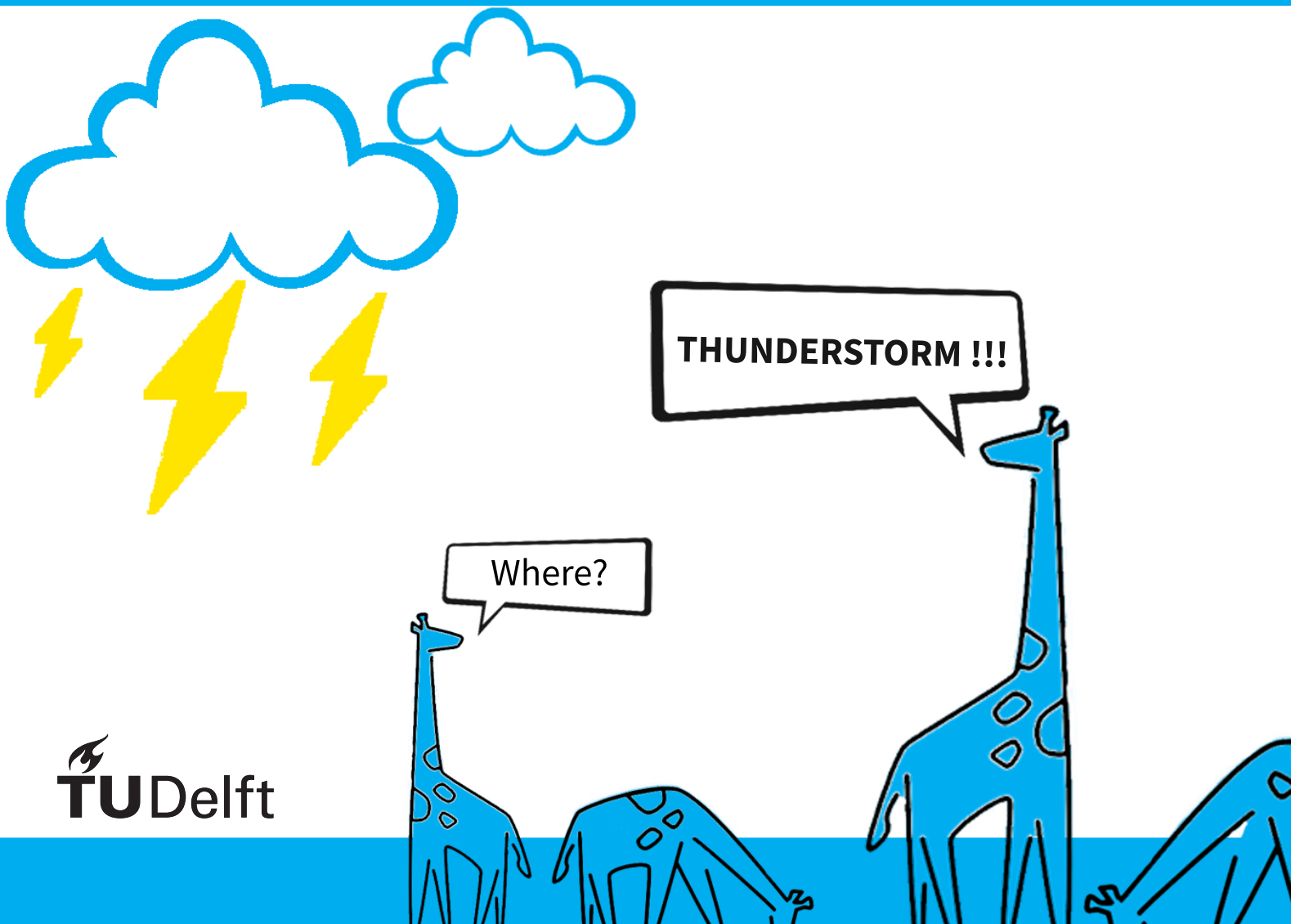
Early Warning Model for Thunderstorms around Lake Victoria

B.J. Magura

Master Thesis for Environmental Engineering
Science Track



TU Delft

Early Warning Model for Thunderstorms around Lake Victoria

by

B.J. Magura

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on August 26, 2020 at 16:30 PM.

Student number:	4173139	
Thesis committee:	Dr. Ir. Marie-Claire ten Veldhuis	TU Delft
	Dr. Marc Schleiss	TU Delft
	Prof. dr. ir. Nick van de Giesen	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TAHMO TU Delft TWIGA

Preface & Acknowledgements

From 2015 to 2018 I worked as a business developer in Sub-Saharan Africa for an Italian solar company. Eager to stimulate solar energy and to understand Africa I set out on a forming and inspirational journey that took me to nine different African countries. I learned how something as simple as a 50 Euro solar kit could allow a child to study at home or a family to gain income by charging phones. I also learned the competitive advantage of the European and American solar companies, and how despite this African entrepreneurs were still able to succeed.

When it was time to move on, I decided to do the master Environmental Engineering but the period in Africa still influenced everything I did. For example, during my master I started Wazo Thrive, a non-profit that focuses on economic empowerment through education and starting small businesses. I also became increasingly interested in global warming and how it will affect our future. Now that I am nearing the end of my master the main lesson I take with me is that switching to renewable energy will not be enough. Changing our economic system to penalize our extreme consumption and polluting lifestyle, instead of stimulating it as is done now, will be vital. Even if this succeeds, adapting to the consequences of climate change will be inevitable.

Moving on to the next stage in my life, I have the deep desire to contribute to both global warming mitigation and adaptation, as well as economic reform. With this thesis report here before you, I hope to take a step in the right direction in global warming adaptation. It is for this reason that I want to thank Nick van de Giesen for giving me the opportunity to work on such inspirational projects as TAHMO and TWIGA. I thank Marc Schleiss for his excellent lectures and insights on machine learning and statistics. A special thanks to Marie-Claire ten Veldhuis, who gave me many valuable tips, ideas, and improvements during our meetings, Skype calls and while reviewing the drafts. I thank all three committee members for their feedback, time, and pleasant cooperation, which continued flawlessly during the Corona crisis.

I was also lucky to have had help from friends. I would especially like to thank Dai for teaching me Latex, Mutale for proofreading my report, and Pieke for the idea of using giraffes as early warning model.

Finally, I would like to thank my parents for their financial support that allowed me to focus on my studies and for their unrelenting interest and encouragement.

B.J. Magura
Delft, August 2020

The work leading to these results has received funding from the European Community's Horizon 2020 Programme (2014-2020) under grant agreement n° 776691. The opinions expressed in the document are of the authors only and no way reflect the European Commission's opinions. The European Union is not liable for any use that may be made of the information.

TWIGA

List of Abbreviations

FAR	False Alarm Ratio
LR	Linear Regression
LSTM	Long short-term memory
NN	Neural Network
OT	Overshooting Tops
PDF	Probability Density Function
RI	Raw Inputs
TAHMO	Trans-African Hydro-Meteorological Observatory
TWIGA	Transforming water, weather and climate information through in-situ observations for geo-services in Africa
XGB	Extreme Gradient Boosting

Definitions

Early Warning Model
Model that aims to provide a warning for upcoming events, in this case thunderstorms, so that precautions can be taken to minimize the negative impacts.
F Score
Indicates the relative importance of model inputs for the XGBoost algorithm by summing the number of times an input was used to make a decision.
False Alarm Ratio
The percentage of alarms that are false. It is defined as false positives over false positives plus true positives.
False Negative
Occurs when no alarm is issued for a thunderstorm.
False Positive
An alarm is issued but no thunderstorm occurs.
Hit Rate
The percentage of thunderstorms that have been predicted. Defined as true positives over true positives plus false negatives.
Hybrid Model
Model developed for this study that combines classification and regression.
Knowledge Inputs
Model inputs based on an analysis of the Spearman correlation and probability density functions.
Lead Time
Time between when the alarm is issued and when the first lightning strike occurs.
Lightning Strike
A cloud-to-ground or cloud-to-cloud lightning strike as measured by the TAHMO lightning sensor.
Model Inputs
The data on which the model trains and predicts the model target.
Model Target
The variable that the model aims to predict.
Neural Network
A machine learning technique based on optimizing weights and biases in neurons to minimize the difference between model target and prediction.
Raw Inputs
The 9 variables that are measured by the TAHMO station, namely pressure, lightning distance, lightning strikes, precipitation, radiation, relative humidity, temperature, wind direction, and wind speed.
Storm Intensity
Defined as the number of lightning strikes occuring during the lifetime of the storm.
Thunderstorm
A period when the TAHMO lightning sensor measures lightning strikes. A storm ends when there has not been a lightning strike for one hour.
Training Set
Past measurements that are used to find the relationship between model inputs and model targets.
True Negative
No alarm is issued and no thunderstorm occurs.
True Positive
An alarm is issued and a thunderstorm occurs.
Validation Set
Measurements that are kept separate and are used as a final test to see how the model performs on unseen data.
XGBoost
Machine learning technique based on combining decision tree's.

Summary

Introduction
This report describes the development of an early warning model for thunderstorm occurrence around Lake Victoria. It is the first model to use the TAHMO lightning sensor data to predict thunderstorms. This study hopes to contribute to the TWIGA focus area of increasing disaster resilience through forecasting and early warnings and could be used to reduce the impact of thunderstorms for the communities around Lake Victoria, and to aid the flights of Zipline's medicine-delivery drones.

Literature Review
Lake Victoria is one of the most lightning active places in the world. The main mechanism of thunderstorm formation is differential heating and cooling between land and lake. This causes nighttime thunderstorms over the lake and daytime thunderstorms over the land. Four existing early warning models have been investigated. Two models used weather station data, one satellite data, and the last one balloon soundings. The study areas were Lake Victoria, Switzerland, and Northern Italy. Data records from 9 to 11 years were used and the lead times ranged from 10 minutes to 6 hours. Two models used linear techniques and two models non-linear techniques, namely a neural network and an ensemble of decision trees. The models achieved hit rates of 0.77, 0.83, 0.85, and 0.89. The false alarm ratios were 0.03, 0.29, 0.60, and 0.94. In this study the effect of different prediction techniques, both linear and non-linear, as well as different model inputs will be further investigated. Moreover, this study also aims to predict the intensity of the approaching storms, which could be useful in guiding emergency precautions.

Method
From the 20 TAHMO stations on the Kenyan side of Lake Victoria, TA00173 is chosen to build the model based on the 3 year data length and the the 2.7% missing values. The station measures 9 variables, namely pressure, lightning distance, lightning strikes, precipitation, radiation, relative humidity, temperature, wind direction, and wind speed. The data is divided into a two year training set from June 2017 until May 2019, and an one year validation set from June 2019 to May 2020. The model is tested with two different sets of model inputs. The first set contains 19 model inputs, selected from the 5-minute interval measurements, standard deviations, gradients and past 24 hour values, based on the Spearman correlation and probability density functions (knowledge inputs). The second input set contains the 5-minute interval measurements of the 9 variables (raw inputs). Three model techniques are tested, namely a linear regression model, a neural network, and an ensemble of decision tree's (XGBoost). The three techniques and two input sets lead to six model configurations. A modelling approach is developed which combines classification and regression into a single model. The model target is the sum of lightning strikes in the next six hours at 5-minute resolution. The model predictions are summed at this resolution and once the threshold is exceeded an early warning is issued. The value of this threshold influences the hit rate, false alarm ratio, and lead time. The average model prediction six hours before the first lightning strike gives the thunderstorm intensity. The advantage, compared to for example the set-up of the other Lake Victoria model, is that the temporal resolution is maintained because the aggregation occurs after the prediction and not before. This also leads to a more precise lead time. The algorithm is newly developed and should be reviewed by another party.

Results
The raw inputs score best on the number of true positives. The improvement compared to the knowledge inputs is 14%, 10%, and 6%, for the linear regression model, the neural network, and the XGBoost, respectively. The knowledge inputs score best on the number of false positives. The improvement compared to the raw inputs is 0%, 14%, and 15%, for the linear regression model, the neural network, and the XGBoost, respectively. The knowledge inputs also increase the lead time by 5%, 6%, and 11%, for the linear regression model, the neural network, and the XGBoost, respectively. The non-linear techniques improve the true positives by 4% for the neural network and reduce them with 1% for the

XGBoost, compared to the best linear regression model. The effect on the false positives is larger, namely an improvement of 40% for the neural network and 35% for the XGBoost model. An increase in lead time of 19% for the neural network and 10% for the XGBoost is achieved. The knowledge inputs combined with non-linear techniques have slightly smaller errors in the intensity prediction of the medium and large thunderstorms.

The neural network with raw inputs gives the highest hit rate (0.91) of all six configurations. The neural network with knowledge inputs has a lower hit rate (0.81) but the lowest false alarm ratio (0.43), the highest average lead time (220 minutes), and the lowest mean absolute error and mean absolute relative error for the medium (51-100 lightning strikes) and large thunderstorms (101-889 lightning strikes). The neural network with knowledge inputs therefore performs best on three of the four criteria. This model predicts 220 out of 273 thunderstorms, issues 164 false alarms and has 806 true negatives. The model has an mean absolute error of 58 lightning strikes and a mean absolute relative error of 5. Compared to the existing early warning models, the neural network with knowledge inputs ranks fourth out of five on the hit rate and third out of five on the false alarm ratio. The model does this with a data length record of three years compared to at least nine year of the other models.

Conclusion

From the existing early warning models it is concluded that regardless of the type of data source, area, and prediction technique, it is possible to predict the majority of thunderstorms. However, with the exception of the Switzerland study which used a very short lead time, the challenge is to reduce the false alarm ratio. This study finds that for predicting thunderstorms a linear technique and current weather station measurements are sufficient, but using non-linear techniques and past temporal weather station measurements reduces the false alarms and improves the lead time. This effect is also seen for the intensity prediction but no conclusions are drawn at this point due to the small differences and overall lacking skill in this area. Overall, it is concluded that the classification results of this model show promise but the false alarms are still too high for any practical application. Moreover, at this stage the model is not able to predict if the upcoming thunderstorm will be small or large.

Recommendations

Several recommendations are made to further develop the model. These are combining the neural network with raw inputs and the neural network with knowledge inputs, reviewing the model set-up and algorithm, optimizing the model parameters, trying out a LSTM neural network, conducting a sensitivity analysis on the lead time, data length, and model inputs, combining multiple weather stations, and trying additional data sources such as overshooting tops or numerical weather predictions. To understand if the model also has potential for aiding Zipline's medicine-delivery drones, it should be tested on a weather station in a different area. Finally, to learn how the model performs in practice it should be tested on location.

Contents

1 Introduction

This report describes the development of an early warning model for thunderstorm occurrence. The possibility and idea of creating this model originates from the desire to investigate and use the lightning sensor data that is measured by the TAHMO stations. This is the first attempt to use this data for such an early warning model. As a starting point, the Lake Victoria area is chosen for this research because of the dense TAHMO weather station network and the high lightning activity. The early warning model has the potential to reduce the impact of thunderstorms for the communities around Lake Victoria as well as improve the flight safety for medicine-delivery drones. This potential will be elaborated on in Section 1.1. The research fits well within the objectives of the TWIGA project, so a brief framework is presented of both TAHMO and TWIGA in Section 1.2. To introduce some overarching structure, the aim of the research is described in Section 1.3 in the form of an objective, research questions and model criteria. Finally, Section 1.4 gives a brief outline of the chapters for the reader to have a comprehensive idea of what to expect.

1.1. Relevance

The quality of weather services in Africa is still low compared to other regions. For example, the one-day weather forecast skills in the tropics are similar to those at day six in the extra tropics (Haiden et al. 2012). The lack of skill is one of the reasons that weather and climate services are still underused in Africa. In East-Africa the percentage of the population using weather services is estimated to be between 15% and 82% depending on the service and population, with lower numbers in West-Africa (5.6%-76%) and higher numbers in Southern Africa (27%-86%)(Vaughan et al. 2019). The same study indicates that in Malawi, indigenous knowledge and personal experience was found to be more reliable. Global warming is threatening this knowledge with potentially harmful consequences. Another study shows that much of the weather and climate information in Africa comes from global data sets that have a coarse spatial resolution making it less useful for individual users (Singh et al. 2018). Creating weather services that are reliable and actually useful for end-users is for this reason a promising field.

This is also true for the area around Lake Victoria. The lake is the largest in Africa and sustains the livelihood of 30 million people (Thiery et al. 2017). The Ugandan Meteorology Department and the WMO indicate that Uganda, one of the countries bordering Lake Victoria, has 287 thunderstorm days a year (Mary and Gomes 2015). These thunderstorms have severe consequences, for example by posing a hazard to the 200.000 fishermen operating on the lake. The Red Cross estimates that 3000 to 5000 people die every year on the lake (Thiery et al. 2017), although the specific causes are unknown. It is estimated that at least 8 relatives, on average, depend on each fisherman. This, in turn, causes major economic and social implications. Besides affecting the lake, thunderstorms also provide risks for the surrounding land. Between January 2007 and December 2011, 150 deaths and 584 personal injuries due to lightning were reported in Uganda (Mary and Gomes 2012). Another study shows that between 2010 and 2012, the North Eastern Ugandan part of Lake Victoria had 18 deaths and 46 injuries reported and the North Western Ugandan part 22 deaths and 50 injuries (Mary and Gomes 2015). Although not around Lake Victoria, another potential application of the early thunderstorm warning is for the delivery of emergency supplies by drones. In Rwanda and Ghana, the Zipline company uses

drones to quickly deliver medical supplies to remote areas, but they are often lost due to downdrafts. These downdrafts occur during cold pools caused by evaporation of heavy rainfall (Schlemmer and Hohenegger 2014). Since thunderstorms coincide with heavy rainfall, an indication of the likelihood of thunderstorms occurring in the coming hours would be tremendously helpful in deciding if a drone should be launched. It should be noted that although an early warning model for thunderstorms could be beneficial, it is only a part of the required measures to reduce the associated risks. For example, as mentioned by Mary and Gomez, financial constraints often force the people around Lake Victoria to keep working outside, regardless of a thunderstorm occurring (Mary and Gomes 2015).

1.2. Framework

In-situ observations play an important role in the quality of weather and climate services by providing initial conditions for numerical models, validating predictions and offering information on small spatial scales. However, it is generally known that weather stations are sparse in sub-Saharan Africa and the existing ones are often unreliable in communicating their measurements (Dezfuli et al. (2017); van de Giesen et al. (2014)). To tackle this problem, TAHMO was founded in 2014 with the aim to deploy cost-effective and robust weather stations throughout sub-Saharan Africa (van de Giesen et al. 2014). To make TAHMO financially sustainable and to create a large positive impact, the raw weather station data has to be transformed into actionable information. A big step in this direction is the TWIGA project, which is short for transforming water, weather and climate information through in-situ observations for geo-services in Africa (TWIGA 2017). The project runs from 2018 to 2021 and has the objective to develop information services specifically addressed to the needs of the African stakeholders. By involving and promoting businesses, TWIGA not only wants to create sound technical products, but also implement them commercially. One of the TWIGA focus areas is increasing disaster resilience through forecasting and providing early warnings. The development of an early warning model for thunderstorms has the potential to contribute to this area.

1.3. Aim

This research aims to create a completely new early warning for thunderstorm occurrence around Lake Victoria. It will be the first early warning model that uses the TAHMO lightning sensor data. Although the model is new, existing early warning models will be investigated to serve as an inspiration and comparison. The developed model will rely on a data-driven approach instead of explicitly modelling physical phenomena. Three prediction techniques as well as two different sets of model inputs will be tested. Whereas other studies have focused only on classifying thunderstorm occurrence, this research will also attempt to predict the thunderstorm severity. The model will be evaluated on four criteria:

1. Percentage of thunderstorms that are predicted by the model (Hit rate) should be maximized.

2. Percentage of alarms that are false (False Alarm Ratio or FAR) should be minimized.

3. The lead time of the predictions should allow for communication of the warning and taking appropriate safety measures.

4. The prediction should indicate the intensity of the approaching thunderstorm, where large thunderstorms carry more weight.

Based on these criteria the prediction technique and model inputs that give the best performance are selected for a further analysis. To summarize the above, three research questions are posed:

- How do the current thunderstorm early warning models work and perform?

- Which model inputs and prediction technique lead to the best performance of the early warning model based on the four model criteria?

- What are the prediction characteristics of the best performing model?

1.4. Outline

The literature review in Chapter 2 consists of two parts. The first part deals with the basic mechanism of thunderstorm formation around Lake Victoria. This is included to have a general idea of what the model should capture. The second part of the literature review discusses the most relevant early warning models and how they work and perform. These lessons will guide the development of the early warning model and can put its performance in a larger context. Chapter 3 is dedicated to the method of developing the model. It describes the whole procedure from selecting the weather station data to validating the final model. Chapter 4 gives an overview of the results, concretely on how the different prediction techniques and model inputs perform and a more detailed investigation of the best model. In chapter 5 the model is reviewed and discussed. Finally, in chapter 6 the research questions are answered and recommendations are made for future research.

2

Literature Review

With this chapter a foundation will be laid upon which the early warning model can be build. The model aims to predict thunderstorm occurrence based on identifying patterns in past measurements. Understanding how the thunderstorms are formed can provide important insights into the value of the different measured variables. Section 2.1 will show that the basic mechanism of thunderstorm formation around Lake Victoria is differential heating between lake and land, which causes a wind that is forced upward either by convergence or topography. Different studies and figures will be addressed to improve the understanding of this process. There already exist a variety of early warning models for thunderstorm prediction. For this study four models are selected that show the most promise and relevance. They are all relatively recent studies, the oldest from 2007 and the most recent from 2019. One of them uses Lake Victoria as a study area. Two others are included because they solely use weather station data, as is the case with this study. The fourth study is chosen because it utilizes a neural network, a prediction technique that receives a lot of attention in the field of machine learning and artificial intelligence. These models will be covered in Section 2.2, which is also the last section of this chapter.

2.1. Thunderstorm Formation

Lake Victoria is one of the lightning hotspots in the world with over 50 lightning flashes a year per km2 (Mary and Gomes (2015); Albrecht et al. (2016)). This makes it very suitable to test a thunderstorm early warning model. With its surface area of 68,800 km2, over 1.5 times the size of the Netherlands, the lake has a strong influence on the weather. Unequal heating and cooling of the lake and land causes a lake-breeze during the day and a land-breeze during the night. At daytime the surrounding land warms faster than the lake, causing a faster expansion of the land air column that results in an air stream from the land to the lake at high altitudes. This in turn results in high pressure at the lake surface and low pressure at the land surface, which eventually propels an air stream from the lake to the land. At the north eastern part of the lake the moist wind from the lake is pushed upward due to the topography, causing convection and thereby resulting in the formation of large cumulonimbus clouds. The topography of Lake Victoria is shown in Fig. 2.1. During the night the land cools faster which results in the opposite situation. A land breeze converges over the lake, thereby being forced upward and resulting in night-time storms. The land and lake breeze lead to the occurrence of thunderstorms over the lake at night-time and mainly on the eastern and north-eastern surrounding land during daytime. That night-time storms occur over the lake and day-time storms at the north-eastern land, is also verified by 16 years of data from the NASA Optical Transient Detector on board of the OrbView-1 (Albrecht et al. 2016). The study by Thiery et al. (2016), that looked at overshooting tops detected from observations by the SEVIRI instrument, as shown in Fig. 2.2, provides additional confirmation of this process. Overshooting tops are described in the study as dome-like protrusions atop a cumulonimbus anvil and are induced by intense updrafts, thereby providing a proxy for thunderstorm occurrence. Fig. 2.2 shows that the number of overshooting tops is highest over the lake at night and highest on the eastern and north-eastern surrounding land during daytime. Relating this to our research, it means that the TAHMO lightning sensor should also indicate most lightning activity during daytime. That this process is both theoretically well-understood as well as confirmed by the satellite data of overshooting tops, gives reason to believe that thunderstorms

can be potentially well predicted. For example, if the weather station data captures the radiation that leads to surface heating, the low-pressure system as a result of the heating, or the surface wind from the lake to the land, it might also mean that thunderstorms can be predicted. The rest of this study will show if this is indeed the case. Nonetheless it also gives rise to worry. The resulting model might perform well around Lake Victoria but can lack skill to assist the medicine-delivery drones that operate in Rwanda and Ghana. The aim therefore will be to develop a method that can also be applied to other regions but it will be outside the scope of this study to test this.

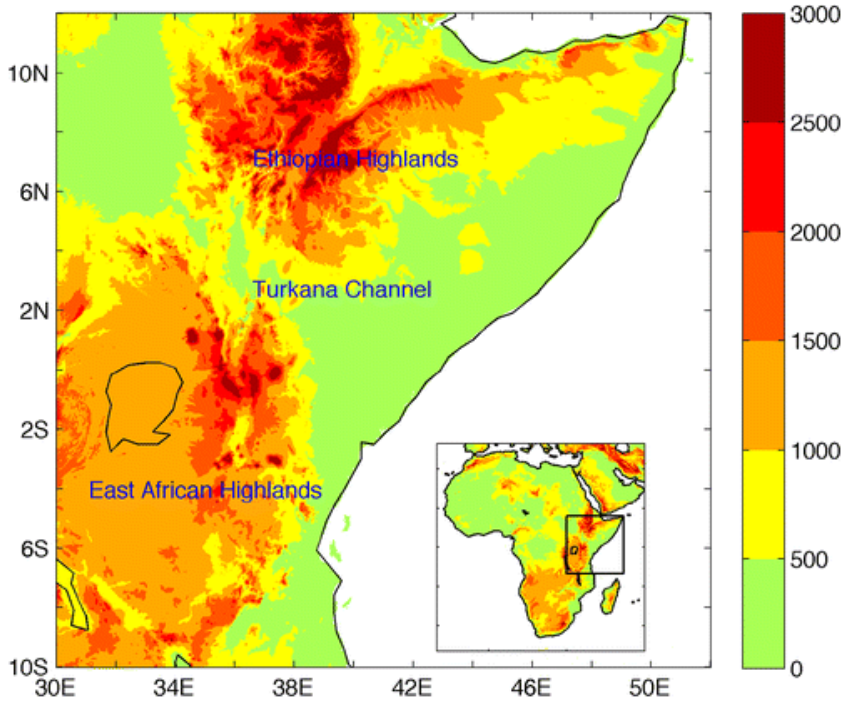


Figure 2.1: Topography of East Africa. Lake Victoria is located at an altitude between 1000-1500 meters and the eastern mountains range from 1500 meters to 3000 meters (Yang et al. 2015)

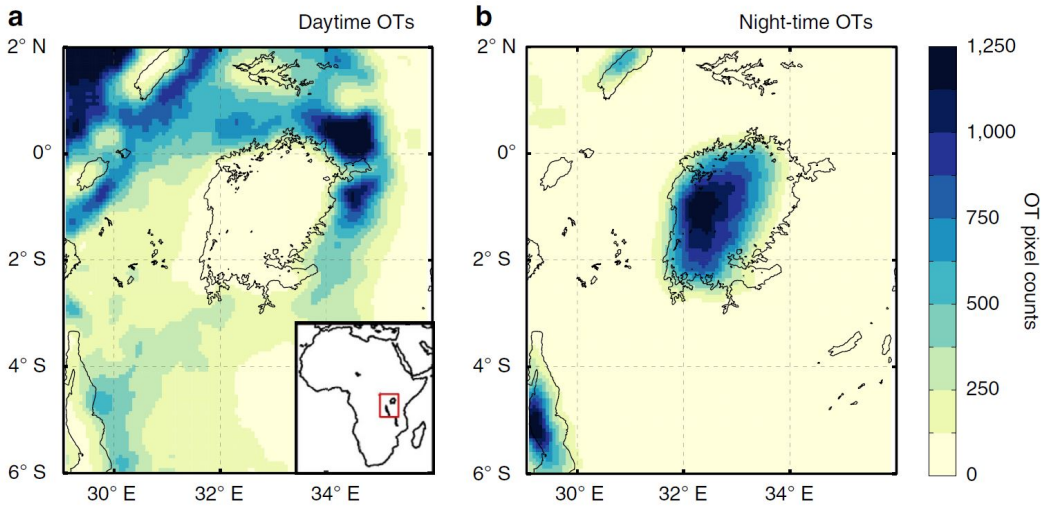


Figure 2.2: Overshooting top detection during 2005-2013, colours show total number of overshooting tops per satellite pixel, over the measurement period (a) from 9:00 to 15:00 UTC and (b) from 0:00 to 09:00 UTC (Thiery et al. 2016)

2.2. Existing Early Warning Models

Developing a new early warning model came from the desire to investigate the lightning sensor data on the TAHMO stations. Moreover, a previous project on neural networks gave the possibility to apply the gained knowledge here. Besides this, the development of the model has to be done from scratch. Looking at the early warning models that already exist and learning from them is therefore essential. As a way to put the early warning model in a more formal setting, it can be characterized as nowcasting. The WMO defines nowcasting as ''forecasting with local detail, by any method, over a period from the present to 6 hours ahead, including a detailed description of the present weather'' (WMO 2017). The four studies that have been selected for this research all fit this definition and will be investigated in this section.

2.2.1. Lake Victoria Model

The first early warning model that is considered is developed by Thiery et al. (2017), and aimed to predict the 1% most intense night-time thunderstorms over Lake Victoria. A logistics regression model was used as the prediction technique. Thunderstorms were characterized by the occurrence of overshooting tops, as identified by an algorithm that takes satellite observations as input. The intensity is defined as the number of overshooting tops over the lake. A nine-year dataset was used originating from the SEVIRI instrument on board the Meteosat satellite. The authors mention that the predictability is derived from afternoon land storms that are often a precursor for night-time storms over the lake. The model target is therefore the number of overshooting tops at night over the lake and the model input the number of overshooting tops during the day over the surrounding land. The model target and model input were aggregated so that there is only one value per night and day. Only the 1% highest nighttime overshooting tops were considered, and a threshold probability is used to balance the hits and false alarms. A lead-time of three hours was used, allowing for communication of the warning and taking safety measures. The model achieved an AUC score of 0.93 with a hit rate of 0.85 and a false alarm rate of 0.13. The hit rate showed how many of the storms that occurred were predicted by the model. Although the model showed significant skills by capturing 28 out of 33 extreme events in the period between 2005 and 2013, the study also mentions that a false alarm is issued almost once a week. Considering that only 28 events occurred, this could be seen as problematic. However, the study also mentioned that most false alarms are still associated with storms, yet with a lower intensity than the 99th percentile. This seems to indicate that the large number of false alarms were related to predicting the 1% most intense thunderstorm. However, it might also be caused by the linear prediction technique or by a lack of model inputs. Since the new model should improve upon this model, non-linear prediction techniques will also be tested. Moreover, were this study only uses one parameter as input, namely overshooting tops over the surrounding land during the preceding day, the new model will use more input variables.

2.2.2. XGBoost Model

The study by Mostajabi et al. (2019) developed a lightning nowcasting model using the XGBoost algorithm (an ensemble of classification trees) with only air pressure, air temperature, relative humidity and wind speed as input variables. The model used the observations from a single weather station to predict the occurrence of a lightning strike. The output was validated by a lightning location system that captured both cloud-to-ground and cloud-to-cloud lightning. The aim of the model was to classify if lightning will occur in the next 10 minutes, 10 to 20 minutes or 20 to 30 minutes. The model achieved a high accuracy with a hit rate of 0.83 and a false alarm ratio of 0.03, for a lead time of 10 minutes. Both a persistence model (hit rate of 0.74 and false alarm ratio of 0.26) and a CAPE model (hit rate of 0.56 and false alarm ratio of 0.86) were outperformed. The performance at lead times of 10-20 minutes and 20-30 minutes were similar with a hit rate of 0.84 and false alarm ratio of 0.04, and a hit rate of 0.83 and false alarm ratio of 0.05, respectively. For this study a dataset between 2006 and 2017 was available and the lightning was predicted within a 30 km radius of the station. This study shows interesting similarities since it only uses weather station data as input, as will also be the case for our model. However, this model only uses weather station data at the current time steps and no information from previous time steps is used. With our model it will be tested if using past information will add to the model's skill. The reasoning behind this is that by for example knowing how the pressure has developed in the past hours, the development of a low-pressure system can be identified. Or an

increase in temperature might point to strong solar radiation that would create an unstable atmosphere prone to formation of thunderstorms. This model uses short lead times, aimed at warning people for lightning strikes, for example to make large gatherings safer. The model that will be developed in our study serves a wider purpose and will therefore aim at a larger lead time. It is interesting to see the low false alarm ratio compared to the Lake Victoria model by Thiery et al. (2017). This might be caused by the shorter lead time (10 minutes compared to 3 hours). The high accuracy and low false alarm ratio also shows the potential of the XGBoost algorithm, which will therefore be selected as one of the prediction techniques in our study.

2.2.3. Neural Network Model

A study from Northern Italy developed a short-term thunderstorm prediction model using an artificial neural network (Manzato 2007). A dataset from 1995 to 2002 was used to train the neural network, and from 2002-2004 to validate the results. The model used balloon sounding-derived indices as model input and the occurrence of at least three cloud-to-ground lightning strikes in the next six hours as output. The best model used a neural network with nine inputs and six hidden neurons. For the validation set a hit rate of 0.89 was achieved and a false alarm ratio of 0.60. The strength of this study is that they used an extensive process to identify the best nine features from a set of 55 variables. A forward selection algorithm was created that identified the best features progressively, starting from one feature and adding single features until no further improvement was achieved. Our study will not use the same process but it will use a more extensive search for the right model inputs, as for example compared to the Lake Victoria and XGBoost study. Moreover, our study will also use a neural network as one of the prediction techniques, to see how it compares with the XGBoost algorithm and a linear model. The authors of this study warn for quick overfitting of the training data, something that should also be avoided in our model.

2.2.4. Weather Station Model

The same research group as the neural network study also developed a model to forecast storm occurrence only using station-derived features (Pucillo and Manzato 2013). A group of features was selected from 33 weather stations that measure relative humidity, temperature, wind direction, wind speed, wind x component, wind y component, heat transport, and moisture transport. The dataset comprised 10 years, from 2000 to 2010. A linear discriminant analysis was used to predict the probability of occurrence of a 40 dBZ and 50 dBZ vertical maximum intensity reflectivity threshold as measured by a radar. The lead time in this case was set to 180 minutes. For the 40 dBZ threshold a hit rate of 0.71 was achieved and a false alarm ratio of 0.31. For the 50 dBZ threshold the hit rate was 0.77 and the false alarm ratio 0.29. This study is interesting because it uses weather station data and achieves a reasonable accuracy and false alarm ratio. The two studies that use weather station data achieve a lower false alarm ratio (0.03 and 0.29) compared to the studies that use satellite data and balloon-soundings (0.94 and 0.60). This seems counter intuitive because the formation of thunderstorms is largely a vertical process and this is not captured by the weather station. Perhaps it is the case that the larger number and variety of variables from the weather station makes up for the lack of vertical information.

2.2.5. Summary of Models

Table 2.1: Summary of the four nowcasting models investigated for this study. * please see the text for how this number was calculated.

Area	Data source	Data length	Technique	Predicted variable	Lead time	Hit rate	False alarm ratio
Lake Victoria	Satellite	2005-2013	Logistic regression	1% highest nighttime overshooting tops	3 hours	0.85	0.94*
Switzerland	Weather station	2006-2017	Ensemble of Trees	Occurrence of lightning strikes	10 minutes	0.83	0.03
Northern Italy	Balloon sounding	1995-2004	Neural Network	>3 lightning Strikes	6 hours	0.89	0.60
Northern Italy	Weather station	2000-2010	Linear discriminant	>50 dBZ radar intensity reflectivity	3 hours	0.77	0.29

The above-mentioned studies are summarized in Table 2.1. Since the false alarm ratio was not given in the Lake Victoria study it was estimated based on 28 true positives and 468 false positives. The 468 is derived from the statement that there was a false alarm almost every week in the 9-year period. Because it is not directly stated the value should be checked with the authors of this study.

3

Methods

In this chapter the method of how our early warning model is developed will be discussed. A closer look at the available data is presented in Section 3.1. Although more TAHMO stations are available, this study uses only one. The selected station is checked for outliers and with linear interpolation measurement gaps are filled. The data is divided into a training set of two years and a validation set of one year. At the end of this section the TAHMO lightning sensor is described. The next step is to decide what the model should actually predict, which is done in Section 3.2. This section also discusses different thunderstorm characteristics retrieved from the data. In Section 3.3 the procedure for selecting the model inputs is explained. Briefly, it is based on the Spearman correlation and distinctions in probability density functions between thunderstorms and no storms. With the model target and model inputs known, the prediction techniques that are tested in this study are presented in Section 3.4. A neural network, a linear regression model and the XGBoost decision tree ensemble are all briefly discussed. In Section 3.5 the model set-up is discussed and how it relates to the conventional set-up. Section 3.6 describes how the model is trained and as a final step validated. The chapter ends with a description of the software that is used in Section 3.7.

3.1. Data

Fig. 3.1 shows the available TAHMO stations on the Kenyan side of Lake Victoria. This area is one of the most active in terms of lightning strikes of all the land surrounding Lake Victoria, as was shown in Fig. 2.2. Moreover, the density of stations is very high, making this a suitable area for research. Each station measures pressure, lightning distance, lightning strikes, precipitation, incoming shortwave radiation, relative humidity, temperature, wind direction, wind gusts, and wind speed.

As a first step in creating the early warning model one of the available stations is chosen. Using multiple stations to validate the results or capture more information, could lead to better results, but this is left for future research. To determine the most suitable station the length of the data record and the percentage of missing values is analysed. All the prediction techniques rely on finding patterns in the past measurements, and the longer the data record, the more patterns become available and the more reliable they are. The missing values are important, because when even one of the measurements is missing, the time-interval cannot be used as input for the model. From the 20 stations available, TA00173 is chosen due to the best combination of data record length and missing values. The lighting sensor on the TA000173 was installed on 12/05/2017, which is also the beginning of the usable data record. The station measures 10 variables in a five minutes resolution. 2.7% of the 5-minute time intervals for this station has missing values for at least one of the 10 variables. It is decided to disregard the wind gust measurements because the firmware version makes it prone to false measurements during heavy precipitation. An overview of the characteristics of TA000173 are given in Table 3.1.

Before the data is used for the early warning model, some pre-processing is conducted. The data is first checked for outliers. A maximum radiation of 1262 W/m2 is found, and although high, it is eventually accepted given the location near the equator and the altitude of 2020 meters. All other values are in order and no outliers have to be removed. The number of missing values in the data is reduced by gap-filling. 10028 time-intervals had at least one missing value, this corresponds to about 35 hours

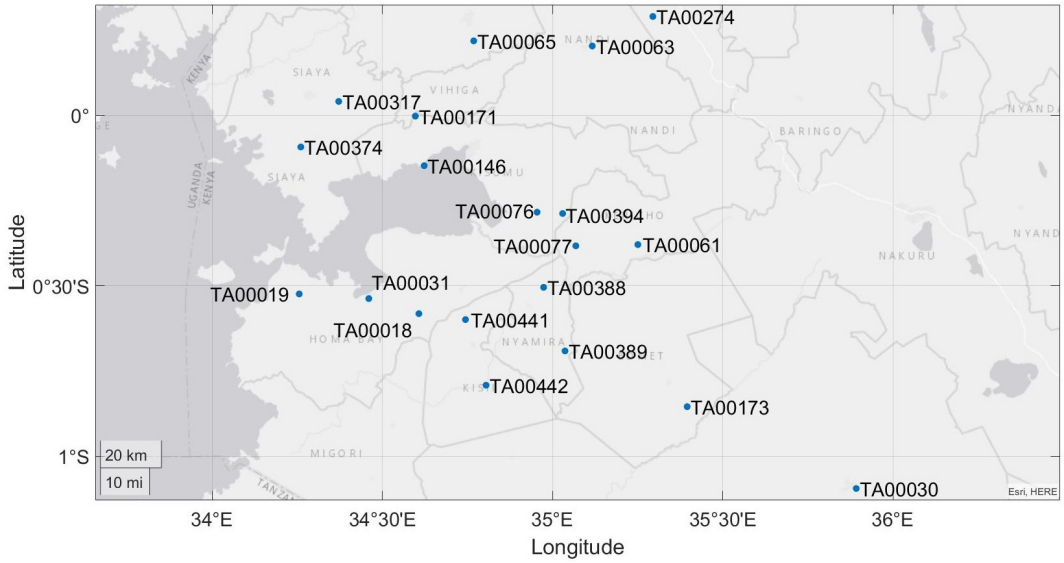


Figure 3.1: Operational TAHMO stations on the Kenyan side of Lake Victoria.

Table 3.1: Characteristics of the TA000173 weather station. *The wind gust data is disregarded in this study.

Location	Longisa High school, Bomet Highway, Kenya
Installation height (m)	2
Latitude (°)	-0.85472
Longitude (°)	35.39565
Elevation (m)	2020
Data record	2017/05/13 - Present
Observation Variables	Atmospheric Pressure (kPa), Lightning Distance (km), Lightning Strikes (-), Precipitation (mm), Radiation (W/m2), Relative Humidity (-), Temperature (°C), Wind Direction (°), Wind Gusts (m/s)*, Wind Speed (m/s)
Temporal resolution (minutes)	5

in a three-year period. Most of them occurred in the wind speed and wind direction measurements. Missing values are filled in using linear interpolation with a maximum gap length ranging from 5 minutes for the lightning distance to 285 minutes for the relative humidity. The gap length is based on the autocorrelation, whereby the maximum gap is set at a correlation of 0.5. This procedure reduces the number of missing time-intervals to 8793, which corresponds to a reduction of 4 hours. At this point the data is divided into a training set and a validation set. With a relative short data length it is important to keep the training set as long as possible. However, to test the skill of the model a representative validation set should also be available. To fulfil the second condition, it was decided that the validation set should contain a full-year. By doing this all the seasons are contained in the final test of the model's skill. This results in a two-year training dataset from 2017-06-01 until 2019-05-31 and a one-year validation dataset from 2019-06-01 until 2020-05-31. For all the remaining calculations in this chapter the training data is used exclusively to avoid any bias towards the validation data. Since the validation data is unseen by the model, it comes closest to testing the model in real-life and the validation results will therefore carry the largest weight.

3.1.1. Lightning Sensor

The TAHMO stations contain a Franklin Lightning Sensor that measures the electromagnetic waves produced by lightning strikes (AustriaMicrosystems 2012). An algorithm detects if the radiation originates from lightning, both cloud-to-ground and cloud-to-cloud, or from man-made sources. To reject man-made sources a threshold is defined, the level of which can be altered. Once the signal is validated as

being a lightning strike, the energy of the radiation serves as the input for the distance calculation. This is based on the relationship that the radiated energy is inversely proportional to the distance squared (Kamau et al. 2015). The output of the sensor is therefore the number of lightning strikes and an estimation of the distance from the station to the lightning strike. The closest distance the sensor is able to measure is 1 km and the furthest 40 km. An important observation here is that the weather variables have different characteristic spatial scale which might lead to problems for the model in identifying patterns. For example, pressure works on a larger scale than 40 km. This could mean that the pressure measurement indicates an approaching storm but when it occurs further than 40 km away, the lightning sensor might not detect it. In this situation a false alarm would be issued. The same could apply to temperature, relative humidity, and wind direction. It could also be the other way around. The station might not measure precipitation although there is a storm occurring, just not directly above the station. Although this could give problems, the hope is that with the right combination of inputs the model will be intelligent enough to identify these situations.

3.2. Model Target

The purpose of the model is to predict an approaching thunderstorm with a sufficient lead time, as well as indicate the intensity of the storm. The model target is set to the sum of lightning strikes in the next six hours. This model target is predicted at a 5-minute resolution. The six hours are selected to maximize the lead time, while still fitting within the WMO definition of nowcasting as given in Section 2.2. Section 3.5 explains how the model target relates to the classification and regression prediction of individual thunderstorms. This section will focus on gaining understanding about the thunderstorm characteristics.

3.2.1. Thunderstorm Characteristics

Within the two-year training dataset, 422 thunderstorms are identified. A single thunderstorm is defined here as the period of time between the occurrence of a lightning strike until there has not been a lightning strike for one hour. This means that multiple thunderstorms can occur during a day, as long as there is a period of at least one hour between them without any lightning strikes. Fig. 3.2 shows during which months the thunderstorms occur and the time of the day when they occur.

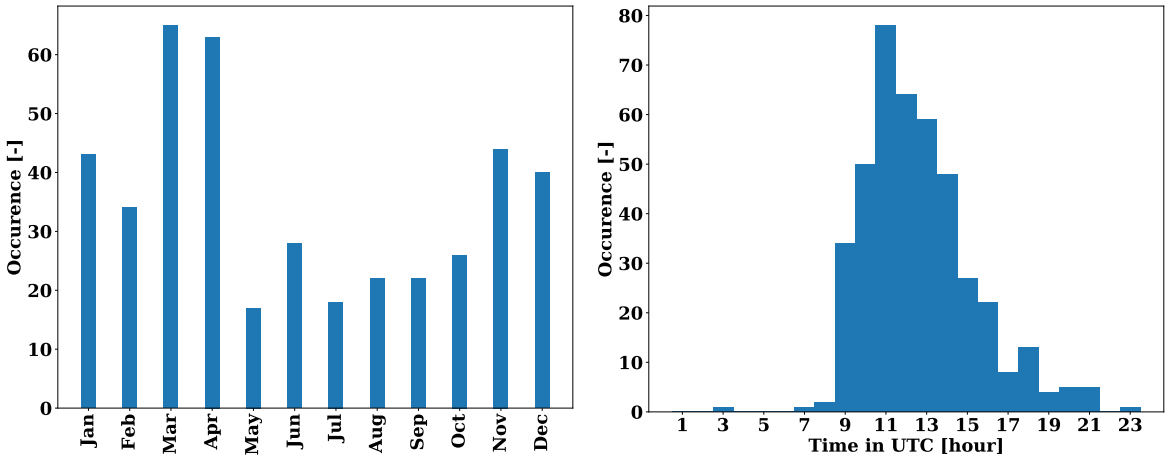


Figure 3.2: The plots show how the 422 thunderstorms of the training dataset are distributed over the months of the year and over the hours of the day.

It can be seen that thunderstorms happen all year round but there is a peak around March and around November. That the thunderstorms happen throughout the year is due to the lake breeze which is not dependent on the seasons. The peaks are explained by the movement of the Inter Tropical Convergence Zone, which causes seasonality due to changes in solar irradiation. This leads to the East African rainy seasons that take place in October-November-December and March (Gong et al. 2016). As was discussed in Section 2.1, during the day a lake breeze is forced upward by the topography and eventually can cause a thunderstorm. Fig. 3.2 shows that thunderstorms occur almost exclusively

during the day. This makes the lightning measurements of the TAHMO station in agreement with the theory. 79% of the thunderstorms, 333 out of 422, occur between 9:00 and 15:00 UTC. Of the 422 thunderstorms only two occur between 22.00 and 06:00 UTC.

Fig. 3.3 shows how the intensity and the duration of the thunderstorm is distributed. Most thunderstorm have less than 100 lightning strikes. There are 89 thunderstorms (21%) with only one lightning strike and 228 thunderstorms (54%) have six or less lightning strikes. There are 45 thunderstorms (11%) with over 100 lightning strikes. The largest thunderstorm counts 599 lightning strikes and lasts 140 minutes. Fig. 3.3 shows that most thunderstorms last only 5 minutes, namely 106 thunderstorms (25%). 238 thunderstorms (56%) last less than 60 minutes. The longest thunderstorm lasts 330 minutes with 383 lightning strikes.

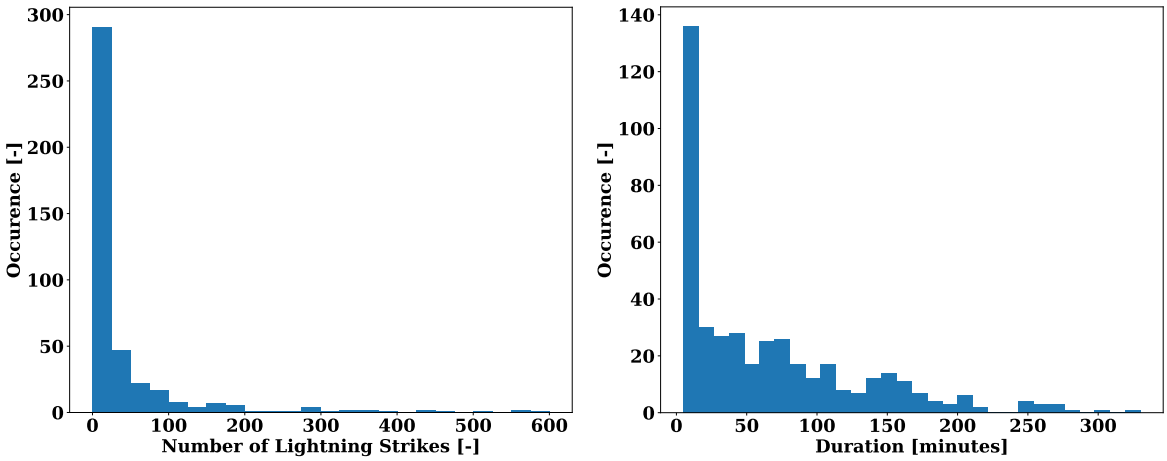


Figure 3.3: These plots show the intensity of the thunderstorm, defined by the sum of lightning strikes, and the duration of the thunderstorms. The duration is the time period between the first lightning strike and the last lightning strike of the thunderstorm. A thunderstorm ends when there has not been a lightning strikes for one hour.

3.3. Model Inputs

The model target is the rolling sum of lightning strikes over the next six hours. To be able to predict this it is important that the model has meaningful inputs. The goal of this section is to identify inputs that are precursors for approaching thunderstorms. The inputs that will be considered are the 5-minute intervals, the standard deviations, the gradients and the daily values of the measurements. The 5-minute intervals show the current state of the atmosphere and could reveal valuable precursors, such as a wind direction from the lake. The standard deviations are investigated because they indicate changes. For example, a fluctuating radiation might point to the formation of clouds, necessary for a thunderstorm to form. The standard deviation is calculated over the past three hours. Gradients point to a drop or increase of a variable. For example, a negative pressure gradient could indicate the formation of a low-pressure system and therefore rising air. The gradients are calculated over the past three hours. Finally, the values of the past 24 hours are also used. The reasoning behind this is that what happened in the last 24-hours has an effect on the coming hours. For example, if there has been precipitation in the past 24-hours the likelihood of a thunderstorm could be higher because there is more moisture available. For the lightning strikes and precipitation the 24-hour sum is used, and for the wind speed the 24-hour maximum. For the other variables the 24-hour mean is used.

With this method there are 36 possible model inputs, 9 for each of the four categories. To select the most meaningful inputs and avoid redundant inputs, the correlations and probability density functions will be investigated. The final selection is made from combining both sets of inputs and removing duplicate and redundant features. To understand if this method provides added skill to the model, it will be compared to using the raw 5-minute interval measurements as model inputs.

3.3.1. Spearman Correlation

Both Pearson and Spearman correlations are tested, and it is found that Spearman gives higher correlation values and also involves skewed distributions such as precipitation and wind speed. For all

four categories, namely the 5-minute intervals, standard deviations, gradients, and daily values, the Spearman correlation is calculated, both between the inputs and with the model target. An input is selected if its correlation with the model target is at least |0.15|. If an input has a correlation higher than |0.75| with another input, only the input with the highest model target correlation is selected. Fig. 3.4 shows the correlation matrix for the 5-minute intervals. Based on the values shown in this figure, the lightning distance, lightning strikes, radiation, relative humidity, temperature, and wind speed are selected. Lighting distance is removed because it has a correlation of 1 with the lightning strikes. Relative Humidity is removed because it has a correlation of -0.78 with temperature and its correlation with the model target is lower. This procedure is repeated for the standard deviations, gradients and daily values. The remaining three correlation matrices can be found in Appendix A.

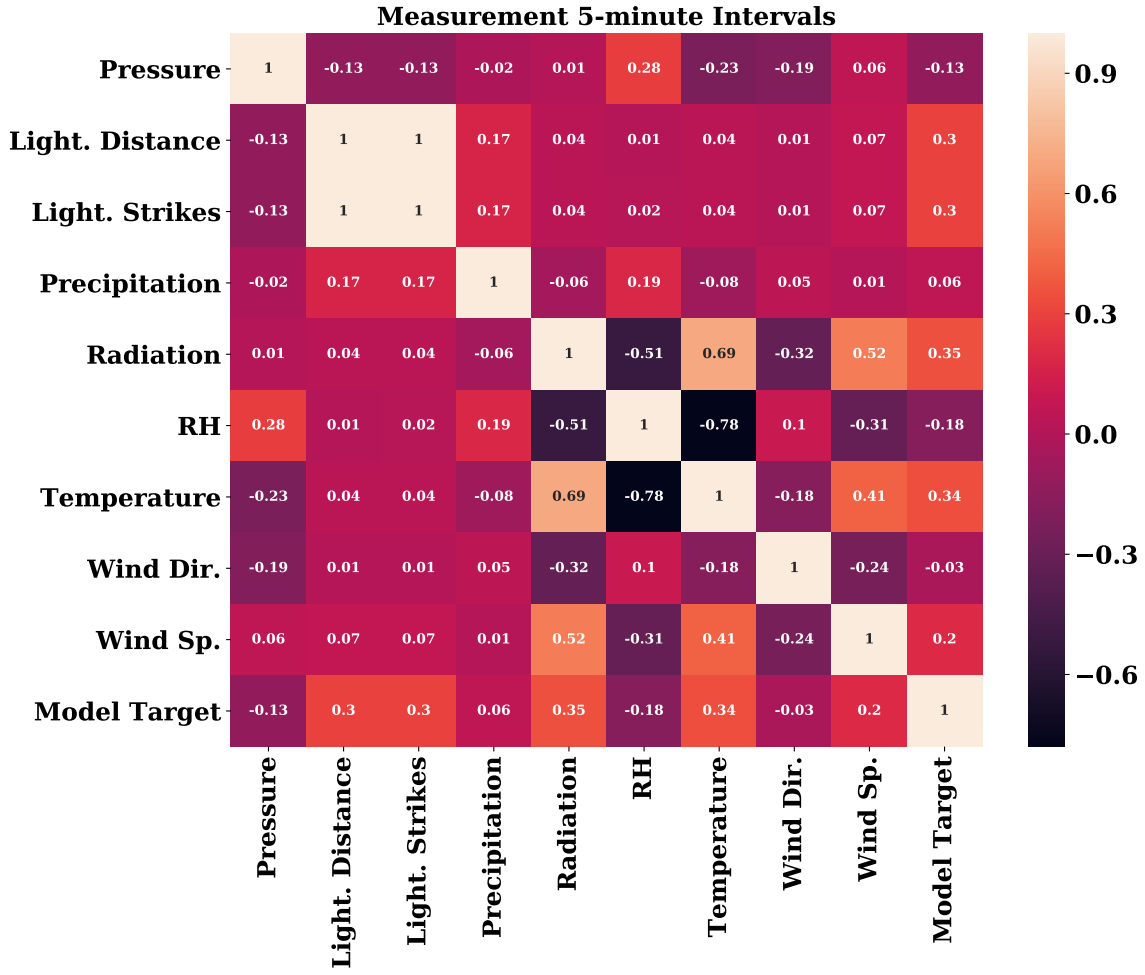


Figure 3.4: Spearman correlation of the 5-minute interval measurements with each other and with the model target. The model target is the rolling sum of lightning strikes in the 6 hours following each observation.

3.3.2. Probability Density Functions

The idea behind using probability density functions (PDF) to select model inputs is that there could be a different distribution when a storm is approaching and when there is no storm approaching. For each measurement a PDF is created with two distributions. The first distribution contains the measurements for which a thunderstorm will occur in the next six hours. The second distribution contains measurements for which a thunderstorm will not occur in the next six hours. If there is a visible difference between the two distributions the model input is selected. Fig. 3.5 shows four probability density functions, one from each of the four categories. The other probability density functions can be found in Appendix A. It can be seen that if there is a thunderstorm in the next six hours the temperature is generally higher. This can be explained by considering that thunderstorms happen during the day

when temperatures are higher. Moreover, a higher surface temperature can lead both to a stronger lake breeze and more convection. When a thunderstorm is approaching the standard deviation of the wind direction, indicating changes in wind direction, tends to be higher. In general the wind is coming from the east at this latitude, the so-called trade-winds, but a lake breeze which is associated with a thunderstorm comes from the west or south-west. Fig. 3.5 also shows that a storm is preceded by a negative pressure gradient, corresponding to rising air and thereby increasing the changes of a thunderstorm. Finally, a higher relative humidity in the past 24-hours is associated with thunderstorms. A simple explanation for this could be the higher moisture availability, increasing the chances of cloud formation.

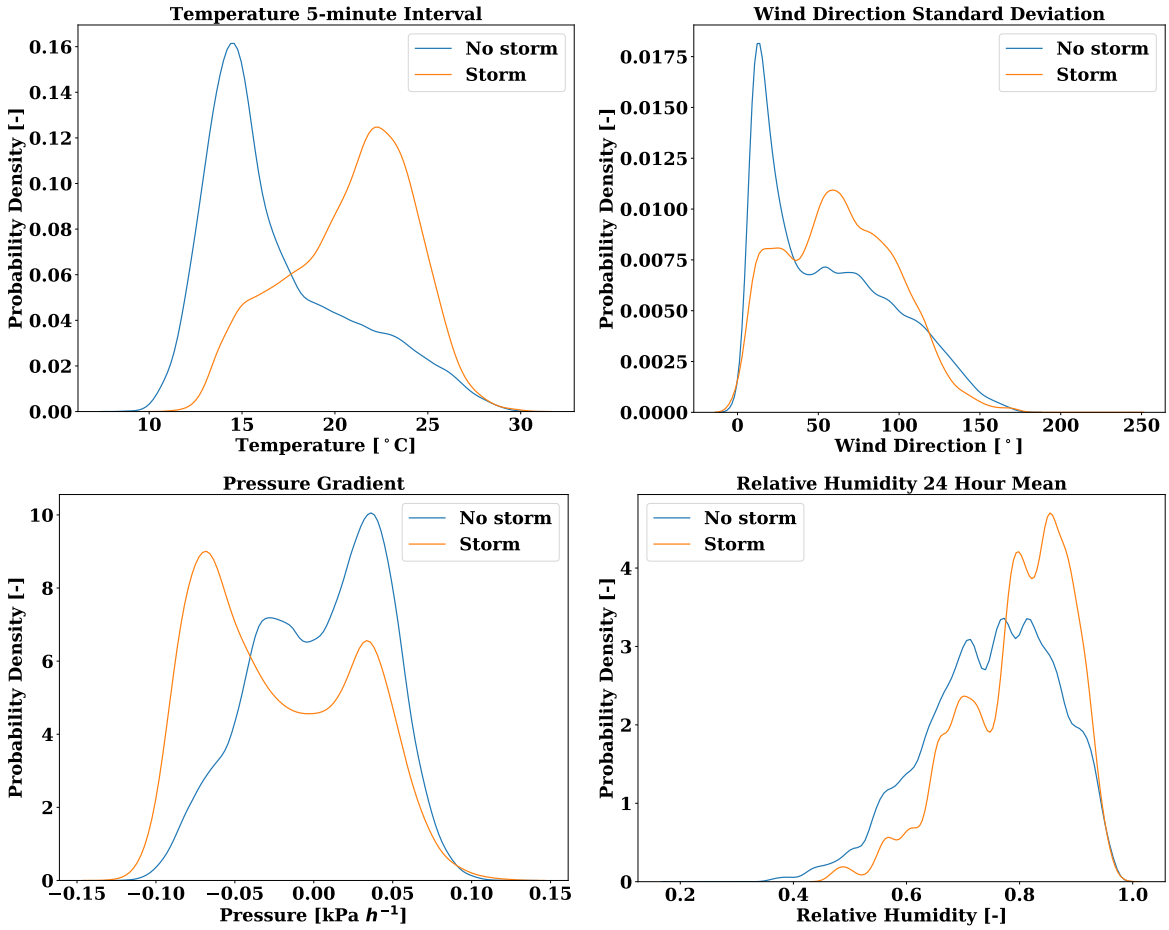


Figure 3.5: Probability density functions for thunderstorm versus no thunderstorm. If the model target is above zero, meaning that there will be lightning strikes in the next six hours, the value is used for the storm distribution. Similarly, if the model target is zero, the value is added to the no storm distribution. One variable has been selected as an example from each of the four categories. The other Probability density functions can be found in Appendix A.

3.3.3. Final selection

By combining both methods a final selection of model inputs is made. The Spearman correlation gives 24 potential model inputs and the probability density functions 16. When the duplicate and redundant inputs are removed 19 model inputs remain. Table 3.2 gives an overview of this, where the x indicates inputs selected by the Spearman correlation and o by the PDFs. The variables in red are removed because there Spearman correlation with another variable is higher than 0.75.

3.3.4. Raw Inputs

To understand if the above procedure gives any benefit in improving the model's predictions, a second model input selection will also be tested. The Switzerland early warning model uses only pressure, temperature, relative humidity and wind speed at the current time step and achieves a hit rate of 0.83 and a false alarm ratio of 0.03 (Mostajabi et al. 2019). It is therefore not unlikely that only using raw

Table 3.2: This table shows the final selection of model inputs. The x indicates a variable selected based on the Spearman correlation. A o indicates it is selected based on the probability density functions. If a variable is red, it means it is removed because it had a Spearman correlation with another variable higher than 0.75.

Variable	5-minute Intervals	Standard Deviations	Gradients	Daily Values
Pressure	o	x o	x o	o
Lightning Distance	x		x	x
Lightning Strikes	x	x	x	x
Precipitation				x
Radiation	x o	x o		o
Relative Humidity	x o	x o	x o	x o
Temperature	x	x o	x o	
Wind Direction	o	o		x o
Wind Speed	x	x		

measurements at the current time steps leads to good results. This means that 9 model inputs will be used, namely the pressure, lightning distance, lightning strikes, precipitation, radiation, relative humidity, temperature, wind direction, and wind speed at the 5-minute interval level. An advantage of this approach is that there are relatively few model inputs, reducing the chance of overfitting and possibly creating a more robust model. In the results chapter it will become clear if this gives better results than the more extensive selection of model inputs based on the Spearman correlation and probability density functions. For simplicity the model inputs containing the 5-minute interval measurements are called raw inputs, and the model inputs based on the Spearman correlation and PDFs are called knowledge inputs.

3.4. Model Techniques

In Section 3.2 the model target was described as being the sum of lightning strikes in the next six hours. Using the inputs found in Section 3.3 the aim of the model technique is to approach the model target as close as possible. The model technique has to find patterns between the model inputs and model targets based on the training data set and apply these patterns on new model inputs. In this study three model techniques will be compared. The first, and most simple, model will be a linear regression model. A neural network will also be tested as was done in the study by Manzato (2007). Finally, the XGBoost algorithm will be tested, inspired by the study of Mostajabi et al. (2019). This study achieved the lowest false alarm ratio, namely 0.03, and also used weather station measurements as input. An important difference however, is that the study focused on classifying lighting strikes in the next ten minutes. The neural network and XGBoost algorithm are both non-linear techniques, where the linear regression is, as the name states, linear. In this section all three techniques will be covered. In the end a brief explanation is given on how the model parameters have been selected.

3.4.1. Linear Regression Model

The linear technique is a regression model with a linear least squares loss function. To avoid overfitting the model utilizes a regularization function that punishes large fitting constants. Equation 3.1 shows the objective function that the model tries to minimize.Y is the model target, X the model inputs, w the fitting coefficients and alpha a constant parameter that controls the degree of regularization. The training of this model consists in changing the fitting coefficients w until the objective function is minimized.

$$\text{Objective function} = (y - X * w)^2 - \alpha * w^2 \tag{3.1}$$

For the models with knowledge inputs there will be 22 different fitting coefficients and for the models with raw inputs 9 different fitting coefficients. Once these are determined they can be used to calculate the model target by adding the product of the fitting coefficients with their model inputs, as shown in 3.2

$$y = X * w \tag{3.2}$$

The advantage of this model is the simplicity since the only parameter that needs to be tuned is alpha. However, a disadvantage is that it is a linear technique and it might not capture the complex phenomena that lead to thunderstorms.

3.4.2. Neural Network

The first non-linear technique used for the early warning model is a neural network with a forward pass and backward error-propagation. In the forward pass the input data goes through the neurons of the hidden layers. The output of each neuron is determined by the activation function, bias, and weight of that particular neuron. All the values of the hidden layer neurons are passed through the neuron in the final layer, which then outputs the final value. The error of this output value is calculated using a pre-defined loss function. This error is back-propagated whereby the bias and weight of each neuron are adjusted with the gradient descend method to minimize the error. Being a non-linear technique, it is expected that the neural network will capture the patterns in the training data better than the linear regression model. However, as also mentioned by Manzato (2007), the neural network can easily overfit the training data. Another disadvantage of the neural network is the number of parameters that need to be tuned, such as the neurons, epochs, dropout rate, and batch size. Finding the right combination of these parameters takes experience and time, or large computational power so that many different combinations can be tested.

3.4.3. XGBoost

The final technique that will be tested is the XGBoost algorithm. This algorithm is similar to the random forest technique since it is built from individual decision trees. The difference is that the random forest technique averages or uses majority rules to combine the individual trees. The XGBoost algorithm works in a forward manner and adds a new tree to the previous tree to improve its result and is thereby combining the tree's along the way. As with the neural network the challenge with this algorithm is overfitting and finding the right combination of parameters. The main parameters of the XGBoost algorithm are the number of trees, the maximum depth of each tree, and the percentage of inputs used for each tree. An added advantage of the XGBoost model is that some insight can be gained in the relative importance of model inputs. Decision tree's are build from nodes, and a node makes a decision based on a model input. For example a node might make a split between temperatures above and below 20 degrees, where each leads to a different number of lightning strikes predicted. The F score calculates how often a model input has been used to make a decision. The higher the F score the more important a model input is in predicting thunderstorms. The F score of both the raw inputs and knowledge inputs will be discussed in Chapter 4.

3.4.4. Parameter Selection

For each of prediction techniques parameters should be chosen in such a way that the optimal result is achieved. A formal optimization would entail trying out different combinations of all the parameters, which quickly becomes unfeasible. For example, if 20 different options would be tested for five parameters, 3.2 million model configurations would have to be tested. This would require computational power that is not available for this study. To still come up with an acceptable configuration each of the three models have been tested extensively while continuously tuning the parameters. The drawback of this method is that no formal optimum is achieved and moreover, with a slight change of model inputs or model targets, the whole procedure has to be repeated. This means that using a formal optimization method could still lead to better results.

To be complete, the choice of parameters for each model is listed here. For the linear regression model it is found that the results are insensitive to changing alpha and the default value of 1 is kept. A neural network with one hidden layer, 60 neurons, 80 training epochs, a batch size of 5000 and a dropout rate of 0.1 is chosen. The mean squared error is used as loss function, Relu as the activation function and Adam as the optimizer. 200 decision trees are used for the XGBoost algorithm, with a maximum depth of 3 and 30% of the inputs for each tree. The algorithm uses a squared error loss function.

3.5. Classification - Regression Model

One of the challenges faced in this research is that the early warning model is completely new and does not build on any previous work. Four other early warning models have been studied and their lessons learned, but this research is mainly based on trying out many different things, discarding models, and continuing with the most promising ones. The model set-up that will be discussed in the first part of this section is the result of this continuous experimenting and shows subtle differences with the existing early warning models. How the model results will be evaluated and compared with the four existing early warning models will be discussed in the second and final part of this section.

3.5.1. Model Set-up

This subsection discusses the model set-up that is developed for this study. Both the conventional approach, as identified in the other early warning models, and the new approach are discussed. For the new approach to work an algorithm is developed that loops through the measurements and predictions to derive the performance of the model. This algorithm is explained and clarified with a plot.

Conventional Approach

The four early warning models discussed in Section 2.2 use a model target that represents a certain intensity value. For example, the Lake Victoria study uses the 1% highest nighttime overshooting tops and the Northern Italy study events with more than three lightning strikes. The model target is set to 1 when the intensity is achieved and 0 when it is not. The model prediction is then a probability between 0 and 1, and a threshold is selected that optimizes the hits and false alarms. The lead time is fixed by aggregating the model inputs and model targets to the desired time interval. The Lake Victoria study provides an example whereby the daytime overshooting tops are aggregated between 13:00 and 18:00 and the nighttime overshooting tops between 00:00 and 12:00. This then gives a lead time of six hours, namely the difference between 18:00 and 00:00. A setback of this method is that the aggregation reduces the temporal resolution of the measurements. In the Lake Victoria study the overshooting tops are available in a 15-minute resolution but this is lost due to the aggregation. Another disadvantage is that it is not clear when the storm actually occurs because the overshooting tops are summed between 00:00 and 12:00. The storm could have occurred anytime during the 12 hour window.

New Approach

In this study a different approach is tested. The 5-minute resolution is maintained and the model target is set to the sum of lightning strikes in the next six hours. A threshold could be applied to every 5-minute prediction but this would mean that there is a potential alarm every 5-minutes which would not be practical. To avoid this the 5-minute predictions are summed and continuously checked against a threshold. Once the threshold is exceeded an alarm is issued. For example, the model might predict 50 lightning strikes occurring in the next six hours at every 5-minute interval starting from 07:00. When the threshold is 200 an alarm would be issued at 07:20 because the aggregated predictions would be 4 times 50 at this time. An algorithm has been developed that loops over the dataset in 5-minute steps and compares the model predictions with the actual occurrence of lightning strikes.

Algorithm

When a lightning strike occurs the algorithm checks if the model predictions exceed the threshold in the period from six hours up to 30 minutes before the lightning strike. When this is the case a true positive is counted. The lead time is then calculated as the time difference between the first lightning strike and when the alarm is issued. False negatives are counted when the threshold has not been exceeded at least 30 minutes before the first lightning strike. The 30 minutes value can be changed depending on the minimum time that is required for taking safety measures. After the first lightning strike the algorithm monitors the duration of the storm and signals the end of a storm when there has not been a lightning strike for one hour. This is to avoid issuing multiple alarms for the same storm or not issuing an alarm for a new storm. After the one hour the analysis is continued and a new true positive or false negative can be counted. The storm intensity is also stored by averaging the model predictions in the six hours preceding the first lightning strike. This is then compared to the number of lightning strikes that have actually occurred during the storm. For example, a storm occurs between 15:00 and 16:00. The average model prediction from 09:00 until 14:55 equals 80 lighting strikes and the threshold is exceeded at 11:00. In this case a true positive would be counted with a lead time of

four hours and an intensity prediction of 80 lightning strikes. The 80 lightning strikes will be compared to the number of lightning strikes that are actually measured between 15:00 and 16:00.

For the true positives and false negatives the algorithm takes a lightning strike as reference point. To determine the false positives and true negative this is not possible. To still be able to count these the predictions are summed over a six hour period. A false positive is counted when the threshold is exceeded during the six hour period and there is no lightning strike during these six hours or the next six hours. A true negative is counted when the threshold has not been exceeded during the six hour period and there is no lightning strike in these six hours or the next six hours. For example, a false positive is counted when the aggregated predictions exceed the threshold between 06:00 and 12:00 but there is no lightning strike between 06:00 and 18:00. Two six hour blocks are considered because the threshold might be exceeded at the end of the first block and a thunderstorm could occur during the second block. With this method only one false positive or true negative can be counted every six hours.

Figure 3.6 gives a visual example of the model set-up in the case of a true positive. The threshold can be defined by the user and will be optimized using k-cross validation, as will be explained in Section 3.6.

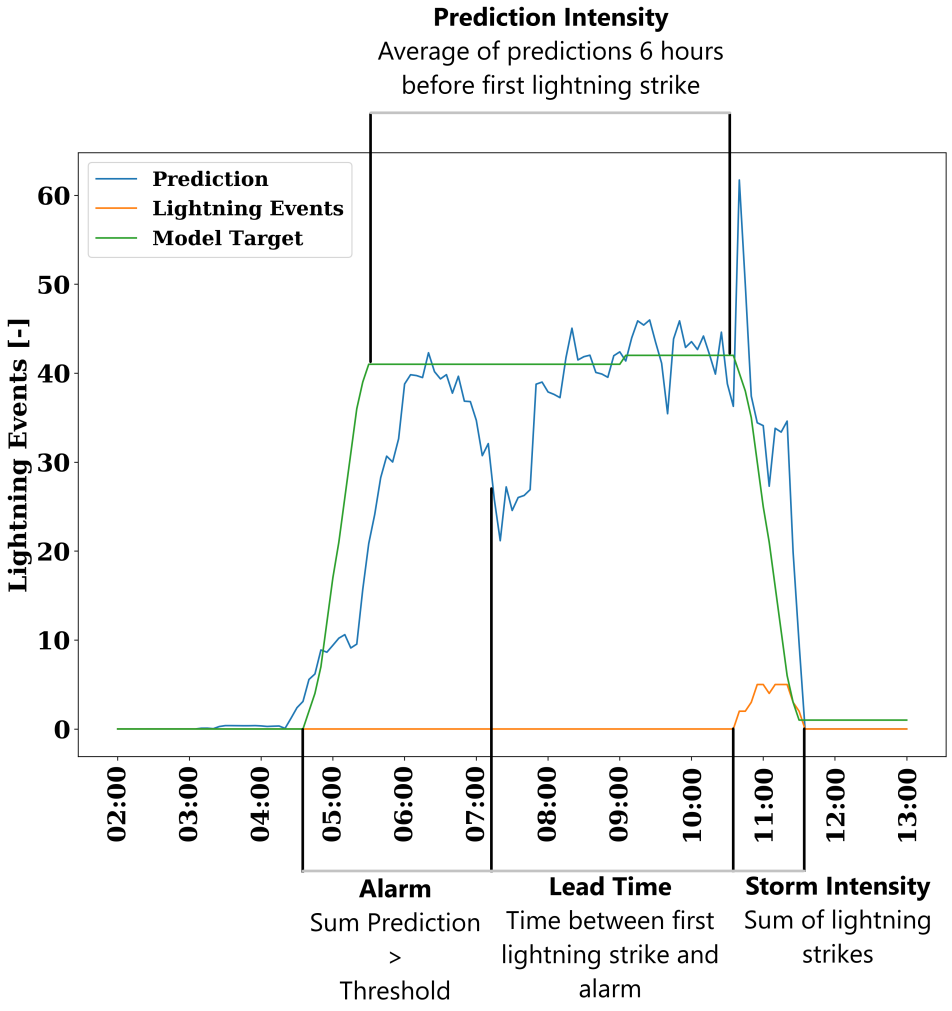


Figure 3.6: This figure shows how the model set-up works with a true positive as example. The blue line in the graph indicates the predictions made by the model, namely the predicted number of lightning strikes in the next six hours. The green line is the model target, namely the actual sum of lightning strikes in the next six hours. The orange line shows the lightning strikes that are measured by the sensor in real time. For this example the threshold is exceeded at 07:10 but this is arbitrary and depends on the value of the threshold.

3.5.2. Evaluation and Comparison

The final outputs of the algorithm are the true positives, false negatives, true negatives, false positives, lead time, and intensity prediction. The six configurations can be compared based on these values because they all deal with the same events. To make a comparison with models from other studies however, the values have to be normalized. For this purpose the hit rate and the false alarm ratio are calculated. The hit rate (eq. 3.3) is the percentage of storms that have been predicted by the model. The false alarm ratio (eq. 3.4) is the percentage of issued alarms that are false.

$$\text{Hit rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.3}$$

$$\text{False alarm ratio} = \frac{\text{False Positives}}{\text{True Positives} + \text{False Positives}} \tag{3.4}$$

A hit rate of 1 means that all thunderstorms are predicted. A false alarm ratio of 0 means that all the issued alarms are followed by thunderstorms. The difference between hit rate and false alarm ratio (H-F) gives a good indication of the classification skill. A H-F of 1 indicates that all storms have been predicted and that there have been no false alarms. A H-F value of -1 indicates that no storms have been predicted and that the model only issued false alarms. Other studies have also used the false alarm rate (eq. 3.5), which indicates the percentage of false alarms over all the times without thunderstorm.

$$\text{False alarm rate} = \frac{\text{False Positives}}{\text{True Negatives} + \text{False Positives}} \tag{3.5}$$

For sparse targets this value seems less suitable since a model with a low false alarm rate can still issue many false alarms. For example, the Lake Victoria model by Thiery et al. (2017) predicted 28 thunderstorms but had a false alarm almost every week over a period of 9 years. This leads to a false alarm ratio of about 0.94 whereas the false alarm rate was 0.13. As stated by Barnes et al. (2009) the false alarm rate and false alarm ratio are often confused. It was found that 10 out of 26 studies used the wrong definition.

It should be noted that the hit rate and false alarm ratio might give an indication of how the skill of the models compare, it is only partially useful because each model has different areas, data lengths, lead times, model inputs, model targets, etc.

The model developed in this study also predicts the intensity of the storm. For all the thunderstorms that have been classified the skill of the intensity prediction will be evaluated using the absolute error (eq. 3.6) and the relative absolute error (eq. 3.7).

$$\text{Absolute Error} = |\text{Prediction Intensity} - \text{Storm Intensity}| \tag{3.6}$$

$$\text{Absolute Relative Error} = \left|\frac{\text{Prediction Intensity} - \text{Storm Intensity}}{\text{Storm Intensity}}\right| \tag{3.7}$$

Both metrics are evaluated because an absolute error of 100 lightning strikes is more severe for a thunderstorm with 10 lightning strikes than a thunderstorm with 600 lightning strikes. The absolute values are considered so that under and over estimations do not cancel each other out when taking the average over all storms. In Chapter 4, the hit rate, false alarm ratio, absolute error, and absolute relative error are given for all combinations of model techniques and model inputs. The prediction intensity skill cannot be compared to the other models because they only deal with classification.

3.6. Training & Validation

Now that the model target, model inputs, model techniques, and the model evaluation are known, the remaining steps are to find the optimal threshold and finally to validate the model using previously unseen data. As explained in Section 3.5 the value of the threshold influences the number of true positives, false negatives, true negatives, and false positives, as well as the lead time, and should be selected for the six configurations. Each configuration, namely the two input sets (raw inputs and knowledge inputs) and three techniques (linear regression model, neural network, and XGBoost), will have its own optimal threshold and validation results. As mentioned in Section 3.1 the data has been

divided into a training set (2017-06-01 until 2019-05-31) and a validation set (2019-06-01 until 2020-05-31). For an unbiased evaluation of the model skill the validation data is kept separately and the training dataset is used to determine the optimal thresholds. However, the optimal threshold also needs to be determined in an unbiased manner and therefore the training dataset is further divided into training and testing days. This division is done using k-cross validation. Four folds are created, each containing a different quarter of the set as testing days and the remaining three quarters as training days. The testing days are uniformly distributed over the training dataset. For each fold a different model is trained and evaluated, according to the procedure shown in Fig. 3.7.

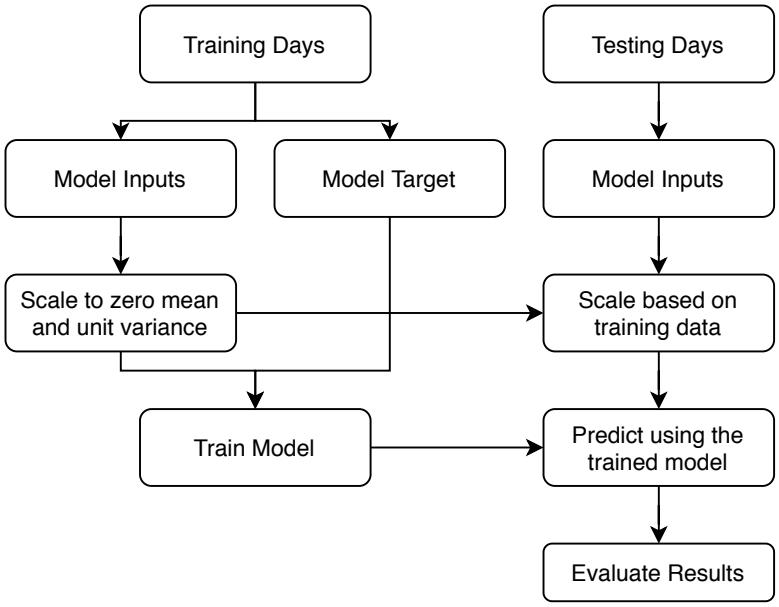


Figure 3.7: This figure shows a schematic overview of how the models are trained and evaluated. The procedure is the same for all three predictions techniques.

The model inputs are scaled to zero mean and unit variance. This is done because the input variables have different units and their absolute values should not influence their importance. Because the results should not be biased towards the testing days, the scaling factors are calculated based on the training days and applied to both training and testing days. The model is then trained by fitting the model inputs to the model target. Once this is done the testing days can be predicted and its performance evaluated. The hit rate, false alarm ratio, and average lead time are calculated by combining the testing days of all four folds. This is repeated for different thresholds and the threshold with the highest value of H-F is selected as the optimum. Thresholds ranging from 0 to 1000, in steps of 50, are evaluated. This threshold does not represent the intensity of a storm but rather the 5-minute aggregated predictions. With the optimum threshold known the model is retrained and reevaluated using the training data as training days and the validation data as testing days.

3.7. Software

The programming environment used for this study is Python version 3.6.8. For the data handling the Pandas and NumPy packages are used. Plotting is done using the Matplotlib package. Each model technique uses a different package. The linear regression model comes from the Scikit-learn machine learning package and is called Ridge. The XGBoost package is used for the ensemble of decision trees. Finally, for the neural network the Keras package is used, which is built upon Google's TensorFlow. All packages are freely available and have many online resources available such as examples and troubleshooting guides.

4

Results

In this chapter the results of the study are presented. Section 4.1 deals with the selection of the optimal threshold for each of the six configuration based on the value of H-F and the lead time. Section 4.2 gives the classification, lead time and regression results for all six configurations based on the validation dataset. Moreover, to gain some understanding of how the thunderstorms are predicted, this section also looks at the relative importance of each model input. This is done using the F score from the XGBoost model. Finally, in Section 4.3 the best model is chosen based on the four criteria listed in Section 1.3 and a further analysis is conducted on its performance.

4.1. Training

The level of the threshold determines how sensitive the model is and influences the hit rate, false alarm ratio, and lead time. The balance between the hit rate and false alarm ratio can be summarized in the difference between both (H-F). Fig. 4.1 shows how H-F and the lead time change with threshold.

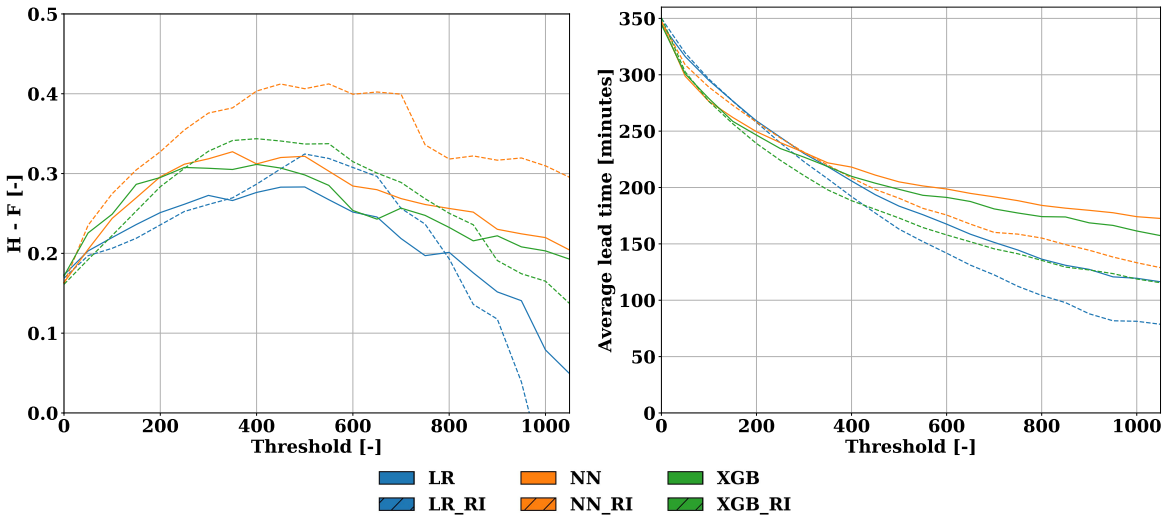


Figure 4.1: The left plot shows how the value of hit rate minus false alarm ratio changes with threshold. The right plot shows how the lead time changes with threshold. The plots show the results for all six configurations. The blue lines correspond to the linear regression model (LR), the orange lines to the neural network (NN), and the green lines to the XGBoost algorithm (XGB). The solid lines correspond to the model inputs based on the Spearman correlation and PDFs (knowledge inputs), and the dashed lines use the raw measurements (raw inputs).

For all six configurations H-F first increases, reaches a maximum and then decreases again. However, the rate of increase and decrease, and therefore its maximum, differs between configurations. The neural network with raw inputs shows the largest difference with the other models and performs best. The linear regression model with knowledge inputs shows the worst performance. The lead time

decreases with threshold and starts around 350 minutes at a threshold of zero, almost equal to the prediction window of six hours. The reason for this is that the evaluation is started six hours before the first lightning strike, and for most thunderstorms the prediction is above zero at this point. At low thresholds the lead time between configurations does not differ much but the gap becomes larger at higher thresholds. The neural network and XGBoost show the slowest decrease in lead time whereas the linear regression model shows the fastest decrease. Table 4.1 gives the optimum thresholds based on the H-F maxima with the corresponding H-F value and average lead time.

Table 4.1: This table shows the optimum threshold based on the H-F maximum for all six configurations. To compare the different models, the values of H-F and the lead time at this threshold are also given.

Model	Threshold	H-F (-)	Lead Time (m)
LR	500	0.28	183
LR-RI	500	0.32	163
NN	350	0.33	222
NN-RI	550	0.41	182
XGB	400	0.31	210
XGB-RI	400	0.34	188

In Appendix B the results are shown both for the training and testing days. This is done to check for overfitting. The figures show that the neural network and XGBoost models with knowledge input set have slight overfitting, meaning that the training set performs better than the testing set. The overfitting is larger compared to the raw inputs. This was expected since the raw inputs contain fewer model inputs and also less redundancy between inputs.

4.2. Validation

With the optimum thresholds selected the models are retrained and evaluated on the validation data. The true positives, false negatives, true negatives, false positives, and lead times of the six configurations are shown in Table 4.2.

Table 4.2: This table shows the classification results and the average lead times for all six configurations based on the validation data.

Model	True Positives	False Negatives	True Negatives	False Positives	Lead Time (m)
LR	217	56	697	273	185
LR-RI	239	34	698	272	177
NN	220	53	806	164	220
NN-RI	248	25	780	190	207
XGB	223	50	793	177	204
XGB-RI	236	37	761	209	183

The true positives are improved by 14% for the linear regression model, 10% for the neural network, and 6% for the XGBoost, when using the raw inputs over the knowledge inputs. Going from a linear technique to a non-linear technique improves the true positives by 4% for the neural network but reduces it with 1% for the XGBoost, compared to the best linear regression model. The model with the highest number of true positives is the neural network with raw inputs. The false positives are reduced by 0% for the linear regression model, 14% for the neural network, and 15% for the XGBoost, when using the knowledge inputs over the raw inputs. Going from a linear technique to a non-linear technique reduces the false positives by 40% for the neural network and 35% for the XGBoost. The model with the lowest number of false alarms is the neural network with knowledge inputs. The average lead time is increased by 5% for the linear regression model, 6% for the neural network, and 11% for the XGBoost, when using the knowledge inputs over the raw inputs. Going from a linear technique to a non-linear technique increases the lead time by 19% for the neural network and 10% for the XGBoost.

The model with the highest lead time is the neural network with knowledge inputs. The values in Table 4.2 are used to calculate the hit rate, false alarm ratio, and difference between them (H-F). These values are graphically represented in Fig. 4.2. The neural network with raw inputs has the highest hit rate (0.91), followed by the linear regression model with raw inputs (0.88), and the XGBoost model with raw inputs (0.86). The neural network with the knowledge inputs shows the best false alarm ratio (0.43). The linear regression model with both model inputs shows the worst false alarm ratio (0.56 and 0.53). Combining both in the H-F value, shows that the neural network with raw inputs provides the best classification results with a value of 0.47, with the second best value being 0.39 for the XGBoost model with raw inputs. The neural network with knowledge inputs (0.38) and the XGBoost with knowledge inputs (0.38) are not far behind. The worst performing model is the linear regression with knowledge inputs, having a H-F value of 0.23. When the lead time is evaluated, the neural network also gives the best performance. The neural network with knowledge inputs gives an average lead time of 220 minutes and with raw inputs of 207 minutes. The linear regression model with raw inputs shows the lowest lead time (177 minutes). Interestingly, the neural network with raw inputs has the highest threshold, namely 550, but shows a higher lead time than the linear regression and XGBoost models.

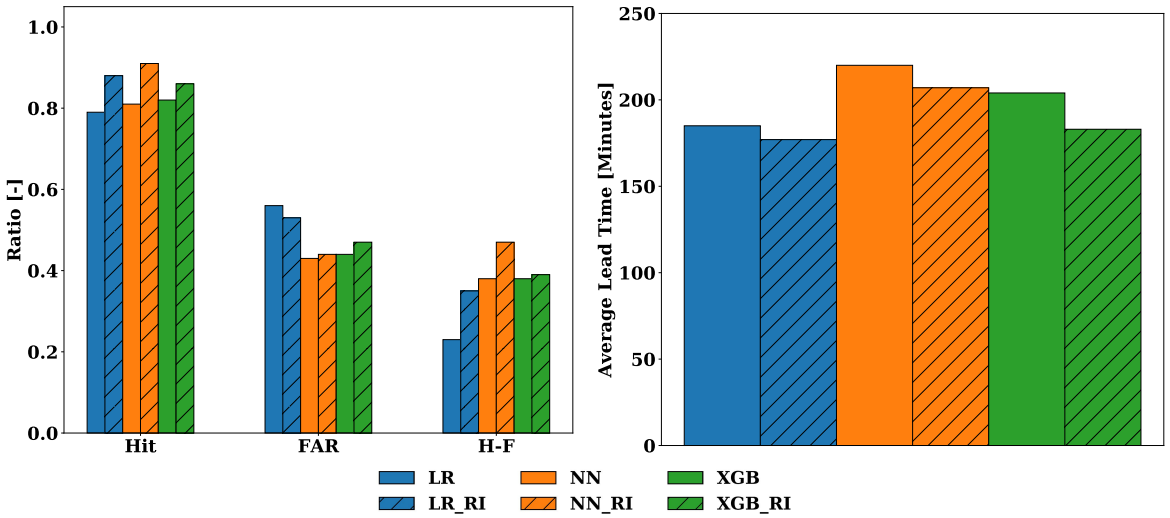


Figure 4.2: The left plot shows the hit rate, false alarm ratio, and difference between them (H-F). The right plot shows the how the average lead time differs between the six models. The results are based on the validation data.

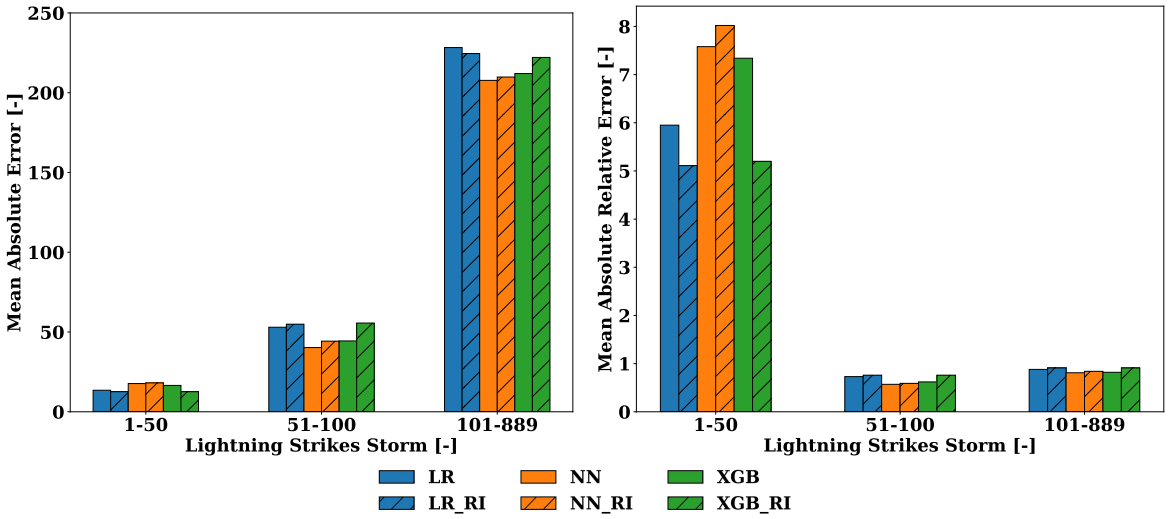


Figure 4.3: The plots in this figure relate to the intensity prediction of the thunderstorms for the validation set. The left plots shows the mean absolute error and the right plot the mean absolute relative error for three different categories of thunderstorms.

Fig. 4.3 give the regression results for all six configurations. Since the thunderstorms have a wide range of intensity, from 1 to 889 in the validation set, they have been divided into three categories. For the lowest category, 1 to 50 lightning strikes, the linear regression model with raw inputs gives the lowest mean absolute error and mean absolute relative error. For the middle category, 51-100 strikes, and the highest category, 101-889 strikes, the neural network with knowledge inputs performs slightly better than the other models. In general it can be seen that the mean absolute error is higher for larger thunderstorms but the mean absolute relative error is lower for larger thunderstorms. To clarify this, if a storm with 800 lightning strikes has been predicted to have 400 lightning strike, the absolute error is 400 but the absolute relative error is only 0.5. On the other hand if a storm with 1 lightning strike has been predicted to have 10 lightning strikes, the absolute error is only 9 but the absolute relative error 9. To understand if the regression performance is useful in practice the best performing model will be further analysed in Section 4.3.

The three tables in appendix B show the results both for the training and testing days of the validation procedure to check for underfitting or overfitting. However, a mixed picture occurs and no clear conclusions can be drawn. The reason for this is that no formal optimization procedure was done, as explained in Section 3.4, and the models use the same parameters for both inputs sets. Ideally, the parameters would be tuned separately for the raw inputs and for the knowledge inputs. Moreover, the model is evaluated on different criteria, and each criteria shows a different degree of overfitting. Nonetheless, the overfitting that occurs is larger than for the training procedure. The testing days of the k-fold cross-validation are evenly distributed over the two years of the training set. On the other hand, the validation data is a completely new and separate year.

To gain a deeper understanding of how thunderstorms can be predicted, the F score of the XGBoost model is investigated. The F score gives a measure of the importance of the individual model inputs by counting how often they have been used in making decisions. Fig. 4.4 shows the F score for the knowledge inputs as well as for the raw inputs.

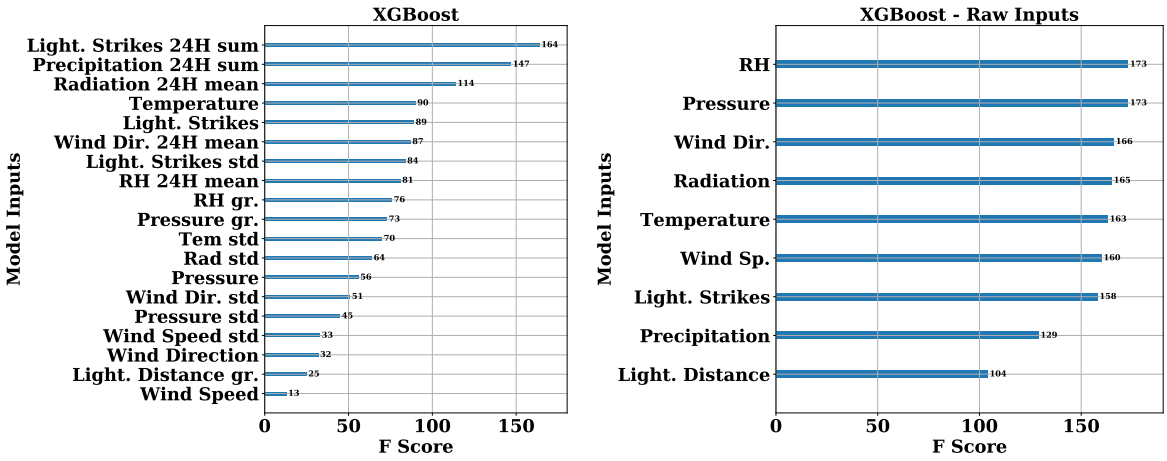


Figure 4.4: This figure shows the XGBoost F score for the knowledge inputs (left) and for the raw inputs (right). The F score counts the number of decisions made per model input. The higher the F score the more important an input can be considered for predicting thunderstorms.

For the knowledge inputs the lightning strike sum in the past 24 hours and the precipitation sum in the past 24 hours, show the highest importance. This indicates that there is value in knowing if a thunderstorm has occurred in the previous day to predict if there will be a new thunderstorm. The inputs based on the past 24 hour show more importance than the 5-minute measurements, gradients, and standard deviations. The F score for the raw inputs show less difference, with the exception of precipitation and lightning distance, which show a low importance. Precipitation and lightning strikes occur during or after the storm, thereby being less interesting in predicting if a storm will occur in the next hours. It is unclear however, why lightning strikes show a higher importance. Radiation and pressure are the most important inputs based on the F score, but the difference with wind direction, radiation, temperature, wind speed, and lightning strikes is small.

4.3. Selection & Analysis of Best Model

In Section 1.3 four criteria for selecting the best model were listed. The criteria included the hit rate, false alarm ratio, lead time, and the intensity prediction. The neural network with raw inputs performs best on the hit rate (0.91) and overall on the classification (H-F=0.47). The neural network with knowledge inputs has the lowest false alarm ratio (0.43). It also has the highest lead time (220 minutes) and the lowest mean absolute error and mean absolute relative error for the medium (51-100 lightning strikes) and large thunderstorms (101-889 lightning strikes). The neural network with knowledge inputs performs best on three of the four criteria and is therefore selected as the best model. The model predicts 220 out of 276 storms that occurred during the one year validation period. The 220 predictions include 43 out the 46 storms with more than 100 lightning strikes. The largest storm has 889 lightning strikes, occurring on the 14th of December 2019. The alarm for this storm is issued at 11:25, 2 hours and 15 minutes before the first lightning strike. 53 storms that occurred are not predicted by the model. Of those storms, only three have more than 100 lightning strikes (103, 117, and 372). 31 out of 53 missed storms have less than 10 lightning strikes and 12 out of the 53 missed storms have only 1 lightning strike. The model issues 164 false alarms. 51 of the false alarms occur on a day where there is a storm. The 164 alarms are distributed over 129 days, meaning that there are 35 days with more than one false alarm. Finally, 806 true negatives are counted. This is when there is no storm and also no alarm. Fig. 4.5 shows how the alarms are distributed over the months of the validation set. The model is able to capture the seasonality of thunderstorms, with peaks around March and November, without having any explicit knowledge of days or months. During 8 out of 12 months more true positives are issued than false positives. During 1 month this is equal and in the remaining 2 months there are more false alarms than true alarms. In the month of January 2020 no false alarms are issued yet 6 thunderstorms are predicted. Fig. 4.5 also shows the distribution of the lead time for the 220 predicted storms. The lead times range from 30 minutes to 350 minutes, with an average of 220 minutes.

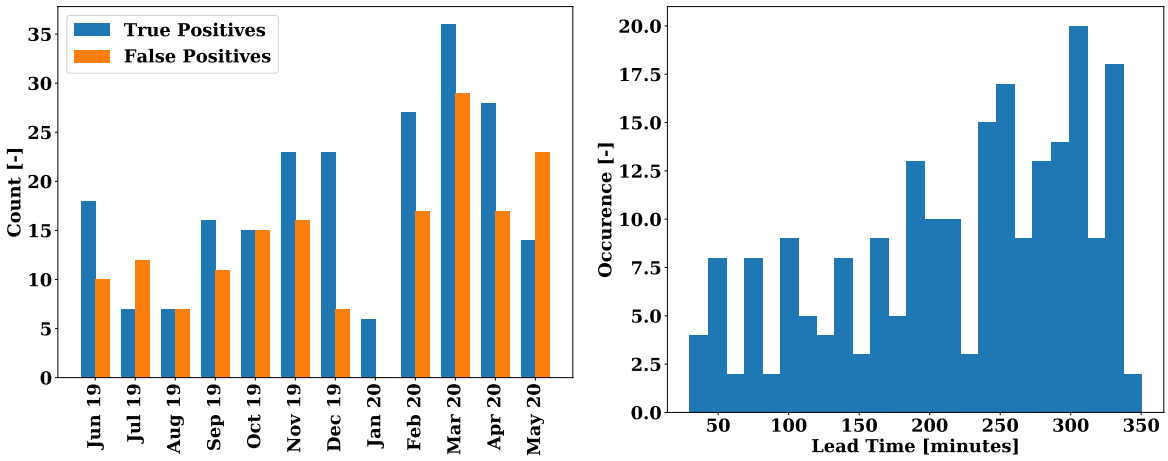


Figure 4.5: The left plot shows how the alarms issued by the neural network are distributed over the one-year validation set. The blue bars indicate the true alarms and the orange bars the false alarms. The right plot shows how the lead time of the predicted thunderstorms is distributed.

The average absolute error of the intensity prediction is 58 lightning strikes and the average absolute relative error is 5. All intensity predictions are between 5 and 92 lightning strikes whereas the storms intensity ranges from 1 to 889 lightning strikes. This means that the model is not able to predict the intensity of the 43 storms that have more than 100 lightning strikes. This becomes clear in Fig. 4.6 where there are absolute errors as high as 870 lightning strikes. The relative errors, shown in the figure, that are above 12 all occur for storms with only one or two lightning strikes. Since the predictions all are between 5 and 92 lightning strikes, the storms that fall within this range are best predicted. The Pearson correlation between the measured and predicted intensity is 0.14 for all storms and for storms between 5 and 92 lightning strikes it is 0.19.

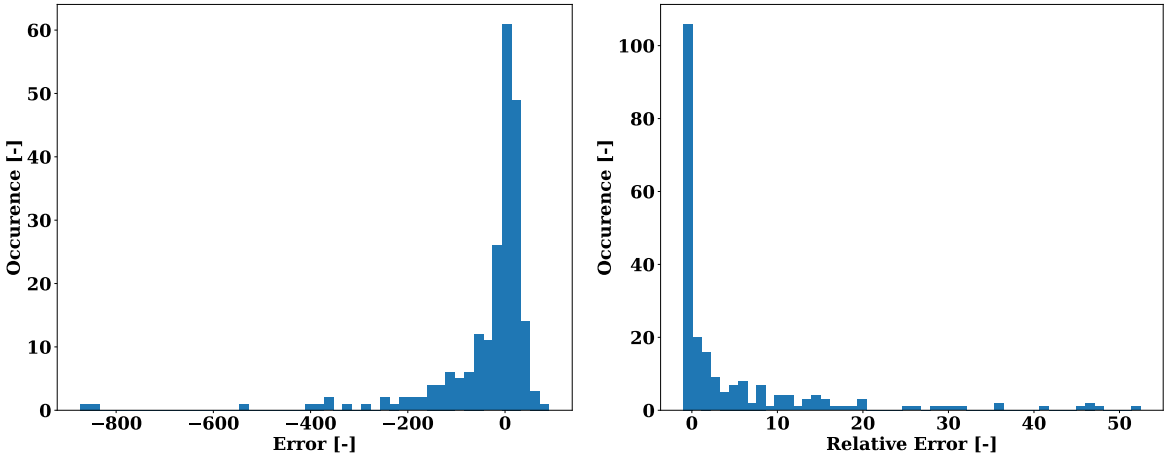


Figure 4.6: The figure shows the error (left plot) and relative error (right plot) made by the neural network with knowledge inputs. The error is the prediction intensity minus the storm intensity. The relative error is the error divided by the storm intensity. It can be seen that the errors are mainly negative and the relative errors mainly positive. The reason for this is that the highest errors occur for the large storms which are under predicted and the highest relative errors for small storms which are over predicted.

5

Discussion

After having presented the results of the early warning model in Chapter 4, this chapter will go on with discussing several points that are relevant to the model. To structure the discussion the points are gathered under eight categories. These are the comparison with the existing early warning models, the false alarm ratio, the thunderstorm severity, the model inputs, the model technique, the model set-up, further improvements, and finally what the model could mean for the drone-delivery company Zipline.

Comparison with other Models

Comparing the model from this study with the existing early warning models is not straight forward because there are fundamental differences with regards to the data, model set-up, and area. Nonetheless, there might be some value in comparing the hit rate and false alarm ratio. The Lake Victoria, XGBoost, Neural Network, and Weather Station model achieved hit rates of 0.85, 0.83, 0.89, and 0.77, and false alarm ratio's of 0.94, 0.03, 0.60, and 0.29, respectively. The value of 0.94 should be confirmed with the authors from the Lake Victoria study because it was not explicitly mentioned. The best configuration, as selected in Section 4.3, achieved a hit rate of 0.81 and a false alarm ratio of 0.43. Based on these percentages the model from this study scores fourth out of five on the hit rate and third out of five on the false alarm ratio. An important distinction here is that the model from this study used a dataset of three years and the other models at least nine years. A longer dataset is likely to improve this model and can be tested by doing a sensitivity analysis using different data lengths. Moreover, the neural network with raw inputs achieved a hit rate of 0.91, ranking it first out of five.

False Alarm Ratio

The high false alarm ratio is a key obstacle to overcome for a real life application. The best configuration issues 220 true alarms and 164 false alarms. During 8 out of 12 months there are more true alarms than false alarms, but in July 2019 and May 2020, the model issues more false alarms than true alarms. This is problematic and should be a priority improvement for the model. Understanding when these false alarm occur should be investigated further. As mentioned in Section 3.1 the spatial scales of the weather variables is different. The lightning sensor only measures lighting within 40 km but other variables such as pressure and temperature have larger spatial scales. Looking at other weather stations to understand if the false alarms are issued for storms further away could increase our understanding. The results also showed that 51 out of 164 false alarms occur on days when a storm occurs. It should be investigated how long before or after the false alarm the storm occurs. Moreover, the XGBoost model achieved a false alarm ratio of 0.03 and used a lead time of 10 minutes. Conducting a sensitivity analysis on this model for different prediction windows (currently 6 hours), could show if if the false alarm ratio can be improved by using shorter lead times.

Thunderstorm Severity

In Section 4.3 it was seen that the model predicts thunderstorms between 5 and 92 lightning strikes, whereas the actual thunderstorms range from 1 to 889 lightning strikes. It is therefore clear that the model is not able yet to distinguish between large and small storms, which results in large absolute

errors for big thunderstorms and large relative errors for small thunderstorms. A more accurate intensity prediction will provide large benefits for this model. Understanding how severe an upcoming thunderstorm is will guide the necessary precautions and minimize their costs. Moreover, with a good intensity prediction the threshold can be altered so that only certain storms are considered, and no alarm is raised for small storms. This can be done based on the preferences of the user. At the moment however, the overprediction of small storms means that they are still included in the predictions, even when using higher thresholds. It is expected that a longer data record can improve the intensity prediction. The Lake Victoria area has a high lightning activity, the two year training set contains 422 thunderstorms, but only 11% of the storms in the training record have more than 100 lightning strikes. A longer data record would contain more severe storms and these could serve as an example for predicting future storms. Although the intensity prediction lacks skill, the current model is already able to classify 43 out of the 46 thunderstorms that have more than 100 lightning strikes.

Model Inputs
In Section 4.2 it was seen that the raw inputs improve the true positives with 14%, 10%, and 6% but that the knowledge inputs improve the false positives with 0%, 14%, and 15% for the linear regression model, neural network, and XGBoost, respectively. By using the raw inputs for the thunderstorm prediction and the knowledge inputs to correct for possible false alarm, it should be possible to improve the model. In principle the model could then achieve a hit rate of 0.91, as the neural network with raw inputs, and a false alarm ratio of 0.43, as the neural network with knowledge inputs. A possible way to combine the models would be to use the output of the neural network with raw inputs as an input for the neural network with knowledge inputs. This would add a second layer that can possibly identify alarms as being false.

Model Technique
Section 4.2 also showed that going from a linear technique to a non-linear technique reduces the false alarms by 40% for the neural network and 35% for the XGBoost model. Using a non-linear technique improves the true positives with 4% for the neural network and reduces them with 1% for the XGBoost, compared to the best linear regression model. The lead time is increased by 19% for the neural network and 10% for the XGBoost. It is likely that the results can be further improved by working on the model technique. For example, the parameters for the models using the raw inputs are the same as for the models using the knowledge inputs. Improvements can be made by tuning the parameters separately for the different model inputs. Moreover, the parameters have not been formally optimized but rather tuned by trial and error. A formal optimization would entail systematically trying out different combinations of parameters, something that quickly leads to excessive computational times and was therefore not done in this study. A more thorough check if the models are overfitting or underfitting should also be conducted. The challenge here is that what is predicted by the model, namely the number of lightning strikes in the next six hours, is not what the model is eventually evaluated on, namely hit rate, false alarm ratio, lead time, and intensity prediction. This makes it hard to see if overfitting or underfitting is occurring. Appendix B shows that almost no overfitting occurs for the k-cross training procedure but a mixed picture occurs for the validation. It is also possible that a better machine learning technique exists for this problem. A long short-term memory (LSTM) neural network uses feedback between current and past temporal inputs that allows it to develop a ”memory”. In this way no explicit temporal knowledge has to be passed in the model, as was done with the knowledge inputs. In this way more subtle temporal changes that lead to thunderstorms might be identified as is possible with the gradients, standard deviations, and 24 hour averages.

Model Set-up
Section 3.5 described the conventional approach for a classification prediction and that it comes with the disadvantage of a reduced temporal resolution and uncertainty in when the thunderstorm actually occurs. A new approach is developed that maintains the resolution and aggregates the values after the prediction and not before. The lead time becomes flexible but more precise and it also adds an prediction on the thunderstorm severity. However, the results showed that the predictions do not capture the thunderstorm severity well. If this would have been the case, the lead time would become equal to the prediction window, six hours in this case, and the threshold only serves to disregard storms below a certain intensity. As it is now the threshold works more as a filter that removes noise and that balances

the hits and false alarms. Improving the regression would make the developed model set-up more useful. Moreover, the algorithm that counts the true positives, false negatives, true negatives, and false positives, is new and should be reviewed by another party. Especially counting the true negatives and false positives is challenging because there is no reference points such as a lightning strike. The way the algorithm works has a strong effect on the final evaluation of the model. For this reason it is essential that the algorithm is eventually tested in real life. Only then is it possible to see how useful the model predictions are and how well the algorithm evaluates them.

Further Improvements
Besides the possible improvement already mentioned such as combining the model inputs, optimizing the parameters, using multiple weather stations, and conducting sensitivity analysis on the lead time and data length record, some further improvements can also be considered. For example, it could be interesting to combine the TAHMO weather station data with other data sources. Using satellite measurements, such as overshooting tops, could provide added skill to the model. The Lake Victoria model used daytime overshooting tops over the surrounding land to predict nighttime overshooting tops over the lake. For this model the overshooting tops counted during the previous night could be added. Moreover, the predictions of a numerical weather model could also be used. The weather model could give an indication of thunderstorm occurrence in the next days and the early warning model could be used as a precision tool to indicate when it will happen.

Zipline
Although the developed model predicts thunderstorm for a weather station on the eastern side of Lake Victoria, it could also potentially aid medicine delivery drones in avoiding bad weather. The drones operate in Rwanda and Ghana, so to understand if this model has possible benefits it should be tested for a different weather station. Section 3.3 describes how the model inputs are selected. An advantage of the method is that it selects the inputs based on the measurements. This means that when a different weather station is used the inputs that arise from the Spearman correlation and PDFs will be different from the inputs for the TA00173 weather station. In principle the model should therefore also be able to predict thunderstorms for a different area. However, it should be noted that the focus area of this study is very lightning active, the 2 year training set contained 422 thunderstorms. This large amount of thunderstorms compensates for the short data length record, which would not be the case when the weather station is in an area with less thunderstorms. Understanding how the area influences the results is therefore an important next step in improving the model. The developed model provides early warnings for lightning strikes but not for the associated winds and precipitation events, which are the source of danger for the drone flights. It would therefore be useful to study the relationship between lightning strikes, precipitation, and wind, and see how this could be incorporated in the model. This study tried to do this by using the Calculated Convective Activity (CALCA), a variable consisting of lightning strikes, precipitation, and wind gusts, as developed by Manzato (2007). However, it was found that at this point the variable did not add any benefit. The prediction skill of the number of lightning strikes should be improved first, before moving on to the more complex task of also predicting wind and precipitation.

6

Conclusion & Recommendation

During this research a new early warning model was developed that predicts thunderstorm occurrence around Lake Victoria. In Section 1.3, three research questions were posed to guide this process:

- How do the current thunderstorm early warning models work and perform?
- Which model inputs and prediction techniques lead to the best performance of the early warning model based on the model criteria?
- What are the prediction characteristics of the best performing model?

Beginning with the first question, the four current early warning models that were investigated all showed different characteristics. The models varied in their data sources, areas, data record lengths, model targets, lead times, and prediction techniques. Even with the large variety in models consistent high hit rates were achieved (0.89, 0.85, 0.83, and 0.77). It is therefore concluded that regardless of the type of data source, area, and the prediction technique, it is possible to predict the majority of thunderstorms using statistical or machine learning methods. Most noticeably, using only ground measurements from weather stations seems sufficient to predict a thunderstorm which is by its nature a vertical process. A larger difference between models was found in the false alarm ratio (0.03, 0.29, 0.60, and 0.94). The Switzerland study distinguishes itself by achieving a ratio of 0.03, possibly due to low lead time of 10 minutes compared to the 3 or 6 hours of the others models. Nonetheless, the other three studies make apparent that the main challenge of an early warning model is to reduce the number of false alarms.

To answer the second research question two different sets of model inputs were tested, namely the raw inputs and the knowledge inputs. Moreover, two non-linear techniques, namely a neural network and the XGBoost model, and one linear technique were tested. Based on comparing the six configurations, it is concluded that for the true positives the raw inputs provide better results compared to the knowledge inputs. For the true positives there is no clear preference for non-linear over linear techniques. For the false positives and lead time, the non-linear techniques with knowledge inputs provide better results compared to the linear techniques with raw inputs. This means that for predicting thunderstorms a linear technique and current weather station measurements are sufficient, but using non-linear techniques and past temporal weather station measurements reduces the false alarms and improves the lead time. This effect is also seen for the intensity prediction, but no conclusions are drawn at this point due to the small differences and overall lacking skill in this area.

Answering the third research question, the neural network with knowledge inputs scores best on three out of four criteria, namely the false alarm ratio, lead time, and intensity prediction. The model predicts 220 out of 276 storms, resulting in a hit rate of 0.81, with an average lead time of 220 minutes. The model predicts thunderstorms between 5 and 92 lightning strikes, whereas the actual storms range from 1 to 889 lightning strikes. Although the intensity prediction lacks skill, the model is able to classify 43 out of 46 storms with more than 100 lightning strikes. During the one year validation period the model still issues 164 false alarms, resulting in a false alarm ratio of 0.43. 51 of these false alarms occur on days with thunderstorms. It is therefore concluded that the classification results show promise but that

the false alarms are too high for any practical application. Moreover, at this stage the model is not able to predict if the upcoming thunderstorm will be small or large.

Compared to the four existing early warning models, the neural network with knowledge inputs scores fourth out of five on the hit rate and third out of five on the false alarm ratio. The model does this with a data length record of three years compared to at least nine year of the other models. The main improvement over the other Lake Victoria model is the reduction of false alarms due to the use of a non-linear technique and more extensive model inputs. However, it is important to note that by focusing on the 1% most intense storms, the Lake Victoria model increases its false alarms. The model set-up developed in this study offers a more precise lead time compared to the Lake Victoria model and it maintains the temporal resolution by aggregating after the prediction.

Recommendation

As a way forward in developing the early warning model several recommendations are made. Combining the neural network with raw inputs and neural network with knowledge inputs should be investigated since it could potentially lead to a higher hit rate while maintaining the lower false alarm ratio. For the model set-up, as described in Section 3.5, an algorithm is developed that compares the predictions to the lightning strike measurements. Since it is newly developed the algorithm should be checked and corrected by a different reviewer. Moreover, the model parameters should be formally optimized and for each technique and model input set separately. Investigating different model techniques, such as a LSTM neural network, could also lead to improved model skill. A sensitivity analysis should be conducted on the model whereby the prediction window, dataset length, and model inputs, are varied and its effect on the performance evaluated. This might provide insight into how the false alarm ratio and intensity prediction can be improved. Combining multiple weather stations could also lead to a better understanding of the model. Further improvements could be realised by adding different data sources, for example, including overshooting tops or numerical weather model predictions. To understand if this model also has potential for aiding Zipline's medicine delivery drones in Rwanda and Ghana, it should be tested on a different weather station. Finally, the only way to understand the potential of the model is to test it on location and see how the predictions match the real life conditions. Before doing this however, more research and development is recommended.

Bibliography

Albrecht RI, Goodman SJ, Buechler DE, Blakeslee RJ, Christian HJ (2016) Where are the lightning hotspots on earth? Bulletin of the American Meteorological Society DOI 10.1175/BAMS-D-14-00193.1

AustriaMicrosystems (2012) Franklin Lightning Sensor ICI datasheet p 27

Barnes LR, Schultz DM, Gruntfest EC, Hayden MH, Benight CC (2009) Corrigendum: False alarm rate or false alarm ratio? DOI 10.1175/2009WAF2222300.1

Dezfuli AK, Ichoku CM, Mohr KI, Huffman GJ (2017) Precipitation characteristics in West and East Africa from satellite and in situ observations. Journal of Hydrometeorology DOI 10.1175/JHM-D-17-0068.1

van de Giesen N, Hut R, Selker J (2014) The Trans-African Hydro-Meteorological Observatory (TAHMO). Wiley Interdisciplinary Reviews: Water DOI 10.1002/wat2.1034

Gong DY, Guo D, Mao R, Yang J, Gao Y, Kim SJ (2016) Interannual modulation of East African early short rains by the winter Arctic Oscillation. Journal of Geophysical Research DOI 10.1002/2016JD025277

Haiden T, Rodwell MJ, Richardson DS, Okagaki A, Robinson T, Hewson T (2012) Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. Monthly Weather Review DOI 10.1175/MWR-D-11-00301.1

Kamau G, Kang'ethe S, Kamau S, van de Giesen N (2015) Design of a low-cost microcontroller-based lightning monitoring device. Kabarak Journal of Research & Innovation 3 (1)

Manzato A (2007) Sounding-derived indices for neural network based short-term thunderstorm and rainfall forecasts. Atmospheric Research DOI 10.1016/j.atmosres.2005.10.021

Mary AK, Gomes C (2012) Lightning accidents in Uganda. In: 2012 31st International Conference on Lightning Protection, ICLP 2012, DOI 10.1109/ICLP.2012.6344235

Mary AK, Gomes C (2015) Lightning safety of under-privileged communities around Lake Victoria. Geomatics, Natural Hazards and Risk DOI 10.1080/19475705.2014.922506

Mostajabi A, Finney DL, Rubinstein M, Rachidi F (2019) Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. npj Climate and Atmospheric Science DOI 10.1038/s41612-019-0098-0

Pucillo A, Manzato A (2013) Usefulness and skill of station-derived predictors in forecasting storm occurrence and intensity. Atmospheric Research DOI 10.1016/j.atmosres.2012.10.016

Schlemmer L, Hohenegger C (2014) The formation of wider and deeper clouds as a result of cold-pool dynamics. Journal of the Atmospheric Sciences DOI 10.1175/JAS-D-13-0170.1

Singh C, Daron J, Bazaz A, Ziervogel G, Spear D, Krishnaswamy J, Zaroug M, Kituyi E (2018) The utility of weather and climate information for adaptation decision-making: current uses and future prospects in Africa and India. DOI 10.1080/17565529.2017.1318744

Thiery W, Davin EL, Seneviratne SI, Bedka K, Lhermitte S, Van Lipzig NP (2016) Hazardous thunderstorm intensification over Lake Victoria. Nature Communications DOI 10.1038/ncomms12786

Thiery W, Gudmundsson L, Bedka K, Semazzi FH, Lhermitte S, Willems P, Van Lipzig NP, Seneviratne SI (2017) Early warnings of hazardous thunderstorms over Lake Victoria. Environmental Research Letters DOI 10.1088/1748-9326/aa7521

TWIGA (2017) Proposal in response to H2020 call: SC5-18-2017 'Novel in-situ observation systems'

Vaughan C, Hansen J, Roudier P, Watkiss P, Carr E (2019) Evaluating agricultural weather and climate services in Africa: Evidence, methods, and a learning agenda. DOI 10.1002/wcc.586

WMO (2017) Guidelines for Nowcasting Techniques (WMO-No. 1198)

Yang W, Seager R, Cane MA, Lyon B (2015) The annual cycle of East African precipitation. Journal of Climate DOI 10.1175/JCLI-D-14-00484.1

A

Model Input Figures

The model inputs are selected based on the Spearman correlation and probability density functions. The figures that are not shown in the main report are found in this appendix. Fig. A.1, Fig. A.2, and Fig. A.3 show the correlation matrices for the standard deviations, gradients, and daily values, respectively. Fig. A.4, Fig. A.5, Fig. A.6, and Fig. A.7 show the probability density functions for the 5-minute intervals, standard deviations, gradients, and daily values, respectively.

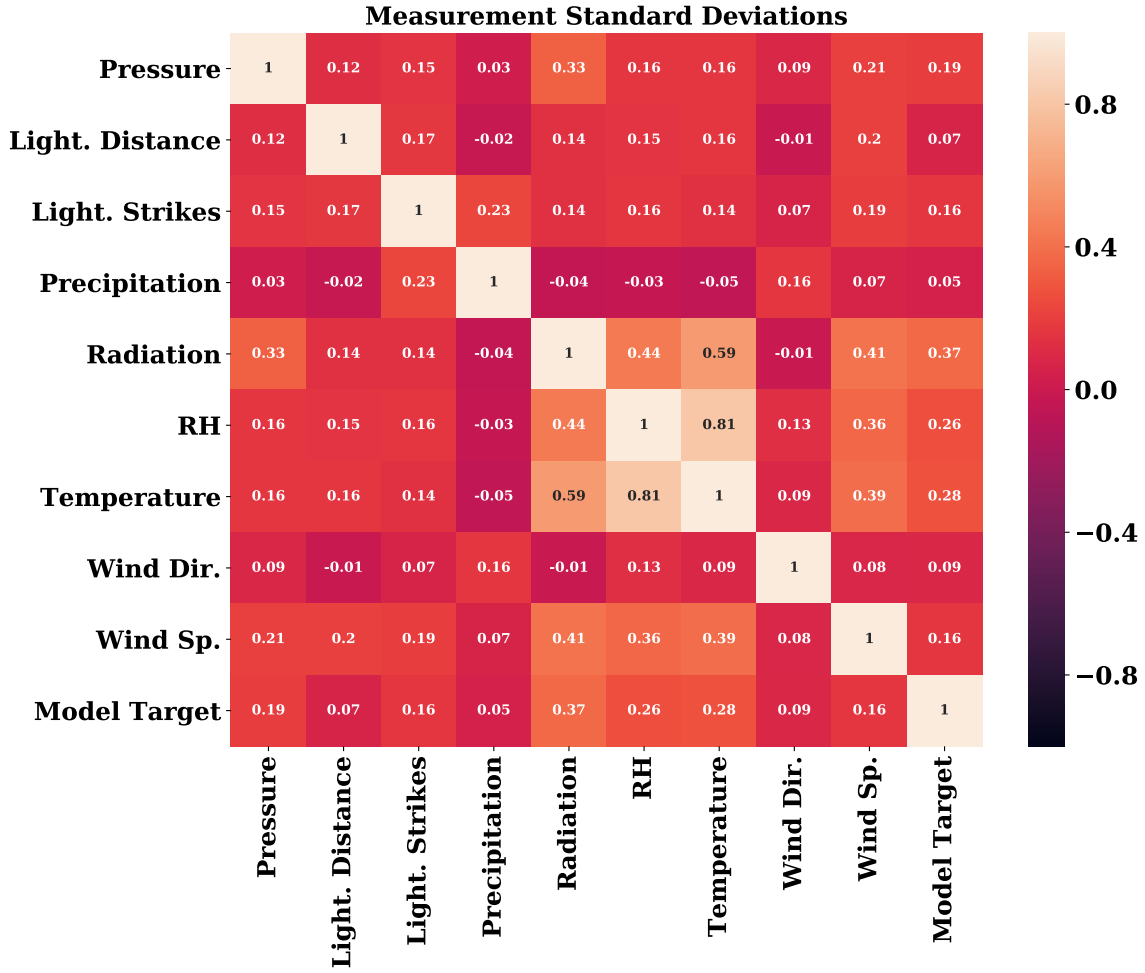


Figure A.1: Spearman correlation of the measurement standard deviations taken over a three hour period, with each other and with the model target.

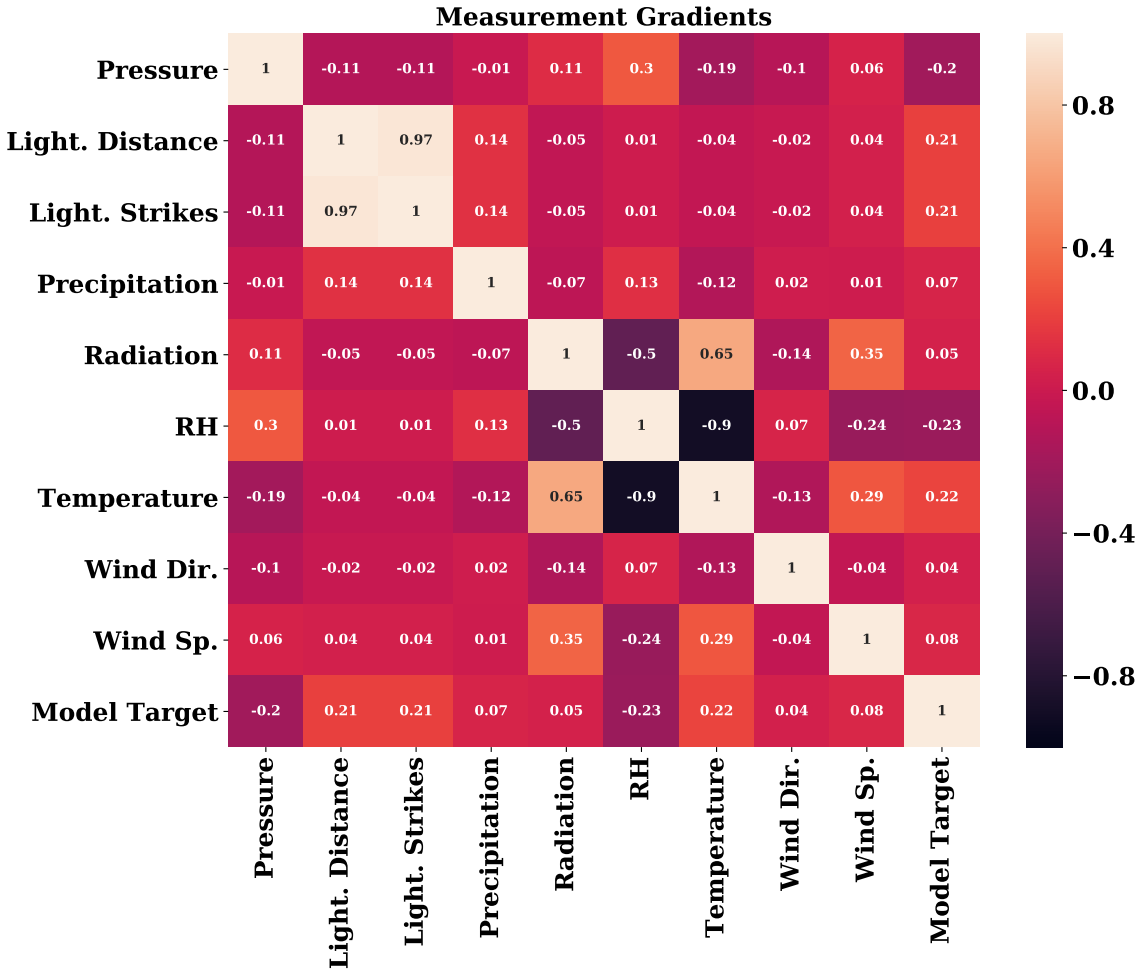


Figure A.2: Spearman correlation of measurement gradient taken over a three hour period, with each other and with the model target

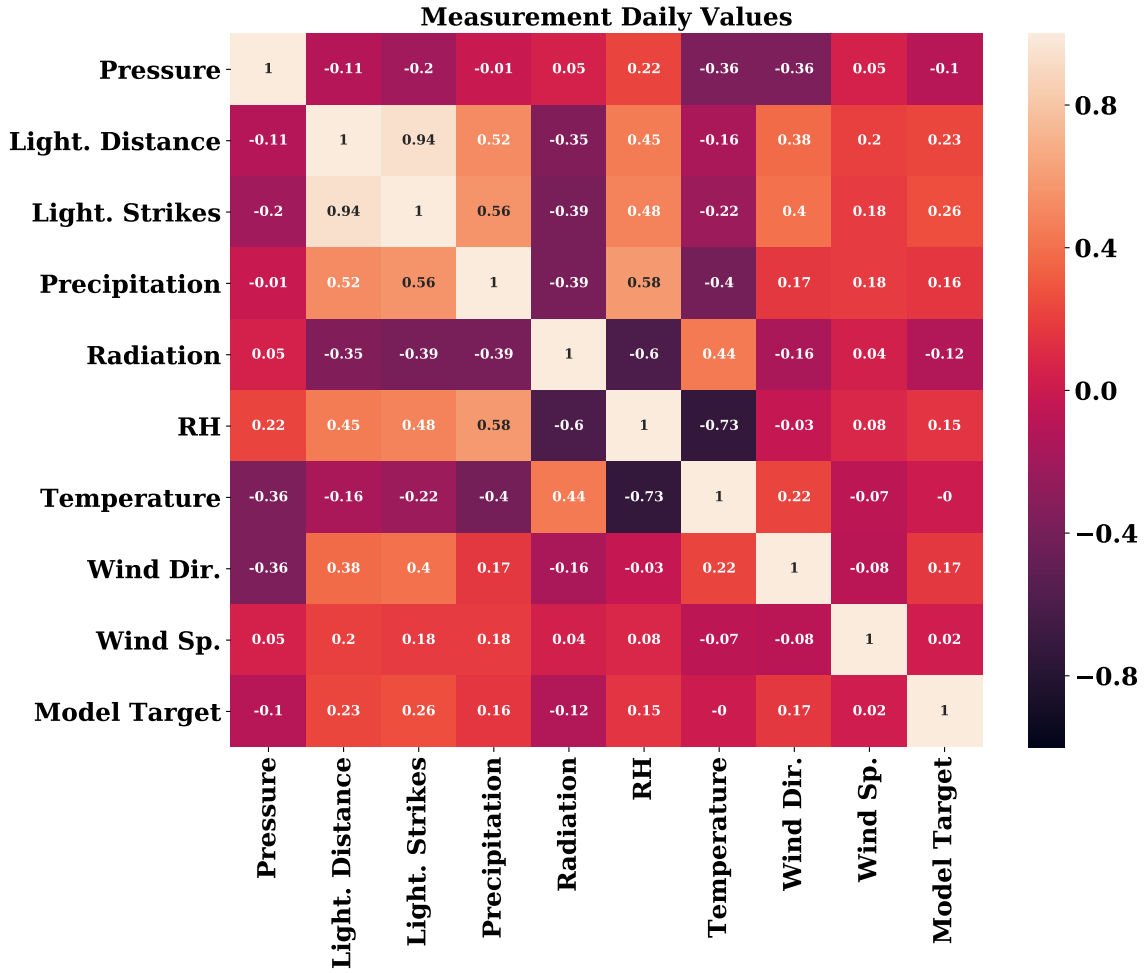


Figure A.3: Spearman correlation of the measurement over the past 24 hours, with each other and with the model target.For the lightning events and the precipitation the sum over the past 24 hours has been used. For the wind speed the maximum value over the past 24 hours is used. For the other variables the mean value is used.

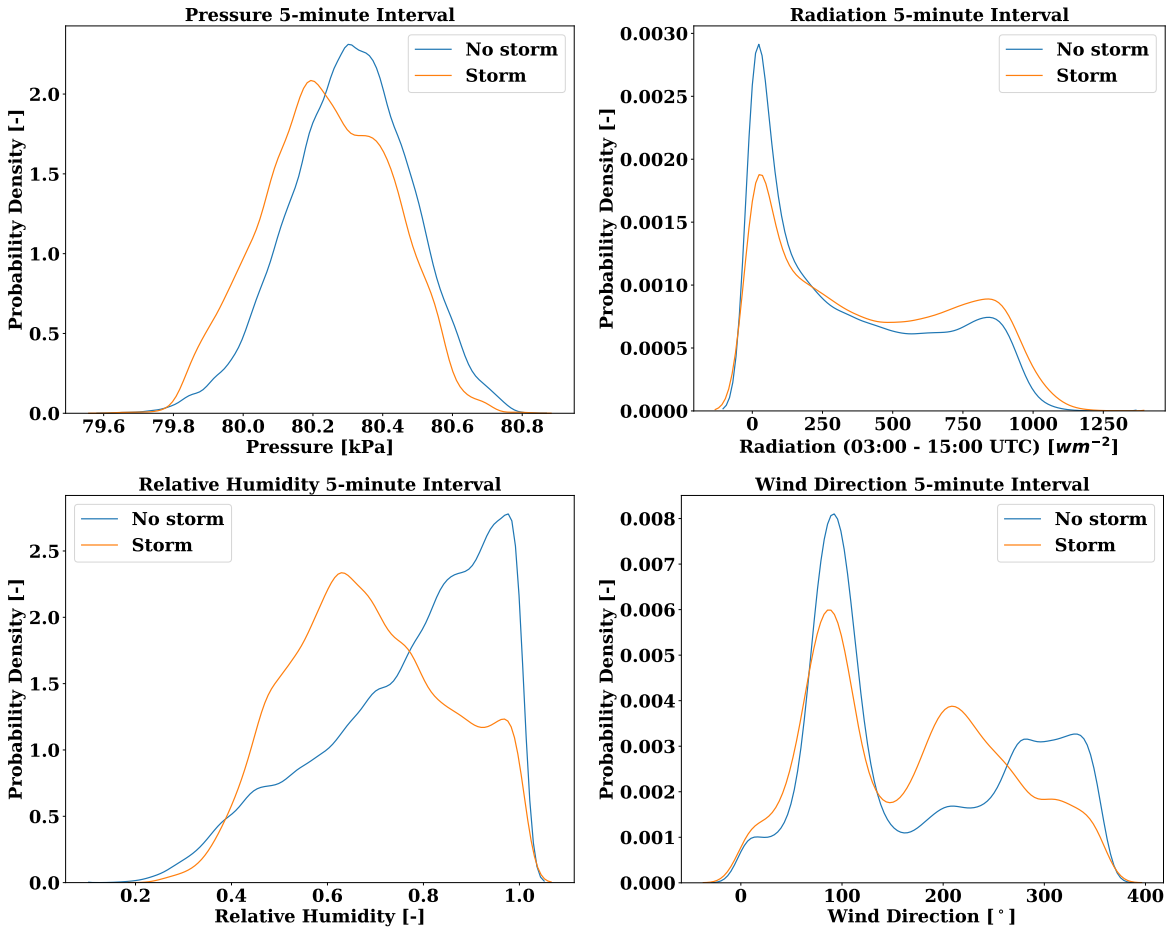


Figure A.4: Probability density functions for thunderstorm versus no thunderstorm for the 5-minute intervals.

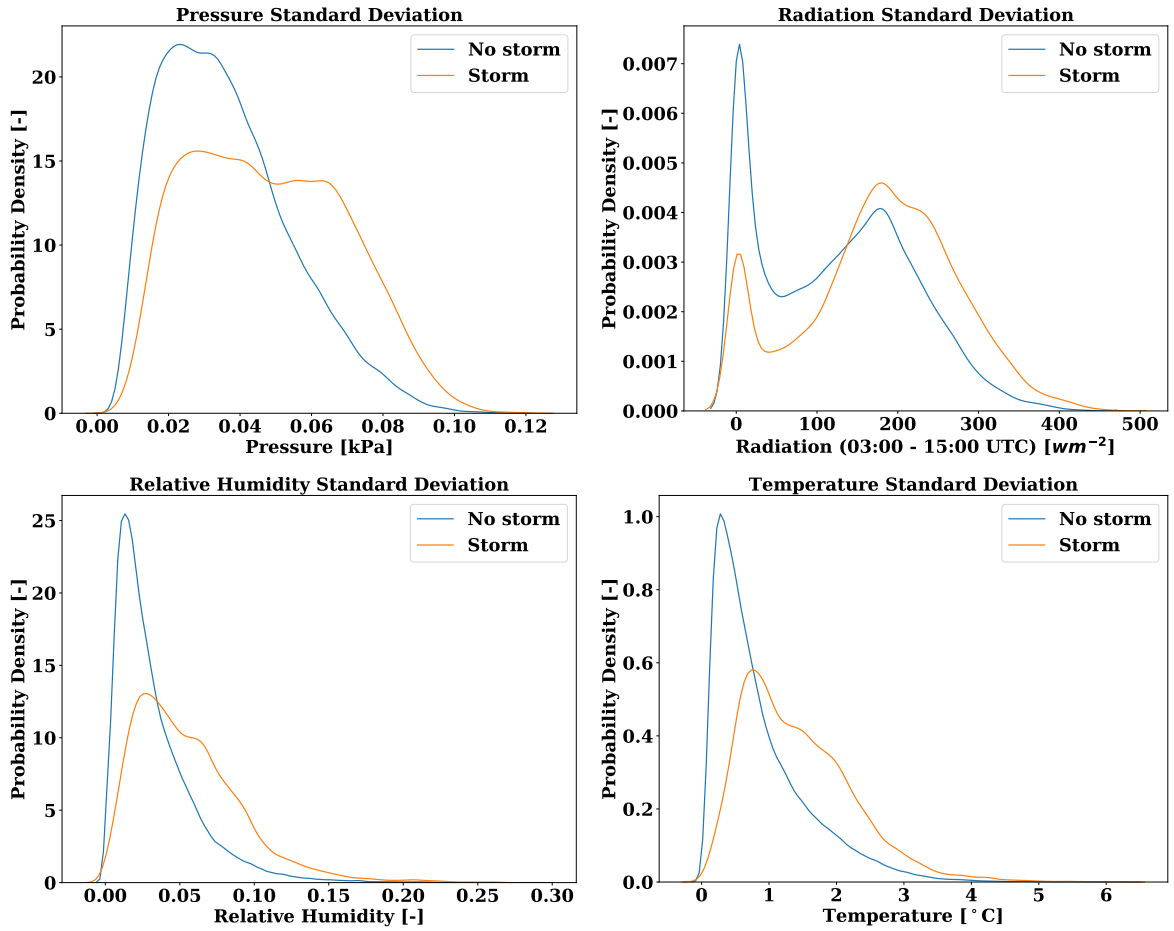


Figure A.5: Probability density functions for thunderstorm versus no thunderstorm for the measurement standard deviations.

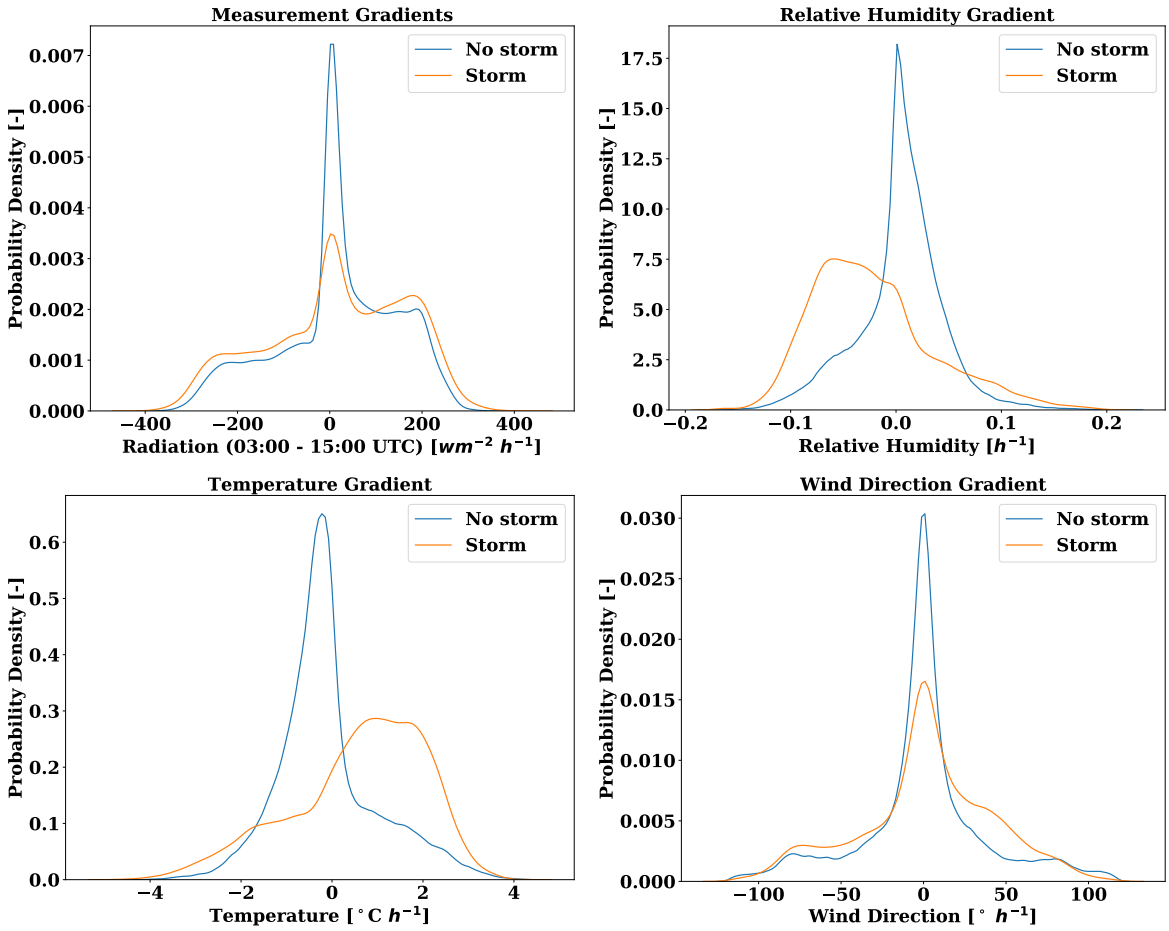


Figure A.6: Probability density functions for thunderstorm versus no thunderstorm for the measurement standard gradients.

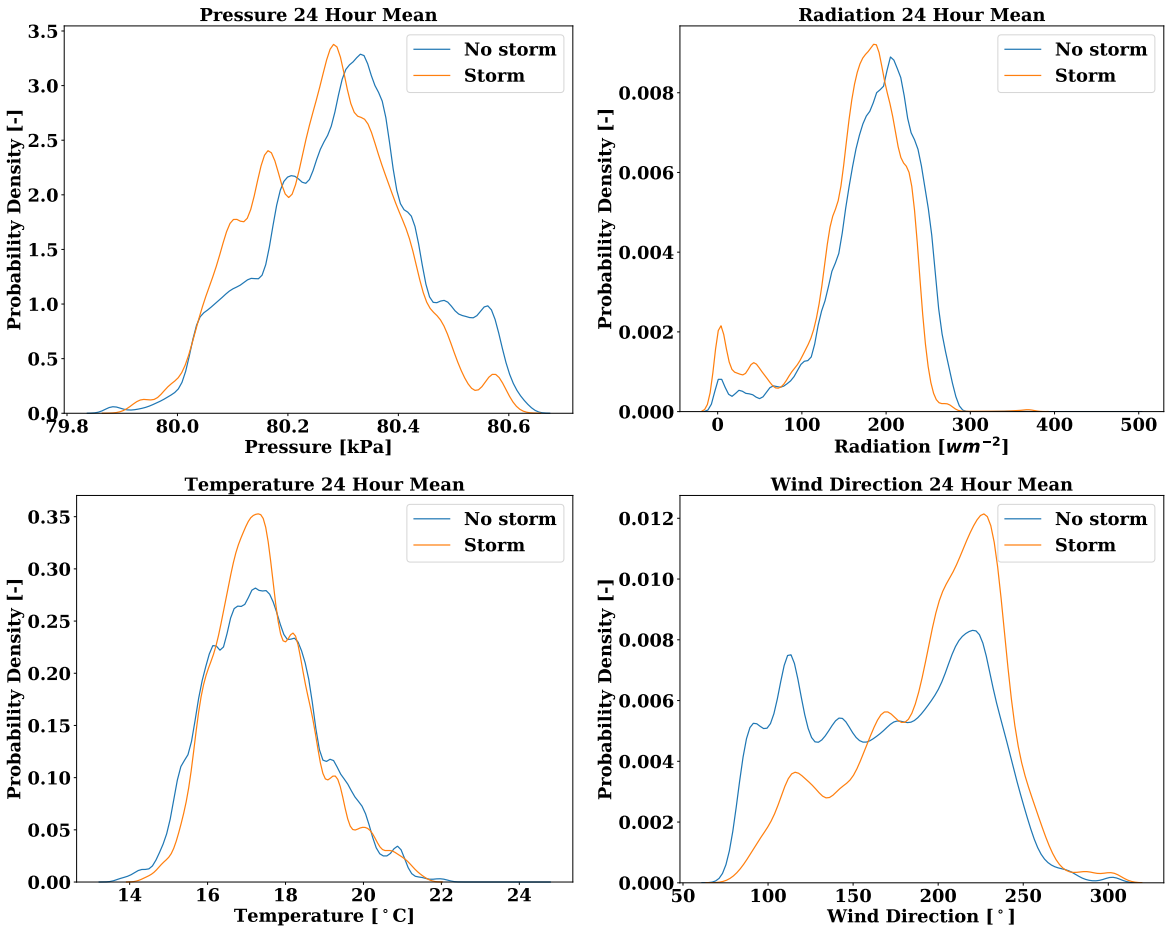


Figure A.7: Probability density functions for thunderstorm versus no thunderstorm for the measurement daily values.

B

Overfitting

This appendix shows if overfitting occurred both for the training procedure of section 4.1 and the validation procedure of section 4.2. For the training procedure, the change of the hit rate, false alarm ratio, and lead time with threshold is shown for the linear regression models in Fig. B.1, for the neural network models in Fig. B.2, and for the XGBoost models in Fig. B.3. The models that show the most overfitting are the neural network and XGBoost with knowledge inputs. Both models show more overfitting compared to the raw inputs. The overfitting shows itself by higher hit rates and lower false alarm ratio's for the training set compared to the testing set. The lead times show only small differences between the training and testing sets. Table B.1, table B.2, and table B.3, show the results for the validation procedure. A very mixed picture occurs, where a better performance is reached by the validation data on some criteria but worse on other criteria. It is also not clear if the raw inputs or knowledge inputs give rise to more overfitting. For the linear regression model and neural network the knowledge inputs show more overfitting, whereas for the XGBoost model the raw inputs. Three reasons are identified for this mixed picture. Firstly, there was no formal optimization for the parameters of the models. Secondly, both model input sets use the same model parameters. In the best case, the parameters would be optimized according to the model inputs. Finally, the model is judged on multiple criteria and they can all show different degrees of overfitting.

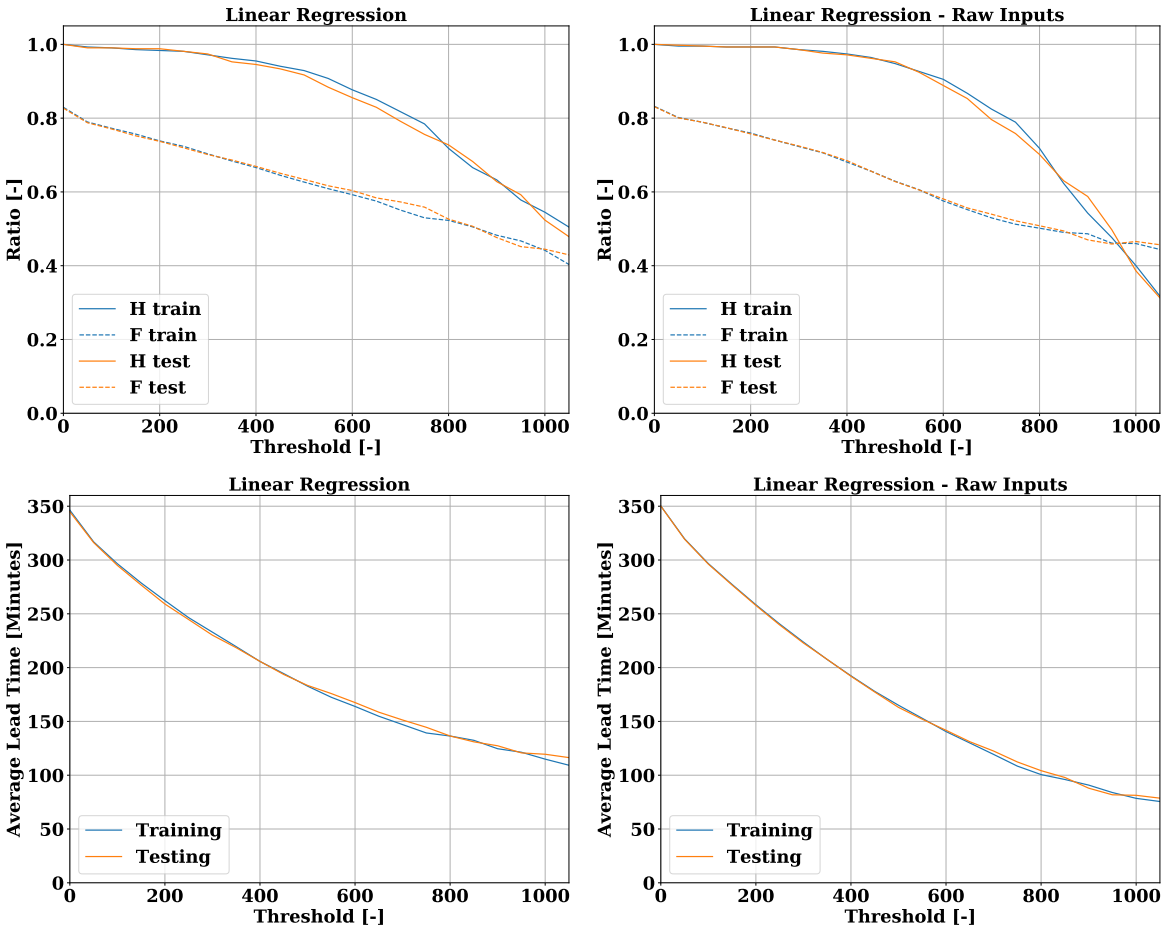


Figure B.1: The upper plots show how the hit rate (solid line) and false alarm ratio (dashed line) change with threshold. Both the training set (blue line) and testing set (orange line) are considered. The lower plots show how the lead time changes with threshold, for the training set (blue line) and for the testing set (orange line). The left plots show this for the linear regression model with correlation and PDF inputs and the right plots for the linear regression model with raw inputs

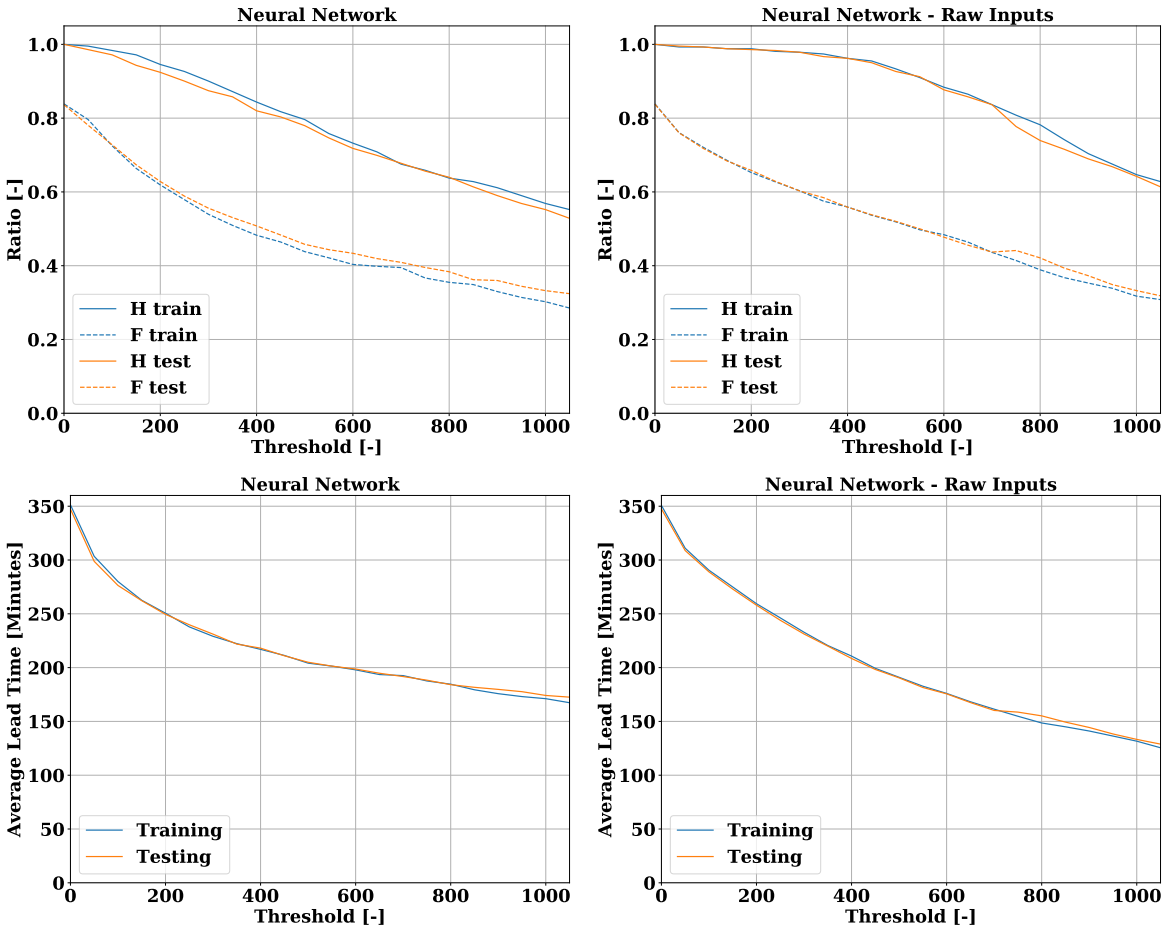


Figure B.2: This figure shows the same as Fig. B.1 but now for the neural network models.

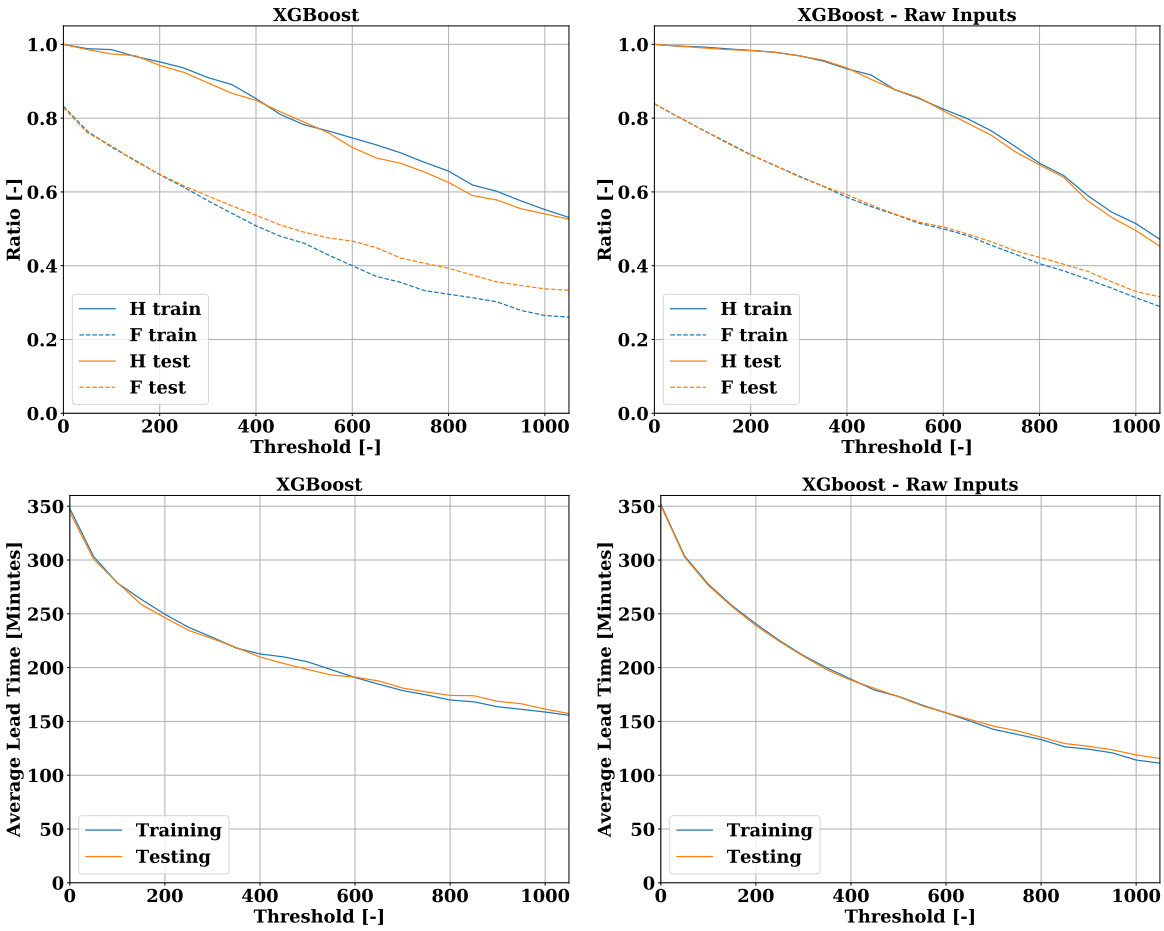


Figure B.3: This figure shows the same as Fig. B.1 but now for the XGBoost models.

Table B.1: Results for the training and validation sets, both for the linear regression model with correlation and PDF inputs and the linear regression model with raw inputs.

	LR		LR-RI	
	Training	Validation	Training	Validation
Hit rate (-)	0.93	0.79	0.95	0.88
FAR (-)	0.63	0.56	0.63	0.53
Lead Time (minutes)	182	185	164	177
MAE	41.03	57.83	40.70	54.81
MARE	4.76	4.39	4.16	3.86

Table B.2: Results for the training and validation sets, both for the neural network model with correlation and PDF inputs and the neural network model with raw inputs.

	NN		NN-RI	
	Training	Validation	Training	Validation
Hit rate (-)	0.86	0.81	0.91	0.91
FAR (-)	0.51	0.43	0.48	0.43
Lead Time (minutes)	225	220	184	207
MAE	44.33	57.60	41.52	55.28
MARE	6.37	5.40	5.62	5.85

Table B.3: Results for the training and validation sets, both for the XGBoost model with correlation and PDF inputs and the XGBoost model with raw inputs.

	XGB		**XGB-RI**	
	Training	**Validation**	**Training**	**Validation**
Hit rate (-)	0.83	0.82	0.94	0.86
FAR (-)	0.51	0.44	0.58	0.47
Lead Time (minutes)	217	204	188	183
MAE	41.94	56.85	40.92	56.95
MARE	5.52	5.26	4.53	3.86