Probabilistic thunderstorm forecasts using statistical post-processing: Comparison of logistic regression and quantile regression forests and an investigation of physical predictors

Edward Groot

Supervisors: Dr. Maurice Schmeits (KNMI), Kirien Whan Ph.D. (KNMI) & Dr. Willem Jan van de Berg (IMAU)

July 11, 2019

Abstract

Probabilities of thunderstorm occurrence and conditional probabilities of lightning intensity over The Netherlands are forecast using statistical post-processing with predictors derived from the operational non-hydrostatic numerical weather prediction model Harmonie, at lead times up to 45 hours. Quantile regression forests (QRF) is compared with logistic regression (LR) for thunderstorm occurrence forecasts and with extended LR for lightning intensity forecasts. Using different sets of predictors that these statistical methods may select, it is demonstrated that pre-selection of predictors based on physical understanding and simultaneously exploiting QRF as machine learning tool can help improving statistical post-processing models. QRF is demonstrated to be beneficial for the predictions, with more skillful forecasts than LR for thunderstorm occurrence. Lightning intensity predictions are influenced by inhomogeneity of lightning detection datasets; despite inhomogeneity, skillful predictions can be made with both extended LR and QRF. The regional maximum of Modified Jefferson index and most unstable CAPE are found as best thunderstorm occurrence predictors and the regional minimum of Bradbury index and maximum of K-index emerge as best for lightning intensity. Neither most unstable CAPE nor microphysical predictors (graupel, snow) are essential for thunderstorm occurrence prediction.

Summary

Severe thunderstorms can disrupt society. Therefore, Royal Netherlands Meteorological Institute (KNMI) attempts to forecast these events and issues warnings for The Netherlands. To assist forecasters with objective probabilistic forecasts, a statistical tool has been developed previously. This statistical tool has been forecasting whether a thunderstorm occurs and additionally the lightning intensity for about fifteen years.

The current statistical tool is outdated; the input of the tool will not be available anymore by 2020. Moreover, the statistical methods have advanced since development of the statistical tool. Machine learning is now widely applied. Weather simulations by numerical weather prediction models have also become more realistic, importantly due to a finer resolution of these models. Therefore, an update of the statistical tool for thunderstorm forecasts is necessary.

To find out the best strategy for making the statistical tool a traditional statistical method, logistic regression, is compared to a machine learning method. The machine learning method is called quantile regression forest (QRF) and consists of many decision trees. Each of the trees split a dataset in many small parts ("leaves") differently as if they are hundreds of different jigsaw puzzles of same picture; all the trees issue a forecast and these forecasts are aggregated into one forecast. Moreover, it is tested whether specific important information about the physical state of the atmosphere in the weather simulation is essential or not. Output from daily weather simulations over the summers of 2015-2017 with the KNMI weather model Harmonie for the next 3 to 45 hours has been statistically connected to lightning detections.

It is found that the machine learning method (QRF) improves the probabilistic forecasts about whether thunderstorms will occur or not, compared to the more traditional method. One of the main reasons for this is that the machine learning tool can exploit multiple variables fruitfully, even if they are statistically strongly related to each other. For forecasts of lightning intensity, it cannot be demonstrated that the machine learning method is better than the more traditional method. A main reason is that the operational set of lightning detection sensors has changed in time. Therefore, the number of detected lightning discharges is much larger in 2016 and 2017 than in 2015 and it makes forecasting the correct (measured) lightning intensity by this system a harder task.

From the physical point of view, it is demonstrated that none of the information represented by individual physical variables is essential to the new statistical tool made. If a certain type of information is removed, it is largely buffered with related information. High lightning intensities are forecast in an air mass that is warm and contains a lot of moisture, in which aspirant storm clouds would have the tendency to rise fast in the vertical direction once they are formed (due to instability). This is consistent with expectations. Lastly, the so-called Modified Jefferson index is the most important variable to isolate potential thunderstorms from non-thunderstorms situations and it is also easy to compute.

Acknowledgements

This thesis has been both a big statistical and meteorological exercise. Before, and during my studies (of both Soil, Water and Atmosphere as BSc. and Climate Physics as MSc.), I have had the opportunity to absorb plenty of meteorological insight and knowledge. My statistical knowledge has evidently expanded more than significantly during this project. For this, I got much help from Maurice and Kiri. They also provided me with much guidance in every step: giving feedback and the opportunity to practice my presentations, such that the presentations did not end up lasting an eternity and Kiri's help during some of my struggles with R.

Regarding the expansion of my statistical knowledge in the past eight months, I am also very thankful to the other members of the statistical post-processing group of RDWK (in particular Kate and Jasper as experts in statistics). KNMI also offers the opportunity to talk with many other experts in operational meteorology and forecasting, in lightning detection systems, weather models, and so on; I want to thank all of them for sharing their knowledge and views with me. The script "Gevaarlijk weer indices" was essential for this project. Maurice has kept this script up-to-date with others and ran it; I want to thank Rudolf van Westrhenen for programming this. In addition, I want to thank Maurice, Lotte de Vos, Hans Beekhuis (all KNMI), Stephané Pedeboy (Météorage) and Egon Wanke (Blitzortung) for the information they provided on lightning detection systems.

Without the friendly IMAU environment, Climate Physics would not have existed. This is a success of the IMAU staff. In particular, I want to thank Willem Jan for his help during the project and giving his views every once in a while. I would also like to thank my fellow master students at IMAU for the nice study environment and especially those that have become my friends. Lastly, I am very thankful to my parents for their long term and persevering daily support during my school and study time.

Becoming an "expert" in meteorology and climate has been my dream for about 15 years and I knew I would likely make it come true by 2019..... Now it is time for another challenge which will further allow me to absorb even more meteorological knowledge; a never-ending journey!

Contents

List of Figures

List of Tables

1 Introduction

1.1 Aim

Thunderstorms are important phenomena to forecast, because these can be hazardous to airplanes due to the strong turbulent motions in and around convective clouds, but also due to the risks involved when persons, buildings or infrastructure are hit. More importantly, accompanying phenomena like wind gusts in excess of 25 m/s, (large) hail and flash floods may disrupt society. These phenomena contribute strongly to European insurance pay outs [Munich Re, 2016].

Recently, numerical weather prediction (NWP) models have increased horizontal resolutions to simulate deep convection explicitly. Additionally, statistical methods for post-processing of NWP output have advanced. We aim to utilise this by developing new probabilistic (severe) thunderstorm forecasting models and compare these with existing model set-ups. The focus is to find out how to make optimal thunderstorm forecasts based on NWP output for up to 45 hours ahead, by first of all comparing the machine learning technique quantile regression forests with logistic regression and second understand which physical variables simulated by the NWP are most relevant. We forecast thunderstorms over The Netherlands in the period from April 15^{th} to October 15^{th}, as the thunderstorm season usually lasts from May until September in The Netherlands and surrounding regions [Taszarek et al., 2019]. The intention is to predict probabilities of thunderstorm occurrence optimally and if thunderstorms occur their intensity, which is based on the lightning detection system that KNMI uses.

1.2 Statistical post-processing: why and how?

Numerical weather prediction model output can be used in its direct form for some specific purposes, although some post-processing has to be applied in all cases: for example re-gridding from NWP levels to near-surface values. Nonetheless, there are several reasons why the numerical weather model will not represent exactly the conditions that are observed at a measurement site:

1. The numerical model represents grid box average conditions, whereas the weather station represents a point value.

2. The model may have a bias (systematic error) for some predicted variable in a certain region.

Furthermore, there are additional reasons to do statistical post-processing:

3. A deterministic forecast with direct model output cannot lead to probabilistic forecasts, without a statistical post-processing step.

4. Statistical relationships can take into account some variability around the direct model output to account for model errors that add some uncertainty to the forecast. Usually the uncertainty is also incorporated by making ensemble predictions.

One could even argue that statistical post-processing is even more crucial for thunderstorms and deep convection, as it is a very non-linear phenomenon for which NWP models have large problems in solving processes and locations of occurrence. Large scale average conditions are typically predicted better (see for example [Bauer et al., 2015]). Therefore, an aggregation step in post-processing model output is very helpful in making lightning forecasts.

For many purposes weather forecasts can thus be improved by not using direct or raw output, but by statistically connecting NWP model output with observation datasets, here lightning detections. This can be done in a very direct way by deriving a statistical relationship, for example between temperature records at a weather station and some temperature output for a nearby gridpoint at surface height. The observed temperature record in that example is a predictand (dependent variable) and the NWP model output for temperature at a nearby point in space is the predictor (independent variable). This statistical relationship can often be improved by including multiple predictors, such as relative humidity and/or wind speed when improving temperature forecasts. In general, it can be said that statistical post-processing

uses a set of variables (predictors, typically model output variables), that may result from multiple NWP models, to predict observational records of the predictand, using a statistical relationship. The prediction based on the statistical relationship that was obtained has to be verified with an independent verification set from the same type of observational record at the same locations as for which training was done. Another improvement can be including ensemble information for statistical post-processing. Since grid boxes, biases and errors in physics vary between different NWP models, one should in principle update a statistical post-processing model when NWP models are updated.

In statistical post-processing, meteorology and statistics go hand in hand. This means that statistical optimisation approaches are applied in standard studies. However, one also wants to take the meteorological relationships into account. Only by using advantages of knowledge in both disciplines, one can come close to good models that work in general cases and take all atmospheric variability and uncertainty well into account. The statistical part involves choosing appropriate statistical models, having well defined, useful and computationally efficient algorithms for fitting predictors, doing reliable cross-validations and bootstrapping to assess uncertainty and modifying meteorological variables such that they are optimally in line with assumptions of statistical models, for example using mathematical transformations. Cross-validation is defined as the procedure of making (often three) subsets where clearly dependent samples are kept in the same subset and then using these subsets for both training and testing statistical models. Each of the subsets is then once used as verification set and twice as training set. Thereby spatial correlations, correlations among predictors and correlations in time in predictors and predictand have to be taken into account. The meteorological task involves deriving meteorologically relevant and potentially complementary predictors. Furthermore the meteorological task involves interpretation of what relations among predictors are important and what the statistical models imply for forecasting certain weather phenomena, here thunderstorms and their intensity. It is important to understand that two different predictors with very similar meaning are from the statistical point of view seen as covariates that may also complement each other in some cases and in meteorology, the same two variables can be seen as two different ways of mapping the same type of (potential) atmospheric behaviour.

1.3 Previous thunderstorm post-processing models

First, some recent studies using statistical techniques for post-processing of short-term NWP output and/or satellite and radar imagery for convection-related weather will be reviewed. This starts with a discussion of the previously made and currently still operational thunderstorm post-processing model used by the Royal Netherlands Meteorological Institue (Dutch: Koninklijk Nederlands Meteorologisch Instituut, KNMI), followed by results obtained in studies comparing new statistical techniques for post-processing and more traditional statistical techniques in Section 1.4.

In the 2000s, a thunderstorm forecasting system was built at KNMI based on NWP output, first optimising the choice of input variables for so-called logistic regression equations (Section 3.2.1) from a list of potential predictors and then optimising the statistical equations [Schmeits et al., 2005, 2008]. In this forecasting system, ECMWF convective precipitation is the most important predictor for thunderstorm occurrence forecasts followed by Hirlam instability indices, mainly CAPE (convective available potential energy), Jefferson index and slightly less frequently Boyden index. [Haklander and Van Delden, 2003] assessed instability indices as thunderstorm occurrence predictor over The Netherlands and Boyden index and CAPE of the most unstable layer (MUCAPE) also appear to be highly ranked. They have used radio sounding data to compare instability indices. Bradbury index was most useful as lightning intensity predictor by [Schmeits et al., 2005, 2008]. Aforementioned predictors are briefly introduced in Box 1. Further discussion of thunderstorm prediction and some relevant theory will be the main topic of Chapter 2.

Box 1: A brief description of some important thundstorm predictors
Convective instability indices are predictors for convective weather, such as thunderstorms, that have been developed mainly in 1950s and 1960s, before numerical weather prediction was common. Observations made by a weather balloon would provide information on whether these types of weather could happen later that day. Frequently used variables in convective instability indices are for example air temperature, dew point temperature, potential wet bulb temperature and geopotential height at standard levels. Therefore one could say they are bulk approximations to the state of the atmosphere. A few common examples are now discussed.

The **Boyden index** is defined as

$$Boyden\ index\ = 0.1(z_{700} - z_{1000}) - T_{700} - 200$$

Here, z_{level} indicates the geopotential height of the respective pressure level and T_{level} indicates the temperature at the respective level. Note that thickness between 1000 hPa and 700 hPa layers is proportional to the mean temperature in this layer, which means that Boyden index relates closely to a vertical temperature gradient.

The **Bradbury index** is defined as

$$Bradbury\ index\ = \Theta_{w,850} - \Theta_{w,500}$$

Here, $\Theta_{w,level}$ is the potential wet bulb temperature at the specified pressure level. It describes potential instability of the atmosphere. Potential instability is a type of thermodynamic instability in which conditions are favouring atmospheric convection after a large scale lifting process has taken place (that has brought a lower layer from which convection would initiate to saturation).

The **Jefferson index** is an index that describes instability of the 925 (also used: 850 and 900) to 500 hPa layer in a more empirical way.

$$Jefferson\ index\ = 1.6\Theta_{w,925} - T_{500} - 11$$

It does not directly compare a parcel subject to latent heat release with its environment, but leads to a good discrimination between stable and unstable conditions leading to convective storms if latent heat can be consumed. A **modified** version of this index includes the dew point depression ($\frac{1}{2}(T - T_d)$) at 700 hPa, to account for dilution and cooling of rising parcels by entrainment when mid-levels are very dry.

Convective available potential energy (CAPE) is the vertical integral of the buoyancy of a parcel, assuming an adiabatic process. It has interpretation as the maximum kinetic energy a parcel initially at rest could theoretically acquire, without interaction with the environment and is therefore not a convective index, but has a clear dynamical meaning.

$$CAPE\ = \int_{LFC}^{LNB} g\frac{T_{v,p} - T_{v,env}}{T_{v,env}}dz$$

The **level of free convection (LFC)** is the level where the parcel would be neutrally buoyant, before it becomes postively buoyant. The **level of neutral buoyancy (LNB)**, also known as equilibrium level, is where the parcel would be neutrally buoyant at the top of a CAPE layer. When a parcel is negatively buoyant (cooler than its enviroment), this integral is called **convective inhibition (CIN)**; the upper limit of the integral is then LFC.

A more complete overview of convective indices and their interpretation is given in [Haklander and Van Delden, 2003]. Definitions of all predictors can be found in Appendix A.

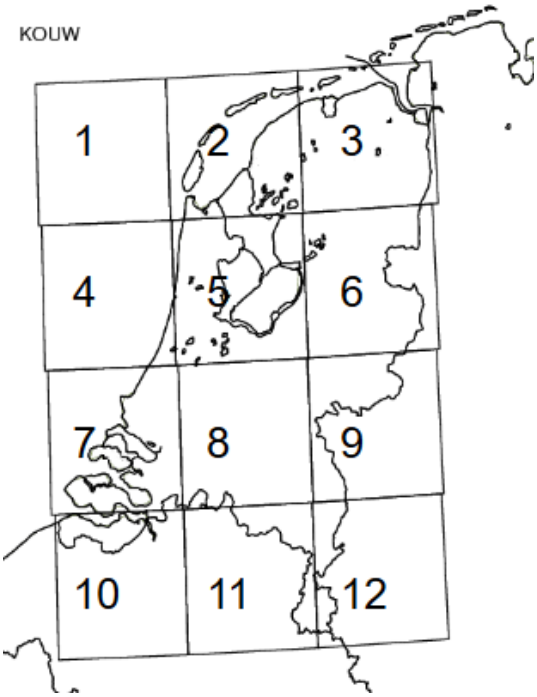


Figure 1: Map of The Netherlands with KOUW-regions drawn as boxes with an indication of the region number.

The intention is to make an update of the currently operational probabilistic thunderstorm forecasting system for weather warnings/alarms (acronym: KOUW), using the same KOUW-regions and somewhat similar predictors. KOUW-regions are twelve boxes that were defined to be used in operational weather forecasts by KNMI, each with an area of approximately $7000 km^2$ and together they cover the Netherlands, bordering parts of Belgium and Germany and the North Sea coastal region (Figure 1). Experiments are done with more potential predictors from NWP output, but in contrary to the 2008 system advected radar and lightning information will not be used as predictor sources.

It is explicitly pointed out that the forecast skill cannot be directly compared with the previous post-processing models for thunderstorm forecasts: first of all, the current lightning detection system "KLDN" that is operational at KNMI is strongly different compared to the former lightning detection system ("FLITS"; see [Noteboom, 2006]). Second, the used NWP model, Harmonie, solves deep convection explicitly; the previous KOUW-system was based on predictors from Hirlam and ECMWF in which all convection was parameterised. Third, the strategy applied to get to final post-processing models is slightly different. One of the consequences of this third point is that no separate logistic regression equations for thunderstorm occurrence are derived for each of the KOUW-regions[1].

Regarding "FLITS" and "KLDN" lightning detections, [De Vos, 2015] has found similarities, but also many differences. The relation between detections obtained with both sets is very obscure, such that both detection systems would classify events differently, even if events are compared to the climatology of their own detection system. The new detection system may have more problems in discriminating between severe and non-severe storms using discharge rates. Results from [De Vos, 2015] have indicated that this was indeed the case for the 2010-2014 KLDN lightning intensities in comparison with corresponding FLITS lightning intensities, observations of hourly rainfall and hourly wind gust at nearby KNMI weather stations.

[1]Both separate fits per region and combined fits for all regions have been experimented with, using the machine learning technique that we apply, namely quantile regression forests. Since the average verification scores were slightly worse with separate fitting, variability between minimum and maximum scores overlapped strongly, because the Dutch topography is relatively homogeneous and because occurrence and non-occurrence would be more in balance, combined fits were preferred.

1.4 Recent developments

However, as mentioned statistical post-processing methods and NWP models have advanced. This means that the two most important factors of added value due to recent developments are the implementation of machine learning methods (in particular quantile regression forests, abbreviated as QRF; [Breiman, 2001] and [Meinshausen, 2006]) and higher resolution NWP output, with related model modifications. Due to the model resolution (2.5 km in Harmonie versus 11 km previously in Hirlam) deep convection is solved explicitly with Harmonie. It leads to an explicit description of vertical velocity along with more direct and detailed description of cloud processes [Bengtsson et al., 2017]. A large dataset with from a non-hydrostatic NWP model has not been used so far to study thunderstorm occurrence predictors as far as we are aware. In addition, it should be realised that the general forecasting performance of NWP models has improved in time [Bauer et al., 2015], which should also contribute to (slightly) better forecasting skills after post-processing when a common reference is used. However, as opposed to these reasons to expect better performance, a reason for which performance at high intensities could be worse, is the non-stationarity of the new lightning detection system data that is operationally used at KNMI.

Comparing QRF and more traditional post-processing techniques like (extended) logistic regression for numerical weather prediction has recently been done by several authors. QRF is a statistical technique, in which typically several hundred decision trees are created with a dataset; each of the trees makes a prediction of the outcome of unseen samples and the whole forest gives a robust prediction.

The problem of severe convective storms has been investigated over The Netherlands from the rainfall perspective by [Whan and Schmeits, 2018]; in this study, similar Harmonie output was used as potential predictor set from an older model version. In their study, it was amongst others found that the QRF method can profit significantly from adding more variables than just precipitation for forecasting storms, whereas more traditional methods usually gain less information from the indirect predictors. For our lightning intensity statistical post-processing study, the distinction between direct and indirect predictors is not clear, because NWP models are not able to simulate lightning discharges explicitly. In other words, direct predictors are non-existent. In traditional statistical post-processing models, the non-linearity in predictors cannot always be fitted appropriately, for example for bimodal or in general multimodal distributions. In some cases this can be partly corrected using transformations. But the non-linearity is no issue for QRF. A predictand can also be bi- or multimodaly distributed, when applying QRF to (ensemble) forecasts [Taillardat et al., 2016]. For the thunderstorm forecasts multimodal distribution could only be potentially present in conditional lightning intensity forecasts, because forecasting it is a continuous regression task, whereas thunderstorm occurrence forecasts involve a binary classification. Additionally, fits that are biased towards certain percentiles (bulk) of the distribution are prevented in QRF; also a result found in the aforementioned study.

A drawback of the QRF method that was found by [Whan and Schmeits, 2018], is that it cannot issue high probabilities for climatologically extreme values. Here $q_{0.97}$, the 97^{th} percentile, was mentioned in particular. This is because QRF is not able to isolate the tail of the distribution, beyond $q_{0.97}$, very sharply from the rest of the distribution. In other words, when an attempt is made to isolate the tail of the distribution (of hourly precipitation) from the bulk with the predictors, the tail is still mixed with a portion of the bulk of the distribution. Note that this happens with other methods as well, otherwise a method would be found superior to QRF and other methods. It is interesting to find out what quantile of discharge rate can be forecast skillfully (and what reliably) using a conditional predictand, as in this study. Nonetheless, since a thunderstorm forecast is made instead of hourly precipitation here, results are not directly comparable. In the QRF framework, the testing that was done by [Whan and Schmeits, 2018] will be expanded, by testing a wider range of settings to optimise the predictions done by QRF.

Promising results have also been found in applying the random forests (RF) technique to the problem of convective initiation, using satellite imagery and NWP output data combined [Mecikalski et al., 2015]. Like QRF, RF is a method that uses hundreds of optimised decision trees based on a dataset to predict the outcome of unseen samples. They conclude that NWP and satellite retrieved information go hand in hand when improving nowcasts. Despite their benefits when nowcasting convective initiation, for longer

time scales (12-48 hours lead time) the satellite imagery is not usable, because even long lived convective systems will hardly survive for such a long time. Therefore, this study will rely on NWP output data. One disadvantage of their approach is that the satellite imagery fields have a very high resolution, which means that there are many data available for each case; this limits the number of cases processed when having the same computational capacity limits, with spatial coherence potentially leading to highly correlated subsamples in the dataset. The approach that will be applied here, namely the lower resolution by aggregating NWP grid cells to KOUW-regions, leads to higher predictability.

A very similar study to [Mecikalski et al., 2015], was done by [Ahijevych et al., 2016]. They forecast initiation of mesoscale convective systems (MCS) over the US with RF. An MCS is an example of a severe thunderstorm case and is intended to be forecast with the post-processing models, although these convective systems do not fully overlap with the high lightning intensity events. Apart from CAPE as an instability indicator, their study found that important predictors are terrain height (not so relevant for the Netherlands), skin/surface temperature, precipitable water, valid time and radar imagery extrapolation. Model precipitation appeared in the middle of the predictor ranking, whereas [Schmeits et al., 2005, 2008] find convective model precipitation as most important predictor. The convective indices in their potential predictor set were much more restricted than ours. Not all of the important predictors are important in every US region. The more homogeneous terrain characteristics in The Netherlands imply that The Netherlands would have a clearly different importance ranking if the same predictors would be used for the same forecasts.

Another important finding in their study was that the ratio between occurrences and non-occurrences of events (that means threshold exceeded versus not exceeded) should not be too small. Their full training dataset, with 0.3% mesoscale convective system initiation samples, was giving worse results than when the ratio was lifted to 30% using a subsampling strategy. In our study, the ratio of 0.3% corresponds to forecasting conditional severe thunderstorm probabilities at extreme intensities ($q_{0.997}$). An additional advantage of the use of conditional probability of ligtning intensities, is that inconsistency in probabilistic forecasts is impossible (for example 35% probability of a thunderstorm and 40% probability of a thunderstorm with at least 50 discharges per 5 minutes is impossible).

Summarised, it can be said that different studies that relate closely to this study have been done. They have statistically post-processed NWP output and in some cases additional satellite/radar data to forecast deep convection and its severity. Satellite and radar imagery are useful tools in the first few hours and beyond NWP output such as instability indices and precipitation are useful. Which instability index contains the most appropriate information, may depend on many factors and notably on the predictand. CAPE is often used and appears informative in these cases. Additionally, (Q)RF has been most extensively applied for the very short term (nowcasting) in the context of atmospheric convection.

1.5 Set-up of this study

Aforementioned summary of findings from previous studies addresses some of the research interests; the general purpose of this study is to understand how thunderstorm forecasts can be improved on the short term up to +45 hours, without the specific interest to improve their nowcasting. This is the main feature of this study and therefore conventional thunderstorm predictors, such as CAPE, Boyden Index and other instability indices, and moisture and precipitation indicators (see Appendix A) are investigated and the added value that combinations of variables and transformed variables can have, for the applied techniques: logistic regression, extended logistic regression and QRF.

The next aim is to understand thunderstorm predictors in the Netherlands and surrounding areas better for forecasting applications, potentially with new predictors. This is mostly in line with the study by [Whan and Schmeits, 2018], but with slightly different experiments and a wider variety of potentially interesting predictor variables. A main focus in the predictor context is whether the microphysics representation in a non-hydrostatic model (Harmonie) and CAPE as an integrated measure of buoyancy when lifting a parcel are of essential value compared to for example non-integrated instability predictors for thunderstorm forecasts.

Having introduced statistical post-processing and some previous studies, the following chapter will cover essential theory on how thunderstorms begin and can be forecast. This is followed by the third chapter on the datasets and specific statistical post-processing methods that are applied. The fourth chapter describes thunderstorm climatology, with some climatology conditional on time and place to assist interpreting the statistical models and some elaboration on homogeneity of the lightning detection data. The results by statistical class of the post-processing models will be discussed in Chapter 5 and in Chapter 6 the physical interpretations and predictor experiments will be presented in more depth. Some extra figures can be found in appendices. Then some case studies are done in chapter 7 to show potential weaknesses of the statistical models. This is followed by some final discussion in Chapter 8 and conclusions (Chapter 9).

For data analysis and model fitting, the programming language R is used; specifically relevant packages are cited.

2 Thunderstorms: physical mechanisms & numerical prediction

After introducing developments in numerical weather prediction from the last years with Harmonie as the example used (Section 2.1), this chapter discusses theory about how thunderstorms and accompanied lightning intensities can be forecast (Section 2.2). In addition, the "most unstable" and "surface-based" layers used are defined (Section 2.3). One of the interests of this study is to find out whether new useful potential predictors can be composed from previously used potential predictors. The three categories of new predictors are described in Section 2.4, with in addition a short summary of available Harmonie variables in that section. Lastly, NWP errors with which statistical post-processing models deal will be discussed as well (Section 2.5).

2.1 Simulation of (deep) convection with Harmonie

In recent years, numerical weather prediction has advanced: non-hydrostatic models are widely used. Since Harmonie is a non-hydrostatic model, it simulates vertical velocity as a state variable and is able to solve deep convection explicitly (see Chapter 1). Shallow convection still has to be parameterised with a resolution of 2.5 km [Bengtsson et al., 2017].

With the microphysical parameterisation Harmonie can simulate various types of hydrometeors within showers separately, namely rain, snow, graupel, cloud ice and cloud water [Bengtsson et al., 2017]. The hydrometeors are involved in lightning initiation according to literature, as will become clear in Section 2.2. With Hirlam and ECMWF, these hydrometeors were not separately simulated and could not be used as separate thunderstorm predictors. As a consequence of improved hydrometeor representation, the interaction of hydrometeors with other processes is also affected. Other parameterisations have also been modified; a detailed description of the version of Harmonie that is used, is provided by [Bengtsson et al., 2017].

Because of the high resolution of Harmonie output, in practice regional minimum, maximum and/or mean of physical fields will be used as predictors for thunderstorm forecasting models. The exact way of deriving such statistical measures from Harmonie output is described in Chapter 3.

Some physical predictors that are identified to relate to thunderstorm occurrence, lightning intensity and/or dynamics of convective cells and are derived from Harmonie output will be printed in *italic* in the next section.

2.2 Theory and forecasting of thunderstorms and deep convection

2.2.1 Initialisation of lightning discharges

Physical mechanism

The initialisation of the occurrence of lightning discharges in the atmosphere is very strongly associated with deep convection and the presence of various hydrometeors in convective clouds, namely ice, water droplets, snow and graupel (soft hail). The first focus will be the understanding of the role of hydrometeors in the occurrence of lightning in convective clouds.

Originally, calculations by [Takahashi, 1978] based on his experiments have confirmed that the interaction of cloud ice and graupel upon collision is responsible for the initialisation of lightning discharges in sufficiently strong updrafts. By accumulating electrical charge in regions with different cloud temperatures, the electrical field that triggers a discharge is thought to be produced. Regions of the cloud around -10 $^{\circ}C$ and around -20 $^{\circ}C$ are often transition regions with a negative charge being observed between these isotherms and positive charge outside this temperature range [Takahashi, 1978]. A discharge can be produced when the negative charge region is opposed on both sides by the region positive charge with sufficiently strong gradients in electrical charge. However, the exact mechanism that leads to a discharge is currently not known [Lopez, 2016].

Literature above suggests three important criteria for thunder occurrence based on the role of convective updrafts and hydrometeors within: presence of cloud graupel and ice content; presence of thermodynamic instability that could cause vertically developed clouds with sufficiently strong updraft; a convective cloud top temperature of about -20 $°C$ or lower.

Numerical modelling with NWP output

Whether the last of the three conditions is fulfilled in Harmonie is not directly represented in the Harmonie dataset, but the first two are; as both cloud ice and cloud *graupel content*[2] are simulated by Harmonie, it can directly be derived whether the first criterion could be fulfilled in the model given a correct forecast. The second criterion of thermodynamic instability is well represented by *CAPE*, because that is a vertically integrated measure of buoyancy of parcels. Additionally, for the third criterion cloud top temperature can be seen as approximately a function of the thermodynamic conditions in the feeding air mass and the height to which the shower could reach; these two are well reflected by the air mass properties in the source air, namely the *potential wet bulb temperatures* Θ_w *in lower layers of 850* and 925 *hPa*, and *level of neutral buoyancy*. Level of neutral buoyancy (LNB) is the top of a positively buoyant layer when a parcel is (commonly adiabatically[3]) lifted; above it rising parcels will decelerate, but cloud parcels with significant upward velocities at this level will clearly temporarily overshoot this level. Some additional processes also play a role in determining the eventual cloud top temperature, such as entrainment and phase transitions occurring in the cloud.

The convective cloud top temperature could in be derived in a more direct way from Harmonie output, because a cloud top temperature is calculated from the pseudo infrared satellite image. Processing it to convective cloud top temperatures would require masking of the top temperatures with vertical velocity field to extract convective clouds and hence combining Θ_w and LNB to approximate it is easier.

Alternative ways to exploit the aforementioned three physical criteria for modelling occurrence of lightning discharges will also be fruitful. [Lopez, 2016] uses primarily CAPE as proxy for updraft velocities and derived graupel content in convective clouds to parameterise discharge densities for ECMWF IFS. In addition, derived liquid water and ice content are used.

2.2.2 Ingredients of deep convection

Physical mechanisms

Three main ingredients have been used in literature to explain deep convection and assess potential initiation ([Johns and Doswell, 1992] and [Doswell et al., 1996]):

1. Steep lapse rates governing thermodynamic instability (covered in Section 2.2.1);
2. Moisture in the layer from which storms are extracting potential energy due to the thermodynamic instability;
3. A lifting mechanism that allows a parcel to be lifted to level of free convection.

This lifting mechanism can both act on large synoptic scale and mesoscale. It is important to realise that both enriched moisture and upward motion enlarge the thermodynamic instability locally and also decrease convective inhibition and related potential suppression of convection by an inversion, by increasing the probability of parcel condensation with subsequent latent heat release and increasing CAPE.

Besides this, deep convection is a very non-linear process which is influenced strongly by neighbouring environment; shallower convection can be a moisture source for the lower part of the free troposphere in a convective environment, but sink for the boundary layer. Furthermore, cold pools are important initiation mechanisms due to which convective cells have interaction (see [Markowski and Richardson, 2010]).

[2] Only graupel was selected as very important, because the frozen cloud content predictors appeared to correlate strongly and not complement each other so strongly

[3] In the applied computations entrainment is taken into account, as described in Section 2.3 and the adiabatic assumption vanishes.

Lifting mechanisms can both arise on small and large scales; extratropical cyclones usually induce low level convergence near their centre and are accompanied by ascending motions, especially if they are deepening or are deep. Using quasi-geostrophic theory, lifting processes that are involved can be understood. Warm air advection (or thickness advection) and cyclonic vorticity advection, which occur near centres of low pressure systems and mainly at frontal boundaries, can be identified as diagnostics for large scale lifting. In addition there are small scale lifting processes, which may relate to subregions of large scale lifting and to differential diabatic heating of an air mass. Differential diabatic heating can lead to smaller scale (mesoscale) low pressure systems and/or convergence zones in the lower troposphere. An example of differential diabatic heating and resulting mesoscale ascent near a heated region is the sea breeze circulation.

Numerical modelling with NWP output

The previous part of this section suggests that besides different indices of convective instability, water vapour itself and enrichment in water vapour can help predict thunderstorm occurrence; moreover predictors can be used to get information about both lifting motions and water vapour content simultaneously. *Water vapour convergence* in the boundary layer is typically caused by convergent flow in the boundary layer, which has a typically higher water vapour content than the free troposphere; regions with water vapour convergence are expected to have (relatively) high integrated water vapour contents or *precipitable water* (PW) as well, because most of the moisture content is typically in the boundary layer. Higher probabilities of hydrometeor contents in clouds and low *convective inhibition* would often be expected in the same regions.

The large scale lifting processes be induced from pressure fields, vorticity fields and temperature fields. Unfortunately, thickness advection and cyclonic vorticity advection are not available from Harmonie[4]. Typically, lifting or ascending motions are also associated with dropping pressure/geopotential heights in the lower troposphere (negative tendency and lower values). Apart from direct diagnosis with vertical velocity or thickness advection and cyclonic vorticity advection, information as water vapour convergence, *mean sea level pressure* and *pressure tendency* could be proxies for lifting mechanisms as well.

2.2.3 Additional potential sources of predictability

Physics

Thunderstorm climatology could give information about likelyhood of thunderstorms, because The Netherlands typically has a northwest-southeast or north-south gradient in occurrence and intensities of thunderstorms [Taszarek et al., 2019]. Additionally, thunderstorms in late summer and autumn are typically different in being concentrated on the coast and less intensive in terms of lightning intensity than severe summer cases (discussed in Section 2.2.5). This means that there are arguably different thunderstorm climatologies in the Netherlands and the proximity of the coast could therefore be a successful predictor.

Apart from the suggested physical fields that can give information on the occurrence or initiation of thunder in a convective storm, the dynamics that is known to explain the initiation and behaviour of deep convection cells can likely help us understanding where thunder could occur. The most important variables for dynamics of thunderstorms are arguably deep layer wind shear and storm relative helicity (from here on just helicity), which explain mode of convection. They can also affect initiation and occurrence of thunderstorms [Markowski and Richardson, 2010]. Therefore they should be physically informative in our study.

Numerical modelling with NWP output

The *coast* variable is determined as being true for KOUW regions 1-5 (North Sea, Wadden Sea and IJssel Lake) and 7 (North Sea coast, southern part); see Figure 1 in Chapter 1. For region 10, one can discuss about whether it is coastal or not, covering the Westerschelde mouth and a tiny bith of North Sea. It was chosen to call this an inland region, as only a very small part is actually sea.

[4]It was investigated whether ECMWF predictors of the indicators of upward motion and ECMWF convective precipitation would add value to the post-processing models, but they were assessed as not strongly complementary to the available Harmonie predictors. Similarly, mean, maximum and minimum of vertical velocity at about 3.5 km height in Harmonie have been investigated, but their information did not add much to already available predictors as column graupel and CAPE.

Unfortunately, the directly derived deep layer shear from near-surface to 6 km is incorrectly formulated in the Harmonie reforecasting output and exactly equal to the 500 hPa wind speed. However, since 850 and 500 hPa wind speed and direction are available, the *bulk shear* over this layer can be derived. This shear measure is less informative than deep layer wind shear with the lowest layer of the atmosphere also included, when convection is triggered from the surface. This is usually from around noon to just before sunset. Helicity over the lower 3000 m is also calculated from Harmonie output.

2.2.4 Potential instability

Physical mechanism

The last point that needs to be made is that CAPE assess conditional instability, which could exist shortly. It can be preceded by purely potential instability. Given that potential wet bulb temperature Θ_w decreases with height, potential instability exists. The instability is released with lifting after condensation in a previously unsaturated layer. It can occur because the upper layer will continue cooling at dry adiabatic pace, while the lower layer already gets saturated and subsequently cools slower with height, due to latent heat release. After condensation of the lower layer the lapse rate increases above that saturated layer, eventually leading to (conditional) instability.

Numerical modelling with NWP output

If the larger scale condensation is not simulated by a model, the conditional instability might never be released in the model but purely potential instability that may be seen by the NWP model and subsequent condensation process would still allow for thermodynamically unstable parcels to be lifted over great depth in reality. Potential instability can be assessed with Θ_w using the *Bradbury index* and is by definition present with conditional instability. In wet climates such as The Netherlands, it is typically strongly correlated with some other 500-850 hPa indices, mainly Adedokun1 Index. The Adedokun indices are in essence just an expression of a temperature difference that a parcel from below would have compared to its environment if lifted adiabatically to 500 hPa. The strong correlation does physically make sense in wet climates; an explanation of that goes beyond the scope of understanding predictions of thunderstorms and deep convection over The Netherlands.

2.2.5 Lightning intensity

Physics

Arguably, lightning intensities can differ among convection modes, since mesoscale convective systems with squall lines can cover a large area of many thousands of square kilometres leading to potentially high intensities, whereas some small so-called single cells can cover much smaller areas and discharges can occur in restricted subregions. The area covered by convective storms is usually the biggest around LNB, due to the expansion of the parcel during its convective rise. This is because the pressure at LNB relative to the pressure in the inflow provides an estimate for parcel expansion during its convective rise. In addition, (storm relative) flow near the top of showers has impact on the area covered by convective cells. Therefore, the area covered is amongst others a function of level of neutral buoyancy and upper atmospheric and storm relative flow. Consistently with aforementioned arguments, one expects that smaller cells isolated cells especially during cold air advection events over the North Sea produce much lower lightning intensities than showers in warm air masses, partly because they are smaller and frequently isolated.

On the other hand, provided that lightning occurs, high lightning intensities are expected to occur when very warm and moist air masses with large latent instability are brought to The Netherlands. In terms of predictors this is amongst others associated with high potential wet bulb temperature (Θ_w) in the lower troposphere and high CAPE; most frequently a tongue of high Θ_w called "Spanish plume" is associated with such setting, also due to the forcing mechanism involved. This has been described in terms of synoptic setting by [Van Delden, 1998] with a case study. The relation between mesoscale convective systems and so-called Spanish plume events over neighbouring UK has been investigated by [Lewis and Gray, 2010] and they state that most MCS events in that region are related to this synpotic setting. The highest frequencies in the UK occur over Eastern Anglia, which is closest to The Netherlands.

Anomalous or "extreme" convective weather is typically associated with MCS systems and also with high lightning intensities in a study over Florida, according to [Williams et al., 1999]. They argue that high lightning intensities and other severe convective weather have a common cause, namely strong updrafts. Additionally, [Púčik et al., 2015] show that various types of severe convective weather over Europe are more likely when CAPE increases, except for wind gusts.

Numerical modelling with NWP output
With the parameterisation that [Lopez, 2016] has made for ECMWF, lightning densities are calculated. This implicitly means that one can assume that it simulates both occurrence and intensity of lightning, with intensity being related to the suggested predictors of hydrometeors and CAPE, the latter being meant as proxy for vertical velocity in convective cells. Lopez also uses cloud base height as giving information about the size of the convective cell in order to parameterise lower intensities in smaller cells, which is consistent with the previous paragraph.

With the physics described in this section (Section 2.2.5), it can be assumed that some prediction for lightning intensity is also possible with predictors that have passed in other parts of Section 2.2.

2.2.6 The elementary predictors

Thirteen important predictors of thunderstorm occurrence, intensity and dynamics have been printed in *italic* in Sections 2.2.1 to 2.2.5. They lead to a physically elementary set of 15 potential predictors that will be used as either proxies or direct information on whether a thunderstorm ingredient is present or convective storms are initiated in Harmonie, because helicity and moisture convergence have been included twice, as both regional maximum and minimum values. For the other predictors, one physically appropriate statistical measure has been selected. For CAPE and CIN, the most unstable layer values have been selected as most relevant, which will be defined in Section 2.3. The list of 15 potential predictors is found in Appendix A.

Other potential predictor sets will be defined in Chapter 3 (Section 3.5), because they have been made with information derived with the statistical methods that are described in Chapter 3.

2.3 "Most unstable" (MU) and "surface-based" (SB) layer settings

Most potential predictors are available as direct model output, or relatively easily computable, for example by taking the differences between two or three model fields. On the other hand, the methods applied to specifically asses CAPE, LFC, LNB and CIN are relatively sophisticated. In general, for these four predictors, the first choice that needs to be made is some initial parcel. Subsequently one can model the buoyancy of that parcel when it would be lifted. CAPE and CIN have been defined in Box 1 (Chapter 1). Important parcels are the "surface-based" parcel and the "most unstable" parcel, the latter is the parcel that has maximum CAPE. This section discusses parcel definitions that have been used to compute CAPE, CIN, LFC and LNB.

Methods applied have been used at KNMI systematically for over a decade. First of all, a mixed layer covering the lowest 500 m of the atmosphere is used as the surface-based parcel to calculate SBCAPE and SBCIN. The most unstable of all 500 m thick mixed layer conditions around the lower 35 layers of the Harmonie model are used for computations of the MUCAPE and MUCIN. These 35 layers are about 3000 m thick in total. After parcel condensation, both SBCAPE and MUCAPE calculations takes into account entrainment, which is assumed to take place above condensation level on a (vertical) length scale of 5000 meters. This implies that the adiabatic assumption commonly defining CAPE actually vanishes and CAPE does not consist of all the convective available potential energy anymore. The CIN is computed over the layer that is below the main belonging CAPE layer, or, if CAPE is zero, over the depth of the atmosphere up to LCL. Although this script mostly produced expected values, some predictor values had to be omitted, because some LFC and LNB values were found to be deep in the stratosphere at the highest model level, which could not be changed with the time necessary run it. Eventually, the values of LFC and LNB above the threshold of 14 km were removed and valid values were used. The script to calculate most

unstable layer values for CAPE and CIN was made for severe weather computations at KNMI [personal communication - Rudolf van Westrhenen] and eventually not modified by us.

When CAPE is not present, the LNB is set to the lowest level where condensation occurs (LCL) and LFC is set to a higher LCL, namely the highest that the script finds during the loop over 35 model layers.

2.4 Potential predictor variables

2.4.1 Variables derived from Harmonie output

An initial set of 67 NWP output variables is created for comprehensive potential predictor sets with variables having physically different interpretations. However, many of these are convective indices which somehow relate to atmospheric conditional, potential or unconditional instability with respect to small thermodynamic perturbations and its potential consequences over depth scales of kilometres/hundreds of hectopascals. Some other variables are indicators for cloud height and content, water vapour and precipitation. Additionally, wind speed and its components at some levels with respective spatial structure (wind shear, helicity) and thermodynamic combinations of temperature, moisture and height are available or computed. While for example convective indices can be strongly correlated, they can also complement each other to improve forecasts for specific cases.

Those of the variables serving as potential predictors are found in Appendix A; variables are selected as potential predictors in several steps, which are described in Section 3.5. This is because the selection procedure involves the statistical methods described in Chapter 3.

2.4.2 New combined predictors from Harmonie output

Three categories of new variables are used. The first category multiplies the regional maximum of Boyden index, minimum Bradbury index or a third variable with abundance of moisture. The moisture that is used is the regional maximum of precipitable water content and its logarithm. Since they combine an instability index with precipitable water (PW), they are called "PWinst"-indices. The third instability component used in "PWinst"-indices is $\Theta_{w,925,max} - \Theta_{w,500,min}$ or "Edward". Subscripts "max" and "min" refer to regional extremes of the predictor. It is similar to Bradbury in that it assesses potential instability (see Section 2.2.4), if no horizontal gradients in temperature and moisture exist. On the other hand, when there are horizontal gradients in potential wet bulb temperature within our regions, it also gives a strong signal. Such condition happens typically when a frontal zone enters the region. So the Θ_w component of this predictor could simultaneously indicate a mixture of potential instability and presence of a frontal boundary.

The reason for taking regional maximum and minimum values of the predictors above and subsequently combining them to new predictors is that when initially experimenting with the predictors, the grid cell values were not yet available. Nonetheless, since predictors described here turn out to be selected in the preliminary experiments, they are kept in the potential predictor sets and not modified to grid cell values. In the end, grid cell values could be used as well, but it was chosen to not do so, because of the huge predictor sets (Section 3.5) that are used, the phase in which the project was and the general purpose of experimenting when predictors look worth using due to preliminary selection.

The second category combines MUCAPE and MUCIN into a new predictor. This predictor is $\sqrt{MUCAPE + MUCIN}$. In addition, the same is done with SBCAPE and SBCIN. They are used to represent an updraft velocity scale in the potential predictor sets in addition to the squared velocity scale represented by just MUCAPE.

In the third category MUCAPE is combined with column integrated graupel, cloud water and snow contents into new predictors by taking their grid cell products. They are eventually power transformed, as indicated by their equations (Appendix A); transformation methods will be discussed in Section 3.1.3. The potential predictors indicate presence of conditional instability and the hydrometeor contents in the model. The idea is that increase in both MUCAPE and aforementioned hydrometeor contents indicate

that thunderstorms are more likely and also more likely to be intensive. It has some similarity to what [Lopez, 2016] has done for ECMWF lightning parameterisation.

The full potential predictor list is found in Appendix A.

2.5 Complicating steps from predictor dataset to predictand

It is important to realise that any predictor can represent both the environments preceding the actual thunderstorm and environments of the thunderstorm and even a mixture of both. To give an illustrative example: helicity on the larger scale can be used to infer the potential of supercells, which will usually give a thunderstorm with high intensity. On the smaller scale helicity can be strongly modified by circulation in and around a supercell system itself [Lilly, 1986]. This can produce even higher helicity very locally. The predictors used by the statistical post-processing model can profit from both the preceding (potential) thunderstorm environment as represented by Harmonie and from the actual simulation of the convective storm. Whether a convective storm is actually produced by the NWP model or not, can in the end be inferred from several variables, such as column graupel.

Imagine now the case that conditional instability is present in Harmonie over a certain region, with most unstable CAPE exceeding 500 J/kg in the NWP model and also in the actual outcome. In this case there are still four options:

1. both in the NWP model and in the lightning detection system, deep convection does not happen (correct rejection).
2. both in the NWP model and in the lightning detection system, deep convection does happen (hit).
3. in the NWP model deep convection happens, but in the lightning detection system it does not (false alarm).
4. in the lightning detection system deep convection happens, but in the NWP model it does not (miss).

If the NWP model is qualitatively good, cases 1 and 2 should more frequent than cases 3 and 4 and the graupel predictor as indicator for presence of deep convective clouds in the NWP is beneficial. However, if cases 3 and 4 happen too frequently, the statistical post-processing models will not select an indicator of deep convection such as high graupel content as important predictor, because it is not informative for the actual outcome due to NWP modelling issues with deep convection. Model errors in a NWP can result from imperfect initial conditions and divergence between the trajectory of reality and that of the model in time, so errors in for example CAPE are likely to be larger at a longer forecasting time than at a shorter forecasting time.

The last important point is to realise that the "truth" value is in the end only based on at least two lightning discharges per six hours within a region. Apart from the potentially strong influence of region boundaries, the threshold of more than one discharge and boundaries of time bins and the accuracy of the actual detection system influence the results.

All of the above stated factors can complicate the classification task of each record, without directly being present in the dataset.

3 Datasets, statisical models & methods

To make a statistical post-processing model based on NWP output, first the (potential) predictor and predictand datasets need to be defined and connected. Additionally, the dataset has to be modified such that the data are in accordance with the needs for optimal fitting (Section 3.1). It includes for example applying predictor and predictand transformations and generate subsets. Some tests are subsequently done to optimise settings. In the next step the methods for initial cross-validation to choose settings selected for final fits and subsequently final cross-validation are applied (Section 3.3).
The procedure is done with different potential predictor sets for the ordinary and extended logistic regression (LR) methods and the QRF method and for two different predictands, for all lead times separately. Based on the final cross-validation, QRF and (E)LR methods can be compared. Then conclusions regarding the relevant research aims can be drawn using verification methods (Section 3.4). Each of the statistical post-processing steps and methods are explained in the remainder of this chapter. Lastly, the potential predictor sets used are defined in this chapter (Section 3.5).

3.1 Definition of the predictions

3.1.1 Predictand definitions

There are two types of predictands, one of which is conditional on the other. The first one is the (non-)occurrence of at least two lightning discharges within a KOUW-region and within a time slot covering six hours, based on the KLDN detection system. These KOUW-regions have been defined in Figure 1 in Section 1.3 and previously in [Schmeits et al., 2005]. The second predictand is only trained on the part of the KLDN dataset in which thunderstorms occur and forecasts their intensity probabilistically, conditionally on the thunderstorm occurrence. The time slots last from 03 to 09 UTC, 09-15 UTC, 15-21 UTC and 21-03 UTC, synchronously with NWP output. For each lead time separate models are trained and tested (evaluated), so each region and time slot can be used multiple times. For verification of thunderstorm occurrence predictions, there is separation between the regions, but for training all regions are pooled, as one dataset. The separate training for each lead time is a general strategy in statistical post-processing called model output statistics (MOS; [Glahn and Lowry, 1972]), as the NWP model output contains more and more errors with a longer lead time, which can for example increase biases between forecast and actual outcome. Additionally, the physical mechanisms that can cause thunderstorms are expected to differ partially between day and night.

The threshold of at least two discharges is chosen, because a single detection is much more likely to be an erroneous lightning occurrence report than an event with two detections, as other features can lead to false detections. However almost all of the lightning occasions lead to multiple detections. In fact, in the detection dataset 92% of all thunderstorm occurrences show multiple discharges. It is important to realise that the relative frequency that lightning is observed differs per region and time slot, which may have effects on the training of the forecasting system and the final product, the MOS forecast. This will be discussed in Chapter 4.

Given that in a six hour period at least two discharges are detected in a KOUW-region, the maximum number of discharges over 5 minute time bins is the second and conditional predictand. The dataset with maximum number of discharges is therefore much smaller and strongly dependent on the climatology of thunderstorms. Note that the characteristics of the detection system have a huge impact on the maximum intensity in the dataset (see [De Vos, 2015]). In addition to differences among different lightning detection systems that can be found, one lightning system is generally not fully homogeneous due to sensor replacements and potential detection efficiency changes.

With lightning intensity as predictand, training and testing again is applied for each lead time and valid time separately, but verification (model testing) is not done for each region separately as for the (non-)occurrence of lightning. This is because the verification sample size would be reduced too much to get reliable scores for only a few samples, if all regions were validated separately (see also Chapter 4).

Thunderstorm occurrence is a binary predictand, for which probabilistic forecasts are made using both QRF and LR. From the statistical point of view, this is a classification task. On the other hand, conditional lightning intensity has a continuous distribution with an infinite number of classes and from the statistical point of view, predicting intensity is a regression task in which the predictand is described as the probability that a lightning intensity threshold is exceeded. Therefore, both prediction tasks require in principle different statistical methods.

3.1.2 Availability and pre-processing of potential predictors

We use the 00z runs of a Harmonie reforecast dataset (2015-2017 period). The Harmonie version used is the upcoming operational version at KNMI, called Harmonie40 HAP2. The output is contained in hourly grib files up to +48 hours and then binned in time bins of 6 hours that do not overlap, namely 03-09 UTC, 09-15 UTC, 15-21 UTC and 21-03 UTC. The resolution of Harmonie is 2.5 by 2.5 km, which allows for explicitly solved deep convection. The Harmonie run of April 23^{rd} 2015 was removed from the predictor (and predictand) dataset, because it was initially not fully processed. Additionally, due to some data processing error, 24 thunderstorm cases in 2015 in KOUW-region 6 have not been taken into account, due to an error in the data processing. Therefore a substantial part of the 2015 dataset is missing for this region. However, on the total of about 2200 thunderstorm cases and about 24.000 non-cases, this is a small number.

Gridded data at the high resolution model grid of Harmonie are first re-gridded to KOUW-regions, with statistical measures of the model fields within a KOUW-region covering all six hours as final potential predictor values. These statistical measures are minima, maxima, means and some quantiles, as can be found in Section 3.5.2.

Reforecast output covering mid-April through mid-October is used, as the warm thunderstorm season usually lasts from the end of April or May until approximately late September [Taszarek et al., 2019] and it is intended to train a model for this part of the year.

The full potential predictor set is found in Appendix A.

3.1.3 Transformation of predictors

In addition to using just the predictor values themselves, some of the predictors are also or solely used in transformed versions. First, all predictors are plotted in scatter plots against empirical conditionally observed lightning occurrence probability in a predictor bin and observed lightning discharge rates. Then, for the predictors showing indications that transforming them makes them better predictors for LR models, we allow a LR model to pick from several transformed versions of the original predictor p: for example $p^{\frac{1}{4}}$ and $p^{\frac{1}{5}}$. In other words, an experiment with multiple power transformations of the original predictor as potential predictors is conducted and the first selected power transformation is chosen as optimal power transformation. This is tested with a dataset containing reforecasts for multiple lead times. It leads to a discrimination in favour of LR, but for QRF transformations are in principle not necessary, due to its splitting approach as described in Section 3.2.3 and because the ranking of predictor samples does not change by transforming.

The mathematical expressions of transformations can be found in Appendix A.

3.2 Statistical fitting methods

3.2.1 Logistic regression (LR)

The logistic regression equation is often used to model the probability that an event happens, using one or multiple predictors. The general expression for this model is Equation (1) [Messner et al., 2014a].

$$P(y = 1) = (1 + exp(a_0 + a_1x_1 + a_2x_2 + a_3x_3 + ... + a_nx_n))^{-1} \quad (1)$$

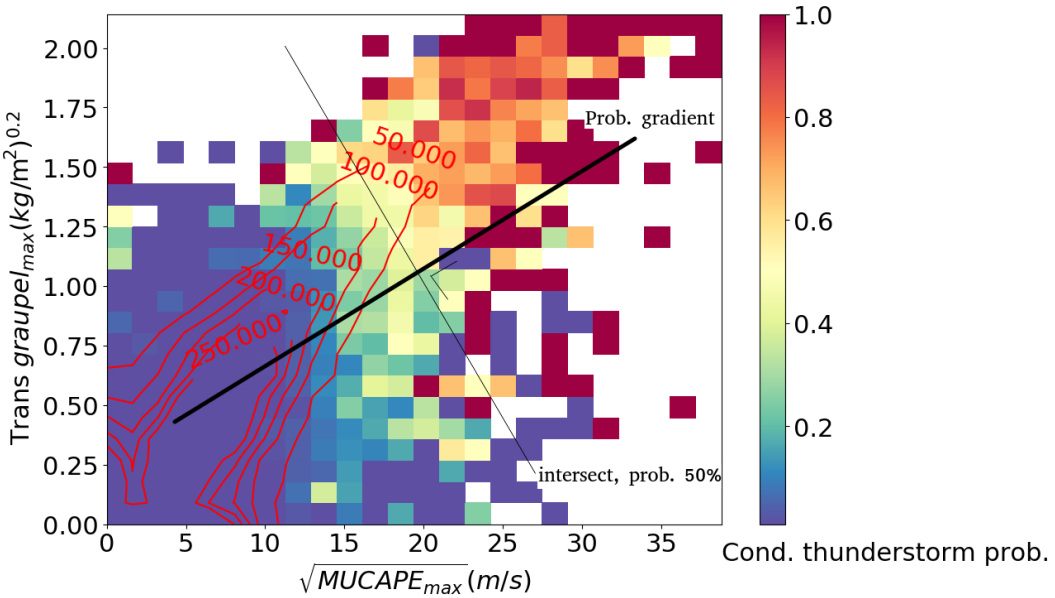


Figure 2: A figure to illustrate what happens if the two potential predictors would be selected, in this example square root of most unstable CAPE and transformed column graupel. Harmonie40 00z output from +3 to +27 hours is aggregated and gridded in 25 by 25 bins for the two potential predictors and the conditional thunderstorm probability per cell is calculated and given by the colour of a cell. Red lines indicate absolute frequency per grid cell. The "prob. gradient" indicates a line along which the conditional thunderstorm probability gradient may point and the intersect indicates where we approximately find the 50% probability of thunderstorms. Note that this figure is actually also a result from this study, but here used as illustration.

Here, $P(y = 1)$ denotes the probability that the event $y = 1$ is observed, x_m denotes a predictor and a_m denotes a model coefficient; here m can run from 1 to n_{LR}. The predictor selection procedure can be interpreted as that it sets some coefficients a_m to zero, which means that the term belonging to it is removed. This selection is done by stepwise selection and subsequently stepwise elimination of terms $a_m x_m$ if necessary[5]. During each step, all predictors are fitted individually to the training data and the one that reduces the squared error of predictions with the training data the most is selected for the LR-model, in addition to previously selected predictors. The selection stopping criterion is the decrease of the Akaike information criterion (AIC; see [Wilks, 2011]) or a maximum tolerable number of predictors, which is set to $n_{LR,max} = 6$. This $n_{LR,max}$ is set to reduce the possibility of overfitting, which will usually occur for large datasets if stopping is solely based on the AIC and additionally in the previous KOUW-system a criterion of maximum 5 predictors were used (see [Schmeits et al., 2005]). If one of the two stated criteria is fulfilled, the selection is stopped.

When a statistical model is overfitting, it fits a relation to residual data from a training dataset that are actually not predictable with the predictor selected to be added to the model. This would likely lead to degrading model performance on an independent verification set.

Models are fit for 1 to $n_{LR,max}$ predictors and results are compared, to select the appropriate LR model for each situation. The ideal number of predictors n_{LR} is found with the initial 9-fold cross-validation as described in Section 3.3.2.

It is important to realise that stepwise forward selection does not always lead to optimal combinations of predictors. To be sure that the best possible model will be fitted, all possible combinations of predictors will have to be tested and verified, but this is an infeasible job when the number of potential predictors is large, even for a computer. An interpretation of this selection procedure as non-optimal is the following: the

[5]An elimination step is attempted for all terms, if all of them are increasing the error criterion, AIC, none are eliminated

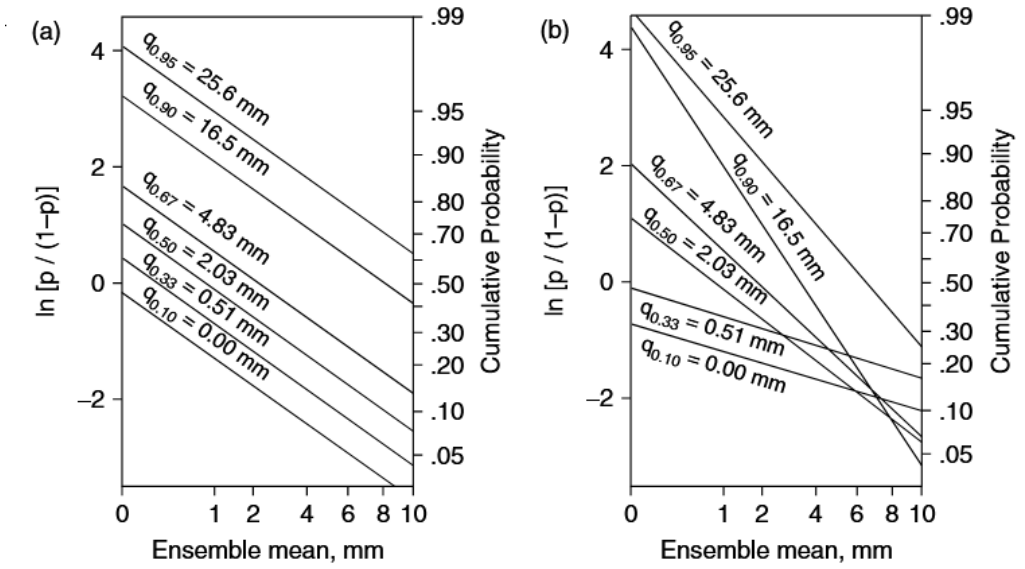


Figure 3: Probabilities of not exceeding a climatological accumulated precipitation quantile over five days as a function of ensemble mean precipitation over five days as predictor. Figure (a) shows a forecast plane made with an extended logistic regression model. Figure (b) shows the same forecast plane with separate ordinary logistic regression models per precipitation threshold. This figure was taken from [Wilks, 2009].

first predictor that is selected, is always the one with the highest correlation between model prediction and outcomes. As a single predictor, it will therefore perform the best of all predictors, given that the predictor is significantly better than the second best predictor. However, some combinations of two predictors without this best predictor might still give more information on the outcome (observations). In other words, those two predictors might be better in a complementary sense and therefore would improve the model. Part of this problem can be tackled with physical intuition, that is by trying to combine predictors that may provide complementary information into new potentially useful predictors. This is what is also done by testing newly invented predictors. Additionally, one can use more sophisticated and computationally costly algorithms for fitting the predictors, such as the so-called LASSO algorithm, but this is not applied. The interpretation of the function that LR fits, can be found in Figure 2. The probability gradient indicated on the graph[6] indicates a potential line along which the thunderstorm probability according to an LR model can increase/decrease with the S-shape of a logistic curve. Perpendicular to this, one can find the line along which the intersect of the graph leads to a 50% probability along a fit that is suggested in this example.

To construct the LR models, the function "StepAIC" from the R-library MASS is used [Venables and Ripley, 2002], with its default minimisation settings.

LR Models can be constructed for any threshold value of a predictand. But for an array of thresholds or full PDF, the extended LR and QRF methods are more appropriate, because the ordinary LR method can lead to inconsistent results as demonstrated by [Wilks, 2009]. These results are shown in Figure 3b: when the ensemble mean precipitation is about 8 mm, the probability of more than 16.5 mm precipitation is according to LR models larger than the probability of more than 0.0 mm precipitation, which is impossible in reality. It is caused by the variable coefficient in front of the ensemble mean precipitation; if coefficients are not exactly equal, the curves of high and low precipitation will always intersect somewhere. By using extended logistic regression, such an issue is circumvented in Figure 3a.

[6]Note that in this example logistic curve red isolines, indicating sampling density in Figure 2, are not taken into account to optimise the illustrated LR fit.

3.2.2 Extended logistic regression (ELR)

The general expression for an extended logistic regression model is Equation (2).

$$P(y = 1) = (1 + exp(g(\theta) + a_0 + a_1x_1 + a_2x_2 + a_3x_3 + ... + a_nx_n))^{-1} \quad (2)$$

All symbols have the same meaning as in Equation (1), but in addition, a function $g(\theta)$ provides a probability density distribution of conditional lightning intensity θ. Combined with the LR result the full lightning intensity distribution, including that for no lightning, is forecast. Hence, this threshold can be interpreted as a threshold intensity, but extrapolation beyond the lightning intensities in the training dataset is not very accurate [Wilks, 2009] and the model skill will not be stable in the upper tail of the distribution used for training. The aim is to skillfully forecast high lightning intensities, which means that the result of this ELR model should especially be reliable for the upper quantiles of observed lightning intensities. For the lower part of the distribution, it is not a direct intention to fit the best model. Usually, $g(\theta)$ is a power function. The shape of this power function is tested and improved applying ordered logistic regression, which is very similar to extended logistic regression, but for any threshold that is trained on, an intercept of the model that replaces $g(\theta)$ is optimized separately. An attempt for finding the relation between these intercepts is then done, like by [Messner et al., 2014a].

When applying the method, a certain set of thresholds for training (T_{ELR}) has to be chosen. A series of thresholds will likely improve the lightning intensity forecasting skill, such that the training profits better from the recorded lightning intensities. Since the intention is to forecast severe cases, the training intensities will be centered on the highest half of the lightning detection set, but training on many of the highest few percentiles as threshold would be based on few data points and may easily induce overfitting for the highest percentiles. Therefore, it is decided after some preliminary experiments to train in a systematic way on every 5^{th} percentile from 50^{th} to 95^{th} from lightning intensity climatology of the corresponding valid time. This climatology per valid time will be described in Chapter 4 and shown in Figure 11.

The variable selection procedure consists again of adding each predictor that is available to the current model with forward selection and selecting iteratively the best remaining potential predictor based on AIC, until a maximum of 4 NWP predictors is included in the model. This $n_{ELR,max} = 4$ is set because the smaller lightning intensity dataset is expected to be more susceptible to overfitting than the larger thunderstorm occurrence dataset and the previous KOUW-system had up to 5 predictors. No elimination is applied in this case. With two NWP variables, all combinations that are possible after elimination have been constructed, when using all options with one NWP variable. Hence, if at most two NWP predictors are used, elimination is not necessary. In some preliminary tests with Harmonie output from another model version and lightning detection dataset, it was seen that typically one or two NWP predictors would be selected based on initial cross-validations, such that elimination is likely unimportant.

To implement ELR, the function hxlr from R-package crch [Messner et al., 2014b] is used, with default minimisation settings.

3.2.3 Quantile regression forest (QRF)

In QRF, there is no dependence on linearity between predictors and response variable, in contrary to logistic regression due to the combination of factors a_ix_i. The tree method splits a (sub)set of training data in two subsets in each step of the building process. During the building process, each of the new subsets is iteratively split until the stopping condition is fulfilled. Such a split is based on whether a value of one of the predictors is exceeded. Before splitting, the algorithm checks all splits possible based on the available records and the available potential predictor variables. To illustrate this: if there are ten observations of five potential predictor variables selected options for the split, nine splits are possible in each predictor variable and in total, 45 potential splits have to be assessed.

To select an optimal split in a tree a statistical measure is minimised. When the tree does a classification task, such as yes or no thunderstorm, the associated Gini index of the outcome variable is calculated for all potential splits (Equation 3). For a regression task, the summed variance of the predictand, here

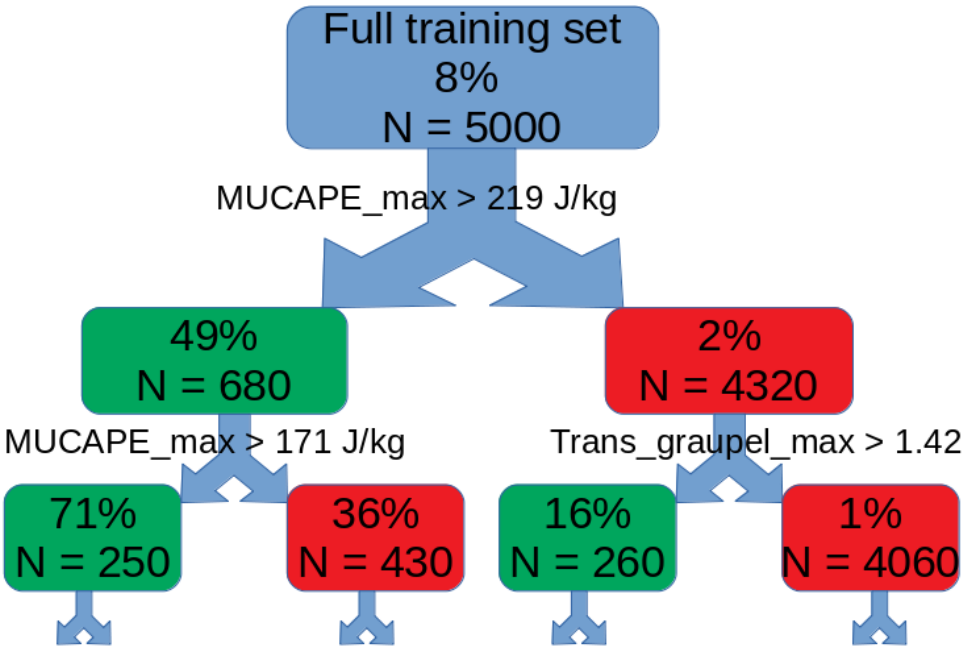


Figure 4: A fragment of an actual tree built in a quantile regression forest. The names at nodes indicate a predictor that is chosen in the tree, the values at nodes are the best splitting values and percentages in boxes indicate the empirical thunderstorm probability in the subset; the set that exceeds splitting value is green, with non-exceeding set coloured red. The number of cases (N) is also indicated in the tree. Note that the full training set is a random selection of the actual full dataset in QRF, which is different for each tree.

transformed lightning intensity, within a branch after the potential split is calculated and minimised to select an optimal split (Equation 4).

$$G(y) = 1 - (p_{y,no}^2 + p_{y,yes}^2)) \tag{3}$$

$$min(Var(y_1) + Var(y_2)) \tag{4}$$

Here, in the lightning intensity prediction, $y1$ refers to the set of lightning intensities on one side of the split and $y2$ refers to the outcome on the other side of the split. In case of classification, $p_{y,no}$ refers to the empirical probability of no thunderstorm and $p_{y,yes}$ refers to the empirical thunderstorm probability in a subset after splitting. The minimal sum of the Gini index (Expression 3) weighted by number of records in a subset defines the best split.

This process of optimal split selection and subsequently splitting is iterated until either the terminal subset has no variance and therefore is fully homogeneous or a certain size of the terminal subset of the dataset is reached. The minimum number of observations that a terminal subset must have before a split is attempted is settable and therefore it is the first so-called hyperparameter that is varied, namely "minimum node size" or s_{QRF}. Values of 3, 9 and 15 are used, because it has to be clearly smaller than the number of records in the datasets to provide representative selections of cases to compare with. The splitting procedure for one tree is further illustrated in Figure 4, with the predictor plane belonging to the same tree shown in Figure 5.

Random forest [Breiman, 2001] is an algorithm that builds many trees for regression or classification tasks. It selects a random subset of the training dataset to build a single tree. The random subsample of records is different for each tree in the forest. Random forest also randomly selects a subset of the potential predictors to try at each split (m_{QRF}). This is a hyperparameter that can be varied and regards the number

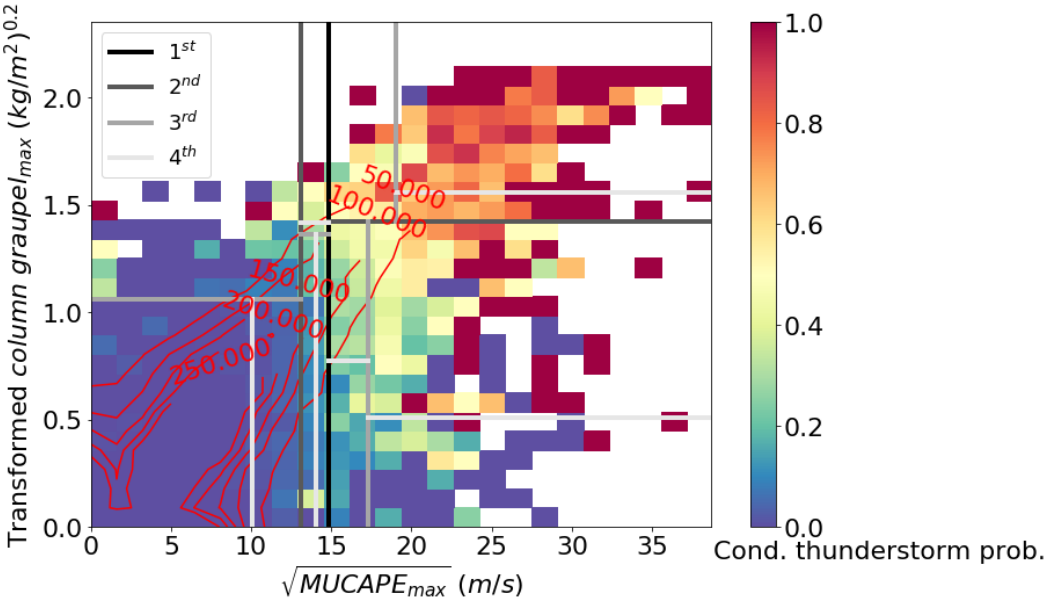


Figure 5: See Figure 2. Here it can be seen what happens when QRF works with these two predictors, with the grey values indicating splitting order. Note that some fourth splits are missing, because a third predictor was selected for these splits. This is the same tree as Figure 4.

of predictors tested only for taking that single splitting decision. In Figure 5 the hyperparameter is equal to two. Besides 2, 6 and 10 are tried, because this number has to be a fraction of all potential predictors that are eventually used for a forest. For a random forest, a specified number of trees is built, which is set to the default value of 500.

Evaluating the probability of an event with such a random forest is done as follows: a sample is sent down all trees by evaluating the decisions for the sample and via the nodes of each tree (see Figure 4). This leads to a certain terminal subset of comparable records for each tree after going down through vertically through tree. From the terminal subsets, random forest uses only the mean of the terminal subset selected of each tree, but QRF [Meinshausen, 2006] does in essence the same and uses information on the outcome of all training samples found at at terminal subsets of each trees. These subsets can give accurate information about the outcome distribution with the sample to be predicted. QRF builds an empirical CDF with all training samples from terminal nodes for regression tasks and similarly an empirical classification probability for classification tasks.

If a sample is in the terminal node of multiple trees, its weight in the empirical CDF thus increases proportionally with its frequency of appearance in terminal subsets. Therefore near neighbours in important predictor variables will end up in multiple terminal subsets, while the sample will likely have an outcome similar to these near neighbours. In essence, many multidimensional stepwise (stairslike) functions are built with a training dataset in QRF.

Stepwise backward elimination is applied for predictor selection, starting with all predictors of a potential predictor set. For the elimination steps, the concept of variable importance is used, which is a measure belonging to each predictor in a random forest that describes how important the variable was for making accurate predictions. Elimination in each step is based on the lowest importance value in the set of predictor variables, which for each variable is calculated by permutation of all records of a certain predictor (randomly reordering its values); it estimates the decrease in prediction accurracy of the part of the dataset that is not used for the tree for which the permutation importance is calculated. But this computation is still sensitive for perturbations and especially in very strongly correlated predictor sets; see discussion by [Gregorutti et al., 2016]. Due to the random nature of QRF, the early eliminations of

Table 1: Hyperparameters that are tested for each of the methods. In this table, n_{set} indicates the number of potential predictors that a potential predictor set contains (Table 2).

Hyperparameter description	**Name**	**Statistical Method**	**Options tested**
Number of predictors	n_{LR}	LR	1, 2, 3, 4, 5, 6
Number of predictors	n_{ELR}	ELR	1, 2, 3, 4
Training thresholds	T_{ELR}	ELR	fixed, based on preliminary tests: $q_{0.50}$, $q_{0.55}$, $q_{0.60}$, $\cdots$ $q_{0.95}$
Number of predictors	n_{QRF}	QRF	1, 2, 3, 4, ... $(n_{set}-1)$, n_{set}
Number of predictors tried for each split	m_{QRF}	QRF	2, 6, 10
Min. sample size necessary to split	s_{QRF}	QRF	3, 9, 15

variables may be somewhat random, since the differences in importance between predictors will be smaller and correlations between predictors have higher impacts when many well correlated variables are present, whereas the last eliminations are likely to have a certain order that is not so dependent on random subsetting for each fit. After elimination of all of the variables but two, the best number of predictors could be derived from the scores of the models in the initial 9-fold cross-validation, in combination with the best performing hyperparameter set. Here, a model refers to all decision trees in a QRF together. The found settings can be used for final cross-validation models.

For the implementation of QRF, the R-package ranger [Wright and Ziegler, 2017] is used.

3.3 Cross-validations and hyperparameter determination

3.3.1 Hyperparameters

For each statistical method there are one or several statistical hyperparameters that can be set. The most obvious one is the number of predictors used in the (E)LR models and other hyperparameters may describe characteristics of the subsample set that is selected and used when fitting, or the way that a model is fitted with QRF; see Section 3.2.3. Therefore, they can be interpreted as meta-parameters for the fitting process. To determine which model settings of hyperparameters to use, the section will discuss the initial cross-validation procedure (Section 3.3.2). Based on the verification performance (Section 3.4.1) of the initial cross-validation the best model settings are chosen and used for a final cross-validation and the actual post-processing model. An overview of the hyperparameters is given in Table 1. Note that the hyperparameter with training thresholds has been set with preliminary experiments.

3.3.2 Initial cross-validation strategy

The initial cross-validation applied to choose hyperparameters is a 9-fold cross-validation. First, all datasets are split into three years and one year is kept apart for final testing, leading to three initial sets with NWP model data to train on. Then, within these three initial sets, a random selection of dates (days) is applied. Each date within a dataset covering two years is coupled to one of the three test sets for model verification and then the other two test sets within these two years of data are combined and used for training a model. Finally, the model is verified with a test set. So 9 verifications with 9 test sets of 9 models for each model setting have been done. This is the initial cross-validation procedure.

The grouping by dates is to account better for dependence: if two neighbouring grid cells end up in different test sets, the model performs likely better due to spatial coherence in some predictor fields and spatial coherence predictand outcome. On the other hand, the random drawing of dates is to take into account variability in the dataset, such as the structured seasonal lightning (intensity) variability.

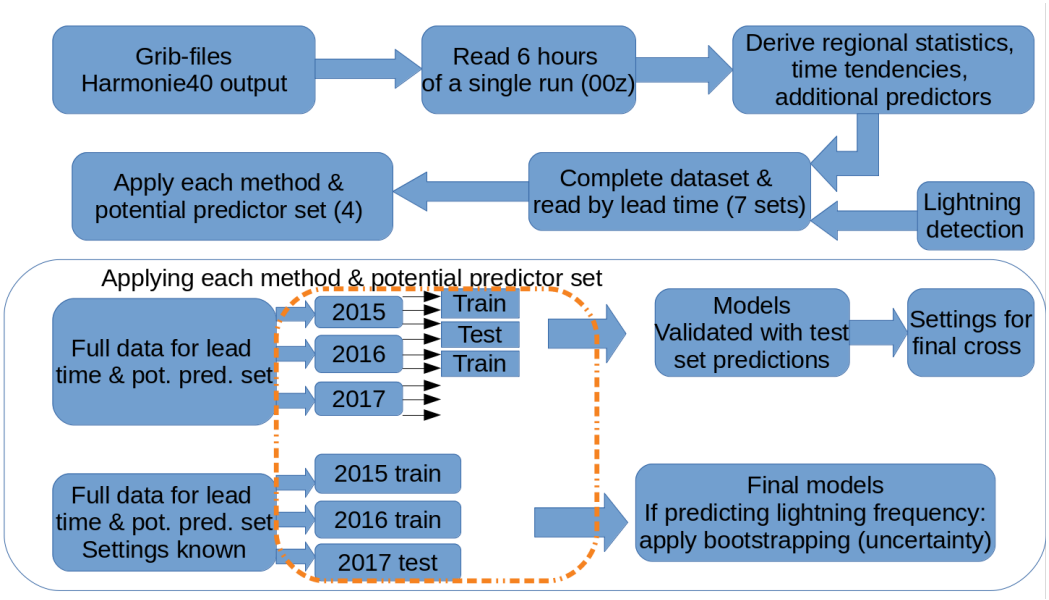


Figure 6: Data flow through the R-scripts. The orange dashed part of the figure indicates where cross-validation is applied by interchanging the test set for each cycle of cross-validation. Note: when applying the methods, the settings are the hyperparameter settings, which is for (E)LR only the number of predictors and for QRF some additional settings, to be explained in Section 3.2.3.

3.3.3 Final cross-validation strategy

For the final cross-validation the statistical models are trained using the best hyperparameters as shown in Table 1. Each year in the final cross-validation is used once used for testing each of the post-processing models, while the other two years are then used for training. For the lightning intensity forecasts, the second approach is used as well: randomly distributing the three years in three subsets and training on two of these subsets, with afterward a verification based on the other of the subsets. One of the two approaches will be preferred and therefore presented as main result (Section 5.4.1). In the cross-validation, the verification set and training sets are again rotated. Like in the initial cross-validation strategy, all records of different regions for one day always ends up in the same verification set, to prevent interference with spatial homogeneity in predictor and outcome (see 3.3.2).

The data flow through the R-scripts until the final cross-validation is visualised in Figure 6.

3.3.4 Impact of cross-validation procedure on predictor selection

The cross-validation procedure leads to a variable set of predictors to be selected, which becomes even less stable when higher correlations among the predictors are present in the potential predictor dataset and the number of observations becomes smaller. However, high correlations also imply that one potential predictor can be replaced by another one without large consequences for prediction. In the results on final fits, importance measures from QRF and forward selected frequency of predictors for (E)LR will be exploit, keeping in mind the correlations if necessary. Additionally, any selected predictor is verified with an independent set in the cross-validation procedure for all methods.

3.4 Verification

3.4.1 Verification methods

Forecasts can be compared to each other using many verification measures; for probabilistic forecasts, the Brier score is a natural one. It is computed as follows: a vector with probability forecasts that were issued is compared a vector to the eventual outcomes of all the events, which is for individual events either 0

(non-event) or 1 (event). The squared difference between all elements is averaged to get the Brier score. Therefore, the Brier score is a mean squared error of probabilistic forecasts [Wilks, 2011]; see Equation (5).

$$BS = n^{-1} \sum_{k=1}^{n} (o_k - p_k)^2 \tag{5}$$

Here, index k denotes an element in the vector with observations (o)/probabilities (p) and n denotes the total number of observations and forecasts in the record. Brier score is a verification measure that can account for resolution, reliability and uncertainty of forecasts: with a high reliability, the forecast probability always resembles the conditional frequency of thunderstorm event observation. In other words, a forecast issued at 10% probability leads to an event that indeed happens approximately 10% of the cases in which it is issued and similarly for other probabilities. On the other hand, resolution expresses the ability of a forecasting model to discriminate between cases and non-cases: forecasts with high resolution issue probabilities that are frequently close to 0% when no thunderstorm observed or 100% whenever a thunderstorm is observed. A forecast with no resolution is the climatological forecast: it issues a forecast with the same probability for every sample. Therefore, in forecasts with high resolution, forecast probabilities are shifted away from the climatological probabilities. Finally, the uncertainty in the Brier score expresses to what extent one can anticipate on the outcome: with a 50% probability according to a climatological forecast, the outcome is uniformly distributed between 0 and 1, which means highly uncertain. This will lead to the maximum uncertainty and a maximum in climatological Brier Score of 0.25: both occurrences and non-occurrences always lead to the same squared error of 0.25. For high or low climatological probabilities, the outcome is relatively certain and this leads to a lower minimum possible Brier score than 0.25.

The Brier score of a model is typically compared to the Brier score of a reference forecast and in the normalised comparison with a reference model it is called Brier skill score (BSS). The climatological forecast is usually used as reference forecast: this type of forecast can be easily issued with historical records of some event. By calculating a BSS, one can see the improvement of a forecasting method compared to climatology as method: the relative decrease mean squared error (Brier score) of a model compared to the mean squared error (Brier score) of the climatological forecast is expressed as the skill score (Equation (6)).

$$BSS = \frac{BS_{ref} - BS_{model}}{BS_{ref}} = 1 - \frac{BS_{model}}{BS_{ref}} \tag{6}$$

Here, it is defined that $BS_{ref} = BS_{clim}$. Sometimes another method such as persistance (forecasting the outcome using the observations of the previous day(s)) is used as reference. The Brier score and BSS can be computed for any threshold in the distribution of observations.

When integrating the Brier score over all possible thresholds with equal weighting, one gets the (unweighted) Continuous Ranked Probability Score (CRPS). Another way of understanding this scoring rule is that for each combination of an observation with predictor values in the dataset there is a cumulative density function (CDF) that describes the probability that a predictand threshold would not be exceeded based on the forecasting model, as a function of predictand threshold: $F(x_i)$. The outcome CDF of an observation, namely $I(y_i)$, is defined by a similar function that jumps from 0 to 1 at the value of observation y_i (see equation 7, with $w(x_i) \equiv 1$). Integrating the area between both CDF curves gives the contribution of this sample to the CRPS, as shown in Figure 7. When averaging the CRPS over all samples, we have a measure of model performance.

Weighting depending on forecasting threshold can also be applied [Thorarinsdottir and Schuhen, 2018].

$$CRPS_{mean} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \int_0^{\infty} w(x_i)(F(x_i) - I(y_i \leq x_i))^2 dx_i \right\} \tag{7}$$

In the weight function $w(x_i) = w(x)$ is independent of the sample, but depends on transformed lightning intensity. It is set to 0.1 for intensities equal to or below 25 discharges per 5 minutes and 1.0 for higher

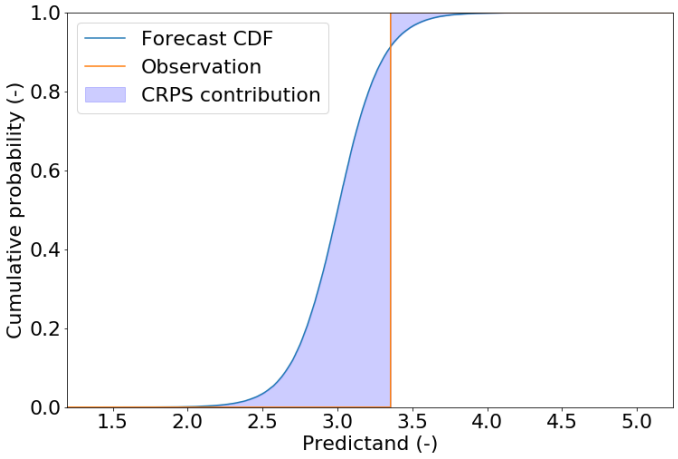


Figure 7: Example of unweighted CRPS contribution of an individual sample (shaded area), given the cumulative forecast distribution and observation (orange jump).

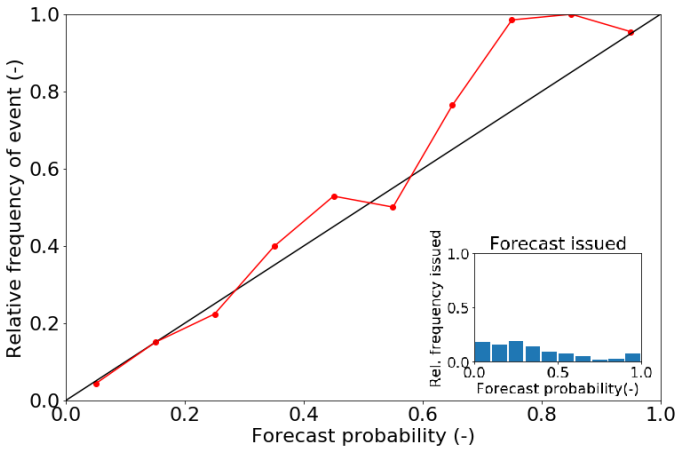


Figure 8: Example of a reliability plot. On the left hand side, the relative frequency of an event in the verification set versus predicted binned probabilities by a model are given. On the right hand side the relative frequency that a model forecasts a probability in the same bins is shown.

intensities, which are of specific interest. In the unweighted CRPS, $w(x) \equiv 1$. Both the weighted and unweighted CRPS will be calculated.

The score as defined in Equation 7 can only be calculated exactly when the full probability distribution of a continuous predictand is defined by a forecast. This means that the score is only useful for predicting lightning intensity. However, verifying the full probability distribution is impossible in practise when applying it to a QRF fit, because we have a finite number of samples in the train set: it returns an empirical CDF, which implies that the PDF is not fully continuous. On the other hand, to compare ELR and QRF in a fair way, the scoring rule has to be implemented the same way in verifying both methods. Therefore the CDF is approximated numerically by calculating some equidistant quantiles of the outcome to in the end derive the CRPS. This approach is also frequently used for verifying ensemble forecasts. The approach is equivalent to approaching an integral with a Riemann sum. In the end, by increasing the number of equidistant quantiles, the integral is computed more accurately, but its calculation requires more computation time if the number of quantiles is increased. Eventually 25 quantiles are used to evaluate the ensemble CRPS.

Similarly to the Brier score, the CRPS can be compared to a reference CRPS obtained using climatology

forecasts, resulting in the CRPSS. Both its weighted and unweighted version can be evaluated. In Equation (6) BS is then substituted with CRPS, as in [Wilks, 2011].

The verification information can be represented in a convenient way, using reliability diagrams. In this type of diagram, the issued probabilities are distributed in bins. For each of the probability bins, the conditional relative frequency of the positive observations is plotted against this binned probability. The forecasts are reliable if a line that connects the points in the reliability diagram lies close to the 1:1-line. However, in some cases, probability bins of the forecast may match with only few observations and the points are then more likely to deviate strongly from the 1:1-line. Therefore, the relative frequency that a certain probability range connected to one of the probability bins is issued, is represented in another part of the same diagram. So with few observations in some probability bins, the reliability diagram can look very unstable, but it may not affect the BSS if it regards a few samples. This happens with a probabilistic forecast for 60-90% in the example of Figure 8. Additionally, this relative frequency diagram can be used to get insight in the resolution of forecasts. If the climatological probabilities correspond to one of the bins in the centre of the diagram, a high resolution model corresponds to a model with high relative frequencies of forecasts issued in the highest and lowest bin.

For verification we use the R-package Verification [Laboratory, 2015]; for the CRPS packages SpecsVerification [Siegert, 2017] and scoringRules (weighted CRPS) [Jordan et al., 2018] are used.

3.4.2 Block bootstrapping to assess the uncertainty in the BSS

The verification scores by region can be used for thunderstorm occurrence forecasts to assess uncertainty in the BSS to some degree. However, as neighbouring regions will be strongly correlated, the uncertainty interval in BSS will be somewhat stronger than when using independent samples for uncertainty. Usually, independence of samples is appropriately assumed in significance tests. For the lightning intensity forecasts, an alternative method that better quantifies uncertainty of the models is applied: block bootstrapping. The blocks consists of observations from a day, as within one day the occurrence of severe thunderstorms among different regions will be strongly connected. This is because they usually extend over multiple regions on one day. You can say that some dependence is still likely to be present among two consecutive days, but this is not so high. Therefore grouping by days is applied and block bootstrapping with dates is done. For block bootstrapping, the procedure applied is to generate a series of random discrete numbers from one to the number of dates in the dataset that is verified, where each date is connected with a random number. The random numbers are converted into dates and these dates are put in the verification dataset. Each time, a few of the dates will occur multiple times in the verification set and each time some will be left out. By applying the procedure 1000 times, one can get an idea of the distribution of the Brier skill scores. From the 1000 empirical skill score results, the 0.025 and 0.975 quantiles are eventually extracted as bounds for the 95% confidence interval of the BSS.

3.5 Potential predictor sets

3.5.1 Overview

This section is to some extent related to thunderstorm theory, but also uses the statistical methods discussed in this chapter and therefore complements content of both Chapter 2 and this chapter. In addition to the "elementary" set of 15 potential predictors (see Section 2.2.6), we start with another very large dataset (Section 2.4.1), from which the eventual selection of potential predictors for the potential predictor sets is described in this section. The potential predictor sets contain predictors that can be used by (E)LR and QRF to build statistical models. By comparing models built with different other potential predictor sets, an idea is given of how complementary various types of physical predictors can be to other predictors, even if they describe the same physical ingredient of convection (Section 2.2.2). All the potential predictor sets used are described in the remainder of this section and summarised in Table 2.

3.5.2 Preliminary potential predictor selection

It could be expected that frequently high or low quantiles within regions are more useful than extremes or mean values of NWP output variables, especially because convection is a skewed phenomenon. Adding extra quantiles to the dataset compared to minima and maxima that were previously used by [Whan and Schmeits, 2018] could improve the statistical models as well, because only extreme values of predictors may give less opportunity to extract complementary information than some slightly less extreme quantiles such as $q_{0.90}$ values within a region. In addition, extreme values of different predictors are unlikely to overlap in space and can be localised outliers. Note that there are about 1200 Harmonie grid cells with six hourly predictor values that are resampled to quantiles $q_{0.98}, q_{0.90}, q_{0.10}, q_{0.02}$, median, mean and extreme values within regions, using nearest neighbour as interpolation method. For some variables, differences between quantiles or extremes may also be interesting.

Since the previous set of quantiles, means, and extremes leads to a potential predictor set of around 600 variables, this set was first reduced by removing very strongly correlated variables. An example is that mean sea level pressure regional mean is very highly correlated with its 6 hour regional extremes, as KOUW-regions are up to almost 100 km in size and the pressure field is typically very smooth. This first step is manually done and has reduced the initial set to approximately 375 predictors. The next selection step is done by making some quantile regression forests per lead time, using its predictor elimination order and permutation importance measure. These are compared to find out which statistical measures of a physical variable are preferred over other statistical measures of the same physical variable. This step is to some extent arbitrary, like the previous elimination step. One could argue about a small bias towards favouring QRF predictions applying this method; however, on the other hand it is very obvious that (extended) logistic regression is much less able to profit from any predictor than QRF. This is because the method uses only few predictors in each model to not overfit. On the other hand, QRF is found to work well with many predictors and is not susceptible to overfitting due to its random subselection strategy, as shown by results of [Gregorutti et al., 2016]. Eventually, a 228 potential predictor set is left. Some physical fields are still present several times in this set. A set of only maximum and/or minimum and means (whichever is physically relevant) of the variables was also made, with the exception of most unstable CAPE and level of neutral buoyancy. These predictors additionally had $q_{0.90}$ and $q_{0.98}$, because they are assumed to be very important. This leads to a much smaller potential predictor set of 91 potential predictors, of which the details are found in Appendix A. Due to a mistake only 90 of the 91 potential predictors were used in the initial cross-validation; see Sections 5.1.3 and 5.3. This is the first full predictor set for final cross-validation.

3.5.3 Smallest set

From the "elementary" set of 15 potential predictors (Section 2.2.6 and Appendix A), an even smaller selection of 4 potential predictors was made by applying the ingredient based approached to severe convection; as best integrated measures of instability MUCAPE is combined with moisture, namely precipitable water. Furthermore a proxy for forcing is added, namely sea level pressure and graupel, representing the actual occurrence and intensity of convection in the NWP. The six hour maximum within the region is used for this four potential predictor set.

3.5.4 Potential predictor sets for physical experiments

A composite set with physically very relevant "elementary" variables for severe convection as well as the most important additional covariate predictors found from selections of the potential predictor set of 91 variables was subsequently made, which contains 40 potential predictors. These potential predictors are shown in Appendix A.

Based on the set of 40 potential predictors, three additional experiments are performed where the additional forecast skill that is gained by including particular sets of predictors is investigated. We remove:

1. The predictors containing a vertical integral of buoyancy (8)
2. The predictors containing a microphysical reservoir (4)

Table 2: Name and description of potential predictor sets that are used.

Name of set	**Size**	**Purpose/remarks**
"228"	228	Presence of multiple quantiles for most predictors
"91"	91	Maxima, minima and/or mean of potential predictors only (except MUCAPE and LNB)
"40"	40	Selection with elementary potential predictors set + QRF from "91"
"15"	15	Elementary potential predictors set
"4"	4	Three ingredients + actual showers
no_CAPE	32	Value of the predictors containing CAPE and/or CIN
no_mph	36	Value of microphysics predictors: snow, graupel, MUCAPE with snow/graupel
no_PWinst	34	Value of newly combined precipitable water - instability predictors

3. The six new predictors with precipitable water and instability index (6) (see Section 2.4.2)

4 Conditional thunderstorm climatology

This short chapter serves as a cornerstone to interpret models that are eventually build and presented in subsequent chapters and to understand the forecast skill of these models. As discussed by [Ahijevych et al., 2016], having strongly unbalanced datasets with only a few per mille of positive outcomes does not lead to optimal statistical models. As a rule of thumb, approximately 5 to 10% positive occurrence in our dataset should be kept in mind to get close to such optimum, although the more balanced outcomes are, the closer one would likely approach an optimal model. This means for predicting lightning intensity, one could say that forecasts are unlikely to work well above the 95^{th} percentiles of lightning intensity climatology (Section 4.3), unless a very powerful transformation has been applied such that ELR allows for extrapolation and unless QRF has some surprising results [Whan and Schmeits, 2018]. Only 19-35 samples exceed that quantile, depending on the time of the day. The dataset will be interpreted in this framework in this chapter. The climatology is very limited in time span and therefore not necessarily representative on long term, as the dataset covers only three years and it appears to be inhomogeneous. Moreover, it covers only the time of year that is studied: April 15^{th} to October 15^{th}. The climatology serves as a reference forecast, upon which improvement can be shown using the skill scores (BSS and CRPSS).

Additionally, the dataset with predictors and thunderstorm occurrence predictand will be exploited to illustrate an example of how predictable lightning is with the predictors used, which can also serve as an empirical guidance for those who are interested (Section 4.2). Furthermore, the issue of homogeneity of KLDN lightning detections is discussed in Section 4.4. With this chapter, the point where a lot of care is required for interpretations of the results is automatically passed.

4.1 The climatology of thunderstorm occurrence

In Figure 9, it can be seen that thunderstorm frequencies over all KOUW-regions average 6% during night and morning (21-09 UTC) and 10% during daytime (09-21 UTC), which makes an average of 8%. During the daytime, a clear north-south gradient in climatology is seen, as the low region indices 1-3 indicate the northern regions and the high indices 10-12 the southern regions. Southeastern and central regions 3, 5, 6 and 8 to 12 all have a climatology with a clear diurnal cycle, whereas the other (they are northwestern) regions have no clear dependence on a diurnal cycle. These four northwestern regions have the lowest thunderstorm frequency during daytime (09-21 UTC). This north-south gradient over The Netherlands is consistent with results based on synoptic stations and lightning detection from [Taszarek et al., 2019] and previous KOUW-study [Schmeits et al., 2005].

For training purposes, it can be said that putting all regions together not only increases the number of cases with thunderstorms; it also means that predictand outcome in the dataset is less unbalanced: there are 6 to 10 % of positive cases for each valid time. With training on individual regions, this would drop to 3-4% in regions.

Note that some thunderstorms in 2015 in region 6 are wrongly processed (see also Section 3.1), which leads to an underestimate of the lightning frequency of 0.01: instead of 8.9% of the time it should be 9.9% of the time in this region. All four valid times were affected by this error.

4.2 Thunderstorm occurrence climatology conditional on Harmonie predictors

Besides a climatology conditional on subregions and time of the day, as described in Section 4.1, another condition can be the air mass which flows over The Netherlands. It can be said that thunderstorms are more likely in tropical air mass than in polar air mass for example (see Figure 10), or that the direction from which the flow is coming can give information on relative frequency of thunderstorm occurrence. Additionally, one can describe the climatology or change in time of air mass that enters the study region or the change of wind direction in time. The combination of the former two types of information can lead to a hypothesis whether frequency of thunderstorm occurrence increases or decreases in time. Nonetheless,

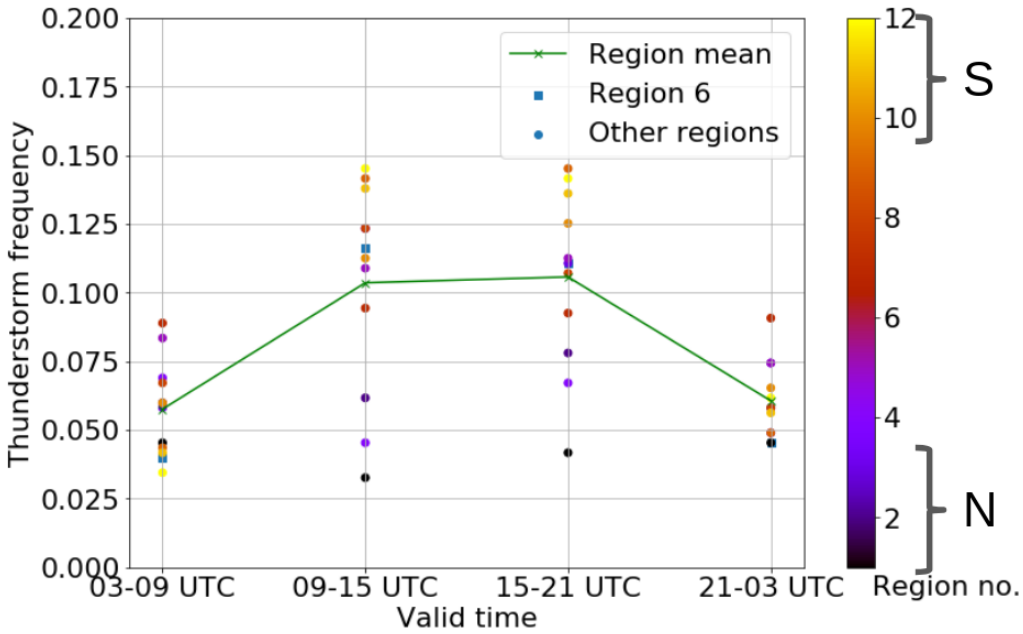


Figure 9: Thunderstorm frequency observed in each region as function of valid time, with the KOUW-region index indicated by the colour. Note that region 6 is affected by a data processing issue (see main text). Region 6 is shown as blue square, in contrary to indications in the legend and other regions, because it has some wrongly processed lightning detections. The northern ("N") and southern ("S") regions are also marked as such.

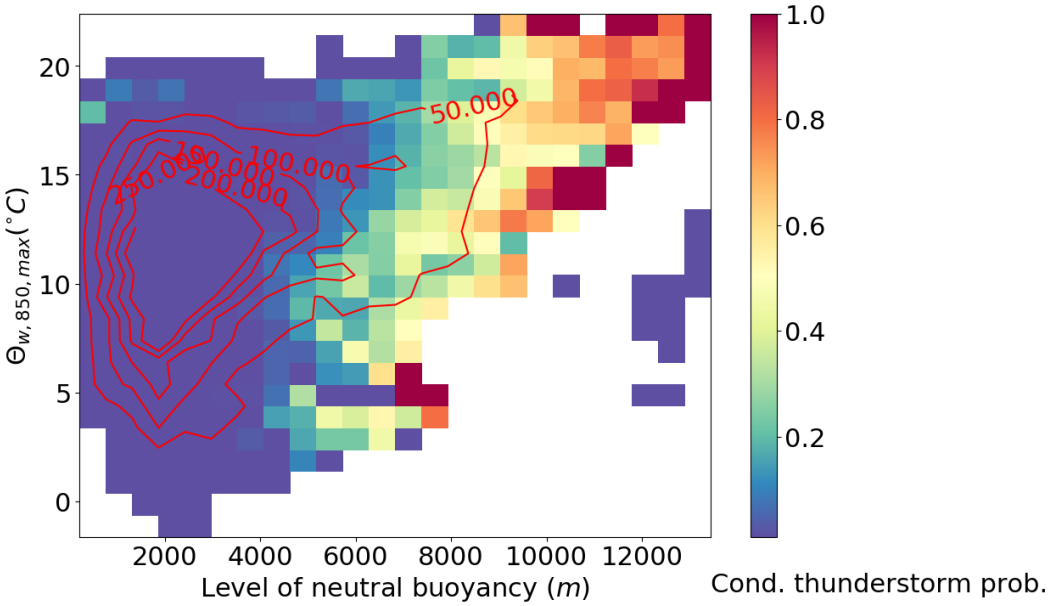


Figure 10: Empirical conditional thunderstorm probability as a function of maximum level of neutral buoyancy or equilibrium level and maximum potential wet bulb temperature at 850 hPa, based on 03-27 hour reforecasts of Harmonie. The red lines indicate that at least 50, 100, 150, 200 and 250 samples are present in a grid cell for empirical probability estimate.

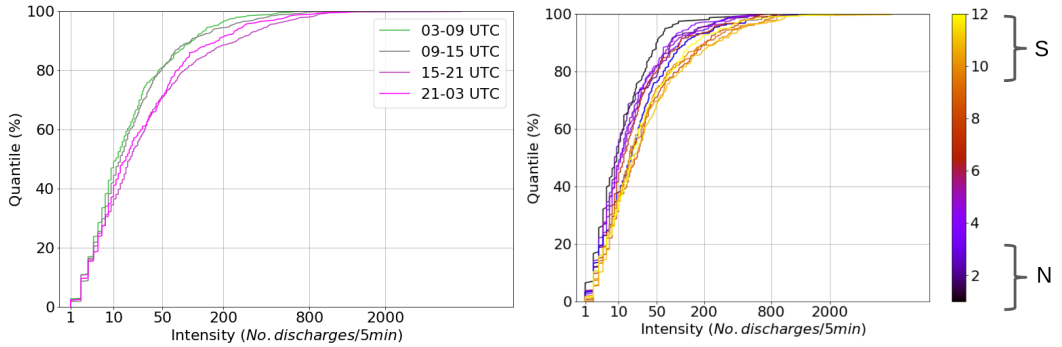
non-stationarity of thunderstorm occurrence frequency conditional on the flow over The Netherlands might mean that such an hypothesis turns out to be falsifiable.

Simultaneously, the short-term forecasting of thunderstorms might profit from the same type of conditional information based on climatology of thunderstorm occurrence. Besides some measures for thermodynamic instability, information such as air mass and flow may be beneficial. This type of information is also informative for understanding thunderstorm forecasting from the meteorological point of view and in statistical post-processing models applied in practice. So ideally, in practice a prediction system based on a dataset with many variables would be used to optimise thunderstorm forecasts and one or two variables would be used simultaneously to build on some physical understanding. Different combinations of two variables will lead to different insights and the real predictions could be based on many combinations simultaneously. Quantile regression forests are very useful for predictions based on potentially many variables and for the interpretation of the results it can be very useful to see why forecasts can profit from the combination of and relation between physical predictors. Additionally, the combinations of variables can help to find out why a model skill in practice behaves in a certain way, as function of the predictors that are present in the model.

Figure 10 shows the empirical thunderstorm probability as function of air mass ($\Theta_{w,850,max}$) and maximum LNB. Isolines indicate the frequency that conditions occur in Harmonie reforecasts. One thing that can immediately be seen is that thunderstorm probability behaves differently as a function of maximum LNB in cold air masses than in warm air masses. That is, thunderstorms typically start occurring in cold air masses (low $\Theta_{w,850,max}$ in Figure 10) with a maximum LNB of about 4500 m, whereas this is slightly over 6000 m in warm air masses (high $\Theta_{w,850,max}$). It can also be seen from the red isolines in aforementioned figure that the relative frequency that potentially thundery conditions occur, is higher in warm air masses ($\Theta_{w,850,max} > 9°C$). A function that is solely based on LNB_{max} would smoothen this effect and the dependency on air mass would be hidden. Probabilities would start to increase slowly when LNB_{max} reaches about 4500 m and keep gradually increasing until an LNB_{max} of roughly 10.000 m. This means that many cases are classified as unlikely thunderstorm cases correctly; this is because LNB_{max} is usually below about 4500 m. However, in the few cases with higher LNB_{max}, the second variable $\Theta_{w,850,max}$ helps improving thunderstorm forecasts a bit my modifying issued probabilities for high LNB_{max}, since it shows whether a cold or warm air mass is present or not. Therefore it can be better inferred what the thunderstorm occurrence probabilities are. The modification would improve overall scoring slightly, since these are a only few samples within the full LNB_{max} distribution, as Figure 10 shows. The post-processing model could improve in particular when $LNB_{max} > 4500$ m, which are cases where the forecast is particularly relevant. Benefitting from many variables is a task that would fit QRF, since LR only takes a few variables into account in each model without suffering from overfitting. Additionally, LR might not be able to pick the ultimate combination of complementary predictors with stepwise selection or only in a fraction of the cross-validations. Although another variable might lead to better discrimination between thundery and non-thundery cases than maximum LNB, this example illustrates how forecasts can profit from many variables in a probabilistic thunderstorm forecasting model.

Besides this, Figure 10 is in agreement with what would be expected based on [Takahashi, 1978]: his conclusion was that thunderstorms are likely to occur when convective clouds stretch out over the temperature range of -10 °C to -20 °C. If we assume that the Dutch summer climate does not likely lead to cloud bases with temperature below -10°C, we can see that as it is anticipated in Section 2.2.1, the combination of LNB_{max} and $\Theta_{w,850,max}$ may give a reasonable estimate whether this criterion is reached. That is, when the potential wet bulb temperature increases, the convective cloud needs to reach higher to generate similar thunderstorm probabilities in the figure. However, the rise of thunderstorm probability is not checked to coincide with level of neutral buoyancy of approximately -20 °C and due to small errors in model forecasts (subtle inversions) or as consequence of assumptions for parcel calculations (Section 2.3) leading to non-optimal estimates of LNB_{max}, the relation between this criterion and arising empirical thunderstorm probabilities will not be one to one.

Note that with a very high LNB_{max}, for example 13.000 m as at the edge of Figure 10, the probability can be equal to 0.0 or 1.0 due to presence of just one or two samples in a grid cell. Additionally, the



(a) Coloured by time of the day

(b) Coloured by KOUW-region index

Figure 11: Cumulative distribution of lightning intensities conditional on at least two discharges being detected. Note that the x-axis is transformed at power $\frac{1}{4}$. The colour of region 6 deviates from the legend in figure b, because it has wrongly processed detections and the northern ("N") and southern ("S") regions are marked as such in figure b.

zero probabilities for $\Theta_{w,850,max} < 12°$ C and $LNB_{max} \approx 12.000$ m are likely related to an error in the LNB_{max} calculation (see Section 2.3). In general the Harmonie output looks very useful for producing thunderstorm forecasts, based on Figure 10.

4.3 Conditional intensity climatology

The conditional climatology of lightning intensities is shown in Figure 11a. The late afternoon and evening hours (15-21 UTC) clearly have a distribution with higher intensities. For the nighttime (21-03 UTC), this generally holds as well, but above $q_{0.80}$ its quantile curve in Figure 11a shifts toward the 03-09 and 09-15 UTC valid times. Between 03 and 15 UTC the lightning intensities are lower, which is consistent with expectations based on previous results (Figure 2b in [Schmeits et al., 2005]).

It is also important to note that there are only 2200 thunderstorm cases in the dataset (see Table 3 in Section 4.4): therefore the upper part of the distribution is wobbly when separating them by region. Only 183 thunderstorm cases per region are left on average, which means that the distribution of lightning intensity appears somewhat unstable and poorly sampled from about $q_{0.85}$ onward. For the four valid times, a somewhat unstable appearance happens between $q_{0.90}$ and $q_{0.95}$.

Looking at the regional intensities (Figure 11b), the N-S/NW-SE gradient in lightning intensities is clear. Region 1 (the most northwestern region) has very infrequent high lightning intensities. It should be kept in mind that this region also has lowest probability of occurrence in general (Figure 9). This means that the conditional $q_{0.97}$ of about 100 discharges per 5 minutes happens in that region only about 0.1 % of the time. This is in three six hour periods out of about 2200 six hour periods in total. In contrary, in some of the southern (yellowish and orange) regions in Figure 11b intensities above 100 discharges per 5 minutes happen 1.5 to 1.9 % of the time (slightly below 20 % of 11 or 12% of the time). The other northern regions 2 to 6 and in addition region 7 are in between the aforementioned extremes. Note that region 6 is again affected by an issue with data processing (see Sections 4.1 and 3.1), such that its climatology differs slightly from the true detection climatology.

4.4 Inhomogeneity of KLDN lightning intensity

For understanding and interpreting the climatology of our dataset, it is necessary to assess how well the dataset with detections represents the truth of lightning. The climatology described in this chapter is strongly affected by detection efficiency changes in time. Documentation provided by Météorage (the provider of KLDN detections to KNMI; personal communication - Stéphane Pedeboy) points at homogeneity issues in the detection dataset. This documentation demonstrates that the detection efficiency in

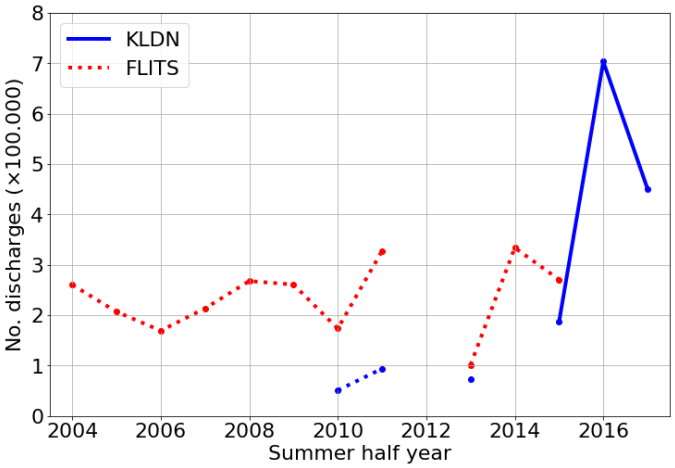


Figure 12: Total number of detections per summer half year as far as available and processed, over all KOUW-regions with FLITS and KLDN. The main source of detection data is drawn as continuous line and other detections are drawn as dashed lines.

the southern and western regions should have increased in the autumn of 2015, due to the implementation of extra sensors in the UK. Among these sensors is one in Shoeburyness, east of London. Additionally, the same happened over the northern regions of The Netherlands in June 2017: extra sensors were installed in De Kooy (near Den Helder) and Eelde (at Groningen Airport).

The consequences for the detection dataset are remarkable: in 2015, there are 186.627 detections in the summer half year (and around 10.000 in region 6 have been processed wrongly!), whereas in 2016 this number increases to 703.368. In 2017, this number is 450.740 over the same period.

Unfortunately, there is no detection system that is evidently homogeneous, such that even trying to make any potentially robust correction for aforementioned inhomogeneities is hardly possible. The reason for this is amongst others that detection systems have differential sensitivity for cloud to ground and intracloud lighting, due to the frequency range over which they detect and differential sensitivity as function of the distance at which a discharge occurs. Even when using a homogeneous detection set for corrections, single events will not be reflected properly in any potential correction of the detection dataset, as can be concluded from [De Vos, 2015]. Lastly, the large interannual variability in lightning activity does not help for the potential of corrections.

However, the KNMI FLITS detections are still available until February 2016. This means that the dataset of full half-year detections covering 2015 can be used. It contains a total of 270.793 discharges, of which 256.117 have been detected that overlap in region and time stamp with KLDN. This number of 270.793 is clearly a positive anomaly compared to years available in FLITS detections, 2004-2015 (Figure 12). These years had a persistent operational sensor set, except some minor disruptions (more than 99.5% complete), which makes the detection set much closer to homogeneous in time. The number of discharges in summer half years within the FLITS dataset are 101.380-333.881, averaging at 215.920. These numbers illustrate that the 2016 and 2017 numbers of discharges in the KLDN dataset absolutely don't fit in the FLITS climatology and that FLITS has even more detections in 2015 than KLDN (when assuming a stationairy climate). In other words: it looks like the 2015 detection dataset of FLITS would be likely underestimating the detections that would have been made by KLDN if 2015 would be homogeneous with 2016. KLDN has even fewer detections than FLITS in 2015, so KLDN likely underdetects in that year.

Therefore, two experiments with adjusted numbers of detections in 2015 are conducted, to investigate potential dependence of the forecast skill on inhomogeneity in KLDN detections. The first one is by replacing KLDN lightning intensities by FLITS lightning intensities in 2015. The second is by simply doubling KLDN detections, which is very crude and arbitrary. However, there are very strong indications

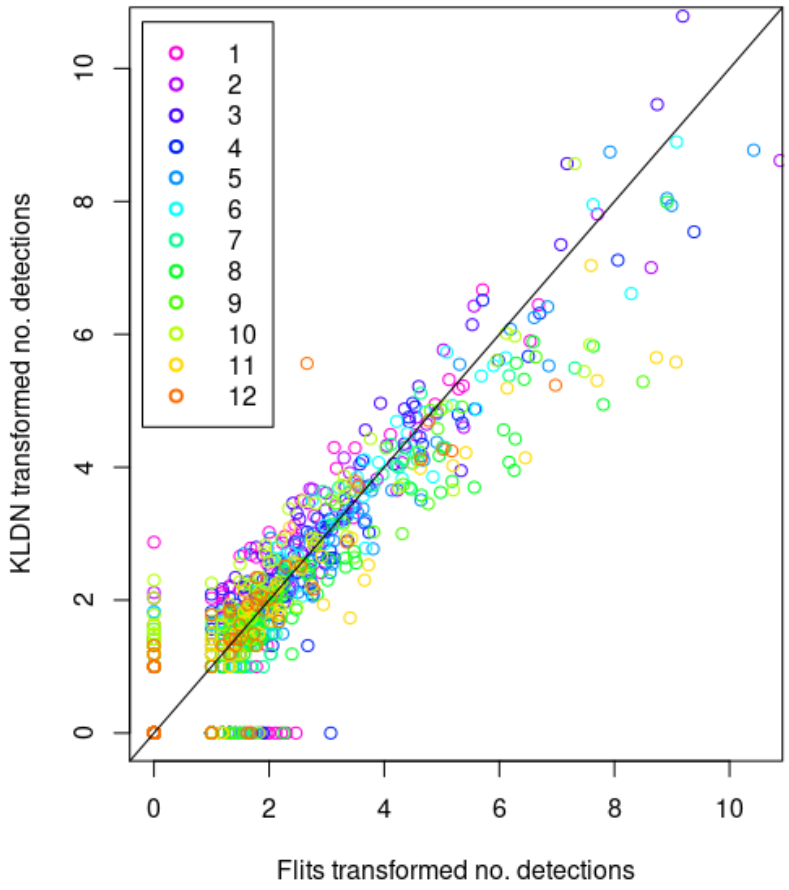


Figure 13: Transformed (at power $\frac{1}{4}$) of lightning detections by KLDN versus coinciding transformed (at power $\frac{1}{4}$) number of FLITS detections over April 15^{th} to October 15^{th} of 2015. The 1:1-line is added for convenience. The region number is shown as colour for each sample.

Table 3: Conditional quantiles of thunderstorm intensity (in discharges per 5 minutes) for four lead times and with the standard KLDN detections, as well as for some perturbation experiments. The 50%, 80%, 90% and 95% values are shown in the table. In addition, for the standard KLDN detections, the number of thunderstorm cases is given.

Valid time (UTC)	KLDN					KLDN, 2015 FLITS				KLDN, 2015 doubled			
	No. cases	0.50	0.80	0.90	0.95	0.50	0.80	0.90	0.95	0.50	0.80	0.90	0.95
03-09	381	11	45	104	164	10	44	104	161	14	58	129	199
09-15	685	13	49	102	228	13	50	110	228	16	60	128	238
15-21	699	19	87	255	437	19	104	289	448	21	113	288	510
21-03	401	16	73	182	292	17	78	185	345	20	92	197	380

that 373.254 discharges in the KLDN lightning intensity dataset in 2015 is at least more likely to be realistic in comparison to the rest of the KLDN dataset than 186.627.

Another reason to use not only FLITS as alternative truth is found in Table 3. It shows that for the valid times between 03 and 15 UTC, climatology when KLDN observations in 2015 are replaced by FLITS are hardly affected in terms of distribution. Though, as shown in Figure 13, the ranking of events will change order.

In addition, FLITS detections of 2015 leads to a slightly different set of thundery cases, whereas doubling KLDN maintains all thundery cases. So with FLITS in 2015, some of the cases, ranking of cases and eventual lightning intensities are adjusted, whereas doubling KLDN is a straightforward and linear adjustment. For the verification, the combination of adjustments makes the adjusted set with FLITS detections in 2015 somewhat less comparable to the reference KLDN than the set with doubled KLDN intensities in 2015.

The intensities shown in Table 3 show also the range of training thresholds used, since the lowest threshold is the median ($q_{0.50}$) and the highest $q_{0.95}$ for training (Table 1). For 25 to 100 discharges per 5 minutes, the two adjusted truth experiments will show whether the skill would have the potential to change and how, if the observations would be more homogeneous. That means: it gives an indication whether some skill in discrimination between severe and ordinary thunderstorms could increase and with a very rough estimate of the change in skill. However, since lightning detection systems are far from convertible into each other, this may be significantly off reality. Besides this, it is of interest whether the upper lightning intensity quantiles and thresholds up to which skill persists change with an adjusted number of detections.

It is emphasised that a correction could also be appropriate for the 2017 network changes in KLDN, in particular in the northern regions, and that a 2015 region-dependent and intensity-dependent correction would potentially remove some non-linearity issues, but this study is in no way an attempt to find out details of dissimilarities in lightning detection datasets. However, it is useful to address dissimilarities shortly. An example of regional dependence of lightning intensity as derived from non-homogeneous detection systems is shown in Figure 13. It can be seen in this figure that region 11 probably has a large difference in detection efficiency between KLDN and FLITS: most of these orange points are clearly below the 1:1-line. Regions 8 and 9 may suffer similar issues and there is some indication of non-linearity, as the deviation from the 1:1-line looks stronger for higher intensities. In contrast, region 1 (purple) seems to have a slightly higher detection efficiency in 2015 in KLDN than in FLITS, with points above 1:1-line.

5 Results on comparison of statistical methods

In this chapter the results of a comparison between the statistical methods are explained in detail. First it describes how the hyperparameter settings have been selected. Special attention is paid to the number of predictors that is included in the fits. Subsequently, the comparison between (E)LR and QRF is done for the two predictands (thunderstorm occurrence and lightning intensity), hereby computing several verification measures.

5.1 Quantile regression forests: hyperparameters

5.1.1 Number of predictors tried for each split (m_{QRF})

For the QRF, there are several hyperparameters that are used in the initial cross-validations (see Table 1). The eventual hyperparameter selection is done with the 90 and 15 potential predictor sets only, because both sets gave very similar results.

The first hyperparameter is m_{QRF}. It appears to be the most important hyperparameter from the tests, considering the results of the cross-validation: 27 out of 28 cross-validation results (7 lead times, 2 potential predictor sets and 2 predictands) lead to the same first selection. The tests of 2, 6 and 10 for m_{QRF} lead to different results, with $m_{QRF} = 2$ usually being clearly better and sometimes equally good as the best of the other two options.

The interpretation of the hyperparameter selection that even with many predictors, there is always useful information contained in any predictor plane. If QRF searches in a six dimensional space, it is searching for a very good split among many potential splits, which in a late stage of building a tree can restrict how random the trees are and enlarge among trees. If trees are more random, various correlated predictors can more likely complement each other than when trees are more similar.

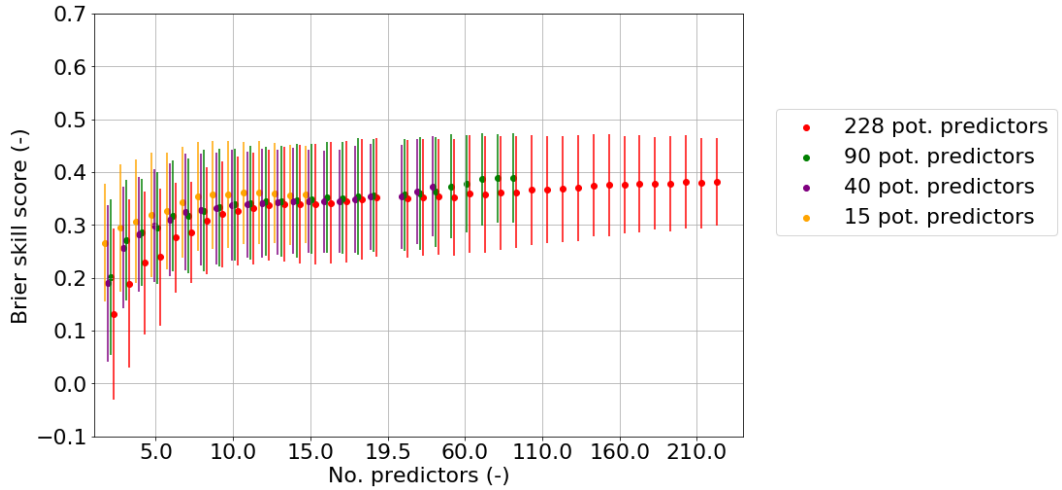
5.1.2 Minimum terminal node size (s_{QRF})

The next hyperparameter is the minimum size of a sample in the terminal node in a tree (s_{QRF}), which turns out to be less crucial in the end: settings 3, 9 and 15 are tested (Table 1). Eventually, for thunderstorm occurrence forecasts, $s_{QRF} = 15$ always leads to better results than any of the smaller values of s_{QRF}. For the lightning intensity forecasts, the signal is slightly less convincing with $s_{QRF} = 9$ as most skillful setting when skillful forecasts can be made for the 15 potential predictors set. For the 90 potential predictors set, there is a mixture of preferences among the seven lead times: s_{QRF} of both 3 and 9 are chosen three times. In the end, the decision that $s_{QRF} = 9$ is the selected setting for lightning intensity forecasts is made, because then the sets of 15 and 90 have the same settings for the two hyperparameters described in this section. With this, we assume that the selected hyperparameter settings are also good for models based on other potential predictor sets. Visualisation of verification scores leading to the selection are not shown in this thesis.

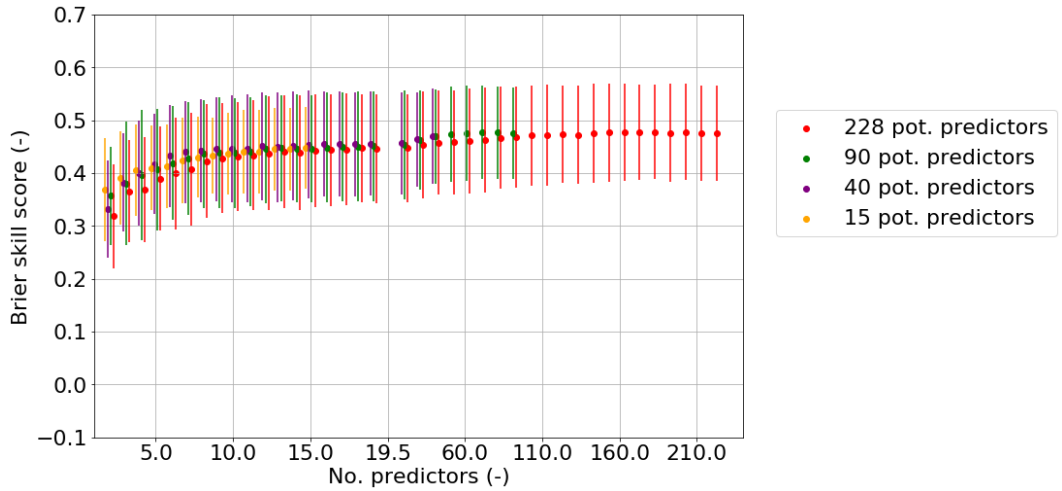
It can be said that thunderstorm occurrence forecasts favour a larger sample size in the terminal node than lightning intensity forecasts, which is partly explained by the larger unconditional training dataset for thunderstorm occurrence forecasts. In the lightning intensity dataset, QRF needs to find more similarity between samples and hence smaller s_{QRF}, because that dataset covers only a region of predictor space in which thunderstorms occur. For thunderstorm occurrence forecasts, it could be interesting to test some larger values of the hyperparameter.

5.1.3 Number of predictors (n_{QRF})

The results shown in this section are based on the initial cross-validation. A comparison is made between the results from different initial potential predictor sets on forecasting thunderstorm occurrence and intensity using QRF. Figure 14 shows the effect of the presence of many or fewer predictors on the scoring of QRF models during the full elimination process for two lead times, initially each 10 elimination steps. For the last 20 predictors the effect of each single elimination step is shown. Elimination is stopped



(a) Forecasts of 00z + 3 to + 9 hours verified.



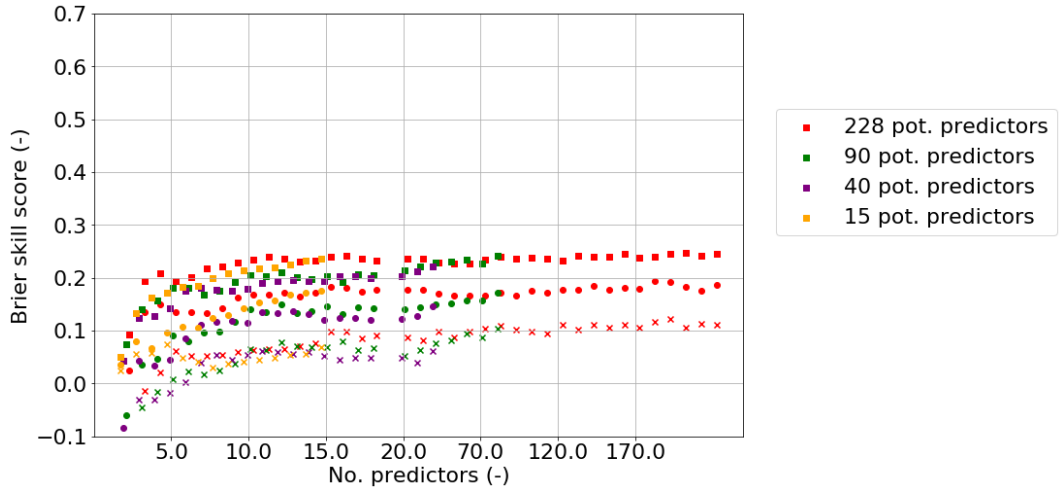
(b) Forecasts of 00z + 15 to + 21 hours verified.

Figure 14: Brier skill score as function of number of predictors for QRF initial cross-validation on probabilistic thunderstorm occurrence forecasts. The mean score over 12 regions is indicated as dot, with error bars indicating $1\sigma_{reg}$ *from the mean score.*

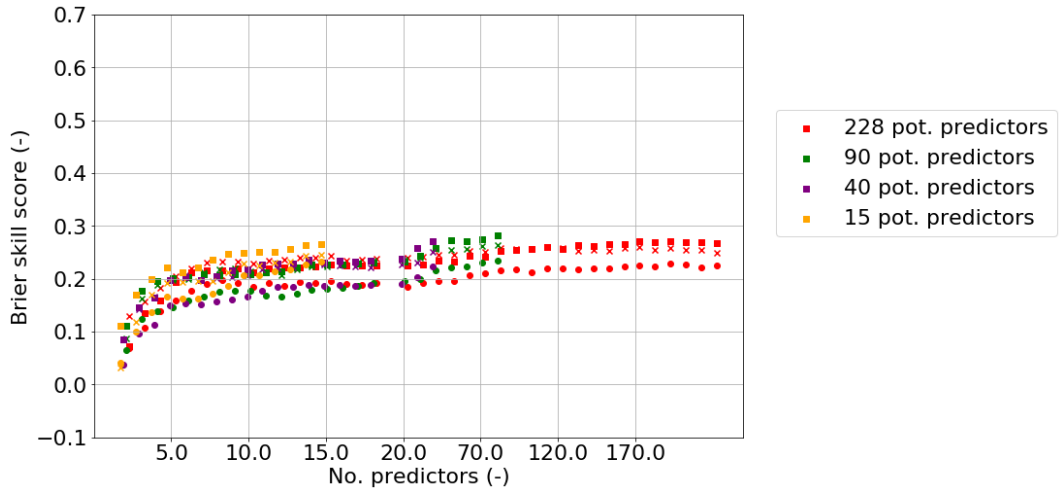
when 2 predictors are left. The most general feature is that starting with 2 predictors (the end of the elimination), the skill seems to go to some saturation level when increasing the number of predictors. However, for the +3 to +9 hours forecast, the maximum skill score with 15 potential predictors seems to be reached at around 10 predictors. In the other cases, the maximum score is typically reached with the maximum number of predictors, but the BSS for 228 potential predictors does not exceed that with 40 and 90 potential predictors when all potential predictors are included as predictors in all sets. Furthermore, with the same number of predictors (after elimination steps), the smaller sets have better skill scores than the 228 potential predictor set.

Error bars give the scoring standard deviation of the 12 regions in Figure 14. These are typically smaller with smaller potential predictor sets, so they are the smallest with 15 potential predictors. This means that eliminating many predictors before having a "best" prediction is ineffective for skillfully forecasting thunderstorm occurrence. It only increases computational costs.

It is essential to note that error bars shown in Figure 14 are by no means informative about the



(a) Verification score (BSS) for forecasts of 00z + 3 to + 9 hours.



(b) Verification score (BSS) for forecasts of 00z + 15 to + 21 hours.

Figure 15: Brier skill score as function of number of predictors for QRF initial cross-validation on probabilistic lightning intensity forecasts. Three regional intensity thresholds are shown: 39 discharges per 5 minutes (squares), 81 discharges per 5 minutes (circles) and 150 discharges per 5 minutes (crosses).

significance of any difference: the 12 KOUW-regions on which the (one) standard deviation is based, can be strongly correlated in thunderstorm occurrence and intensity. That means, if region 5 has a thunderstorm on a certain day, it is very likely that some other regions will have thunderstorms too on that day, especially the neighbouring ones. So for any significance estimate, one would have to correct for correlations between all the regions. Furthermore, if significance tests would be applied to test differences in (optimal) BSS between the potential predictor sets, these differences are not going to be significant, with results being as close as in Figure 14.

The dependence of QRF performance on potential predictor set and its change during the elimination process for intensity thresholds is shown in Figure 15. Although the skill scores are less stable than for thunderstorm occurrence prediction, a saturation of skill with increasing number of predictors can still be identified and it can be seen that the skill scores are more or less constant for the biggest set with more than 90 predictors. Having a potential predictor set larger than about 90 predictors is not likely to help improving lightning intensity forecasts notably. Since (extended) logistic regression does not profit from so many predictors simultaneously being implemented in a model and it only would select the best few

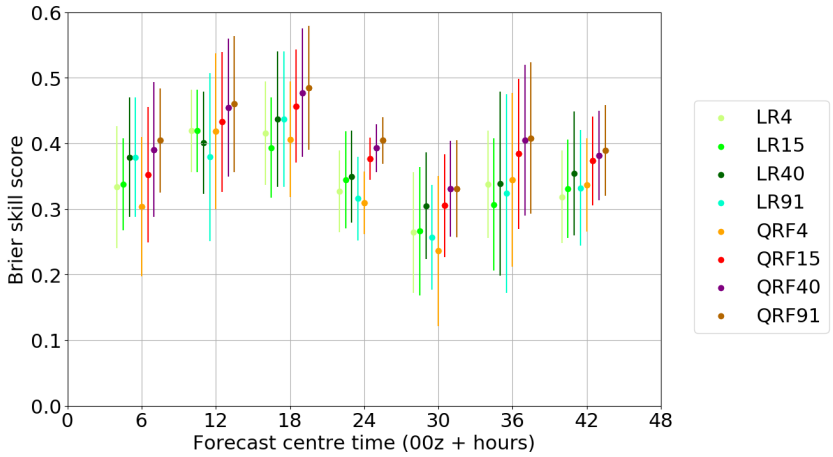


Figure 16: Comparison of Brier skill score as a function of lead time for all methods predicting thunderstorm occurrence, with $\mu_{reg} \pm \sigma_{reg}$ (indicated by error bars). Note that in some cases, two potential predictor sets lead to the same fit for LR.

whereby it profits mostly from complementary variables, for (E)LR it neither makes sense to use very large datasets of 228 potential predictors. Therefore, in the remaining sections, the search for optimal thunderstorm forecasting models is limited to those models using 91 potential predictors at maximum.

Since we have identified the saturation-like behaviour of QRF when it is stepwise eliminating individual predictors and that it does not suffer from overfitting notably, it can be said that using the 91 potential predictor set for QRF is generally at least as useful as a smaller predictor subset of this set after elimination in practice and the full set of 91 potential predictors will be used as biggest potential predictor set from now on: $n_{QRF} = n_{set}$.

5.2 (Extended) logistic regression and number of predictors

For the (E)LR method, the only hyperparameter to be set is the number of predictors. The number of predictors is set to the highest number for which the initial cross-validation verification shows an increase of BSS for most of the regions (LR) or a clear majority of the verified intensity thresholds (ELR) per lead time and potential predictor set. Sometimes, this criterion is somewhat arbitrary to judge, but usually it is relatively clear. In addition, attention is paid to whether the model is consistent with elementary relations, mainly that thunderstorm probability and intensity should increase with increasing thermodynamic instability and with increasing hydrometeor concentrations. In a few cases, a model is rejected due to the opposite relation. This is an indication of overfitting and the number of predictors is in these cases decreased by one (with another check of consistency). For thunderstorm occurrence, it leads to models with 1 to 4 predictors (typically $n_{LR} = 2$) and for lightning intensity, it leads to 1 to 3 NWP output predictors, in all but two cases of those presented in the remaining chapters having 1 or 2 physical predictors. The selected predictors are further discussed in Chapter 6.

5.3 Thunderstorm occurrence forecasts: logistic regression versus QRF

The hyperparameters have been selected now (Sections 5.1 and 5.2) and therefore we continue with the results of the final cross-validation in the remainder of Chapter 5 and in Chapter 6. In Figure 16 the BSS with four potential predictor sets and the two methods for thunderstorm occurrence forecasts are shown for seven lead times. From this figure, it is clear that QRF91 is probably the most skillful method, closely followed by QRF40. Logistic regression is never more successful than these methods in terms of skill score distribution among regions, though no significance testing is done. Among the logistic regression sets, LR40 is usually the most successful method, based on the BSS. For logistic regression, the skill score is not regularly increasing with increasing potential predictor set size; therefore the skill of LR looks less

stable in this figure. An important reason for that is that LR only profits from selected predictors and otherwise suffers from overfitting, whereas QRF exploits the full potential predictor set without suffering from overfitting. However, for QRF it can be seen that for all lead times, the saturation-like increase as a function of potential predictor set size is still present. In other words, the skill rapidly increases when the number of potential predictors goes from 4 to 15 for all lead times. The average BSS typically increases less rapidly from 15 to 40 predictors and only very slightly from 40 to 91 predictors. The uncertainty margins based on the BSS standard deviation among 12 regions are similar between QRF40 and QRF91, except for the +3 to +9 hour forecast, where it shrinks going from 40 to 91 predictors. The last important feature of the figure is that QRF is not better than LR when having only 4 potential predictors, but likely worse.

Besides these intercomparison results between methods and potential predictor sets, it can be seen that the forecasts centered around 18 UTC on the first day are most skillful followed by those centered around 12 UTC on the first day, as expected (highest climatological probabilities). It can also be identified that the average BSS always decreases if one jumps 24 hours forward in time (so for the same valid times); this is also consistent with expectations.

Thunderstorm occurrence forecasts have also been verified with reliability diagrams. Out of 48 combinations of valid time and region, only two combinations turned out to have somewhat diminished reliability at high thunderstorm probabilities (30-90%), namely the extreme northwestern region at night and the extreme southeastern region during daytime. These are largely independent of the potential predictor set and method used and are therefore not shown.

5.4 Lightning intensity predictions: extended logistic regression versus QRF

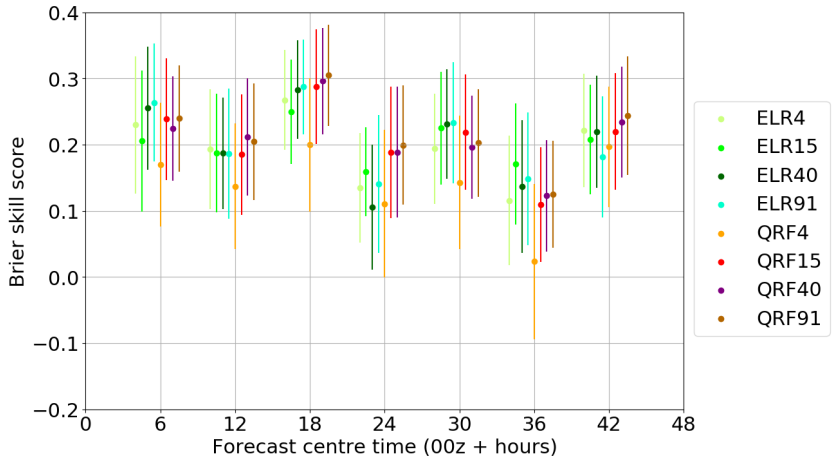
5.4.1 Cross-validation strategies

Due to the inhomogeneity issues described in Section 4.4, the results from the random final cross-validation procedure (but where all regions for a day are grouped) are effectively used instead of the one where we test on a year and train on the other two, for analysing the performance of lightning intensity forecasting models. All lead times indicate similar or better performance with the random cross-validation (see Appendix B). One could argue that randomly distributing days between test and train dataset could improve skill due to some correlations between previous case and current case in both predictor and predictand outcome. However, the predictand (observed transformed lightning intensity) correlations between subsequent and current case within all regions and for four valid times have been calculated. These are 48 combinations, of which 47 range between -0.23 and +0.26, with 15 negative correlations and 33 positive correlations. Part of the correlations are influenced by structural seasonal variations in lightning intensity, as in July and August typically more severe thunderstorms occur. Nonetheless, one region had a strong correlation of 0.68 for one valid time . This is region 1 for 21-03 UTC, where only 25 cases occur with only three cases of interest above 25 discharges per 5 minutes. With the given sign changes in aforementioned correlation coefficients no significant artificial skill is suspected due to the random cross validation strategy.

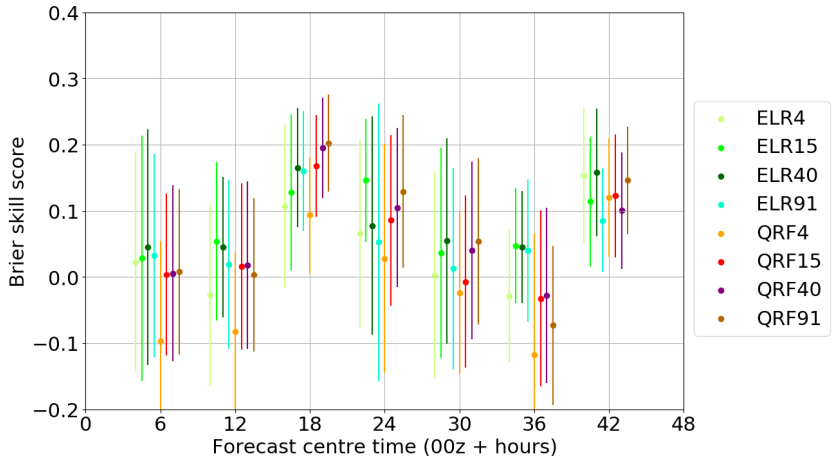
5.4.2 Brier skill scores of four potential predictor sets

The comparison of BSS for four potential predictor sets with the two methods and for seven lead times for lightning intensity forecasts is shown in Figure 17. Now, the confidence intervals indