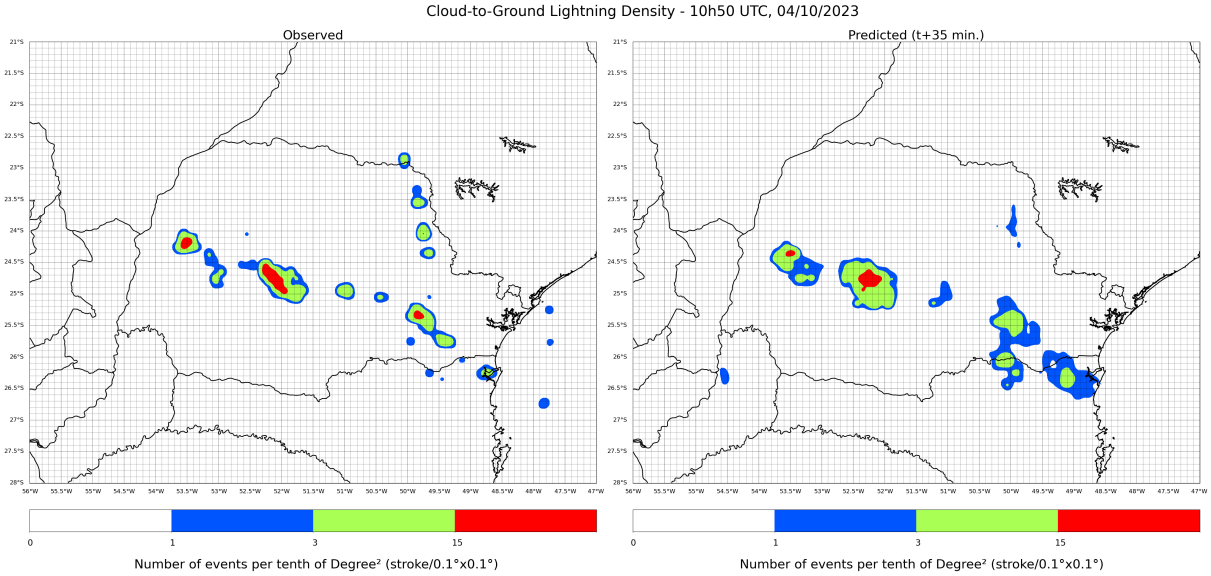

Figure 7.7: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 35 min.

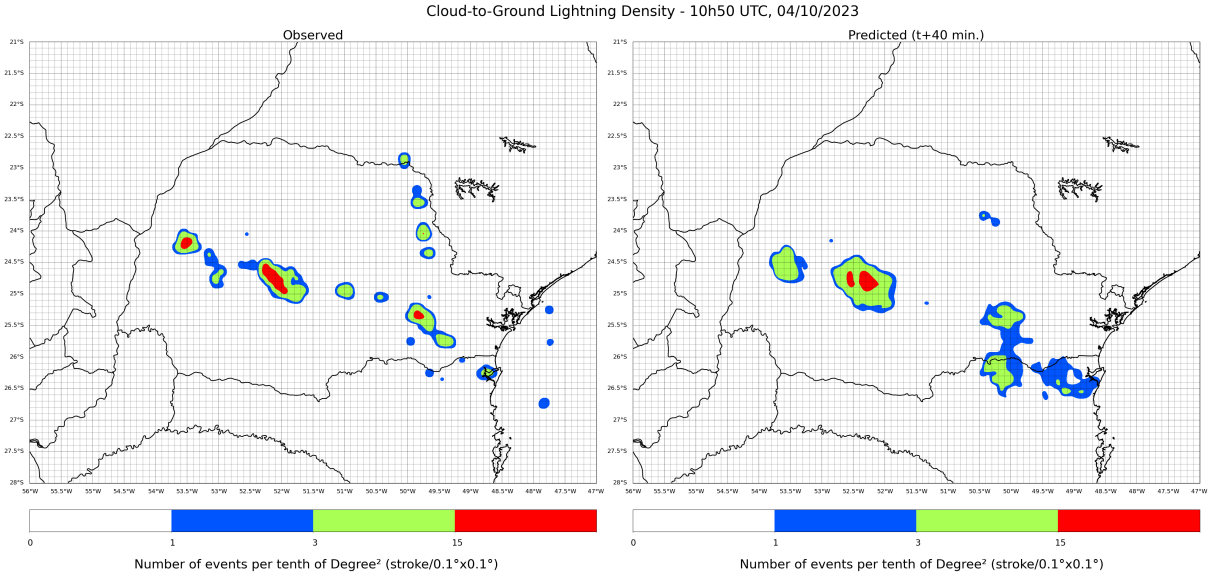

Figure 7.8: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 40 min.

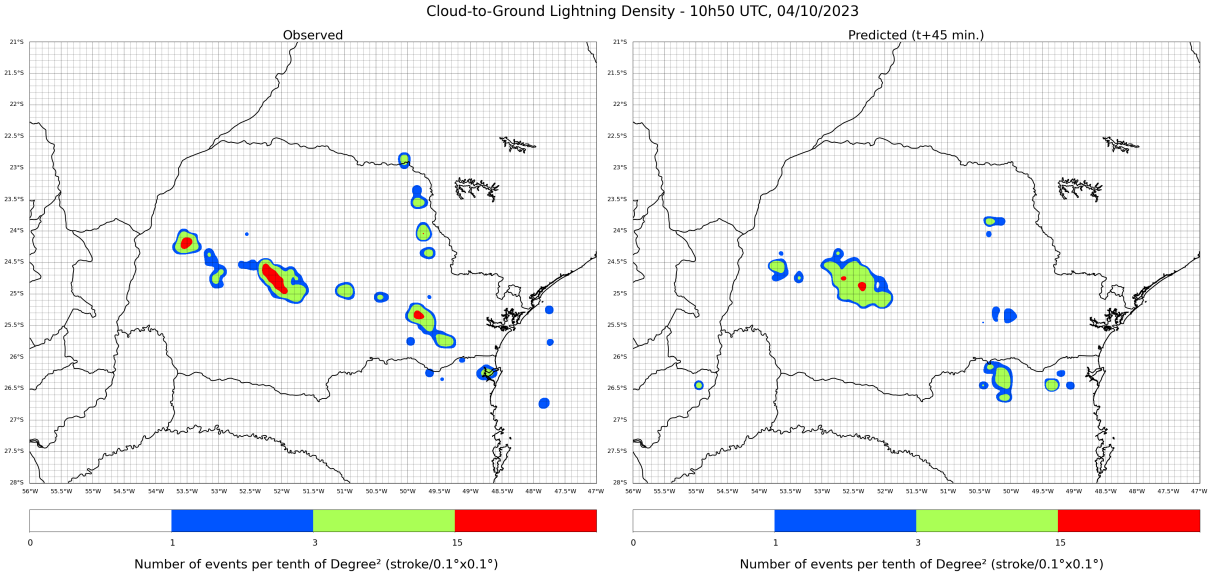

Figure 7.9: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 45 min.

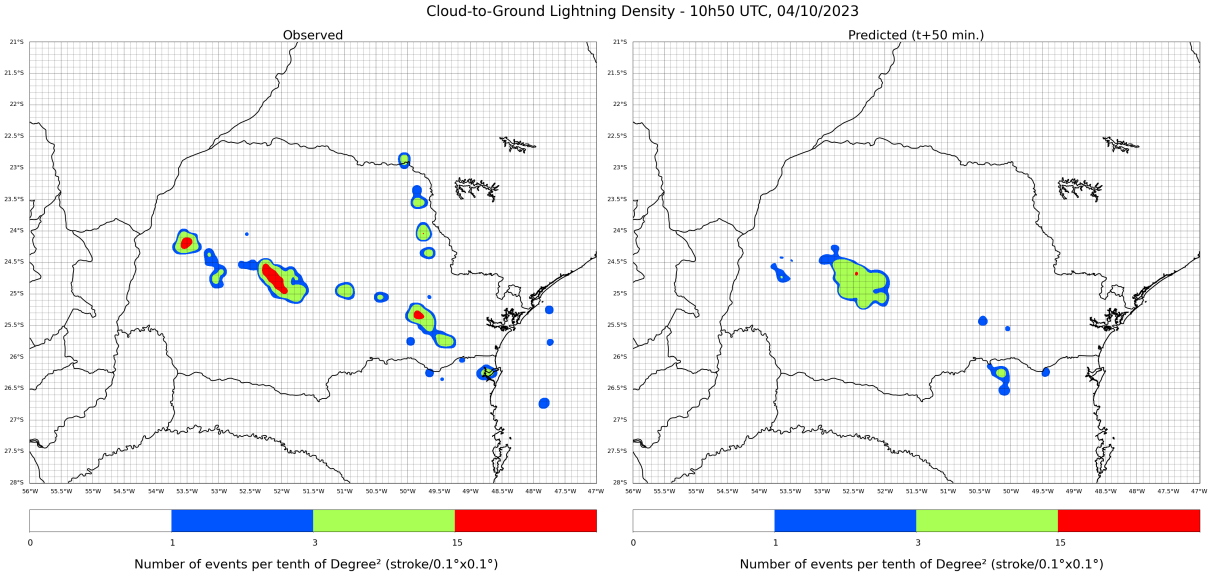

Figure 7.10: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 50 min.

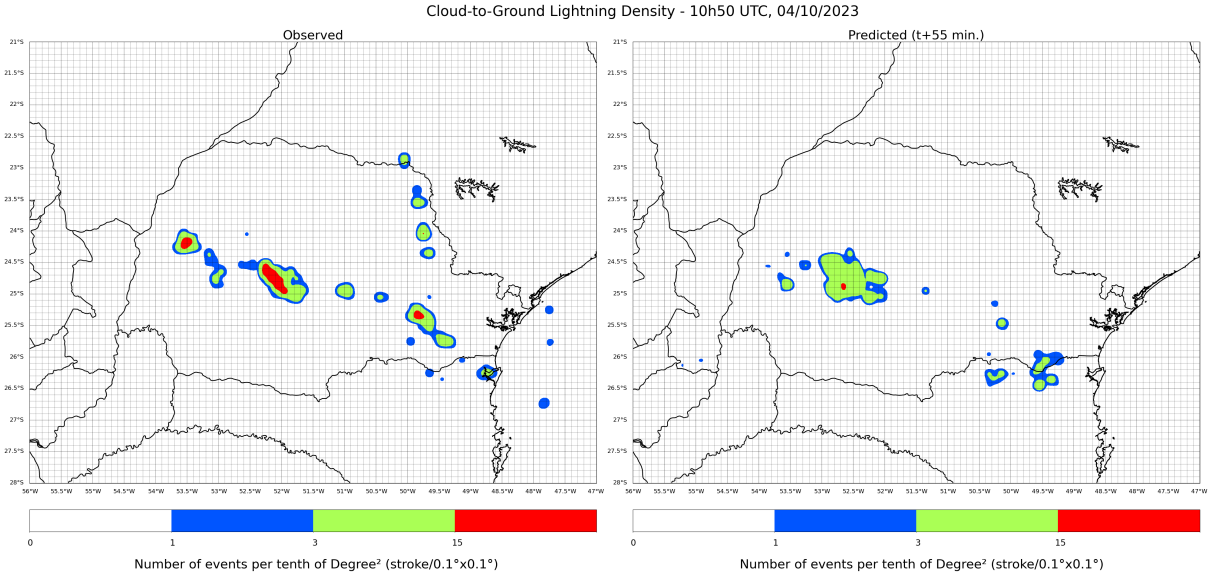

Figure 7.11: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 55 min.

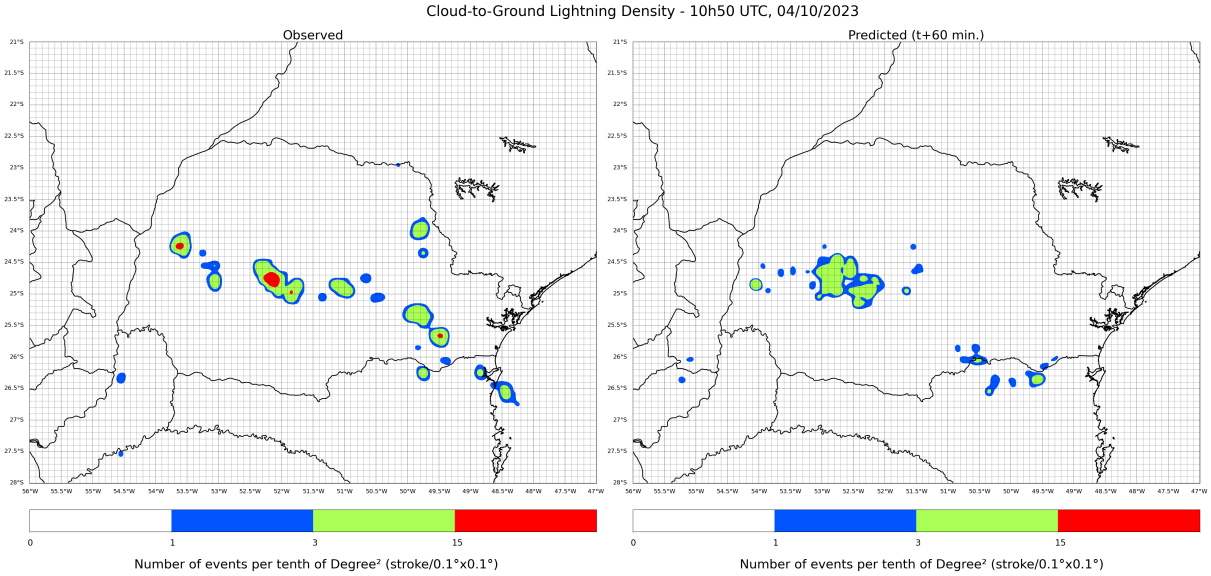

Figure 7.12: The left map displays the observed CG lightning density on the ROI and the right map displays the predicted CG lightning density at 10h50 UTC on 10/04/2023 for a lead time of 60 min.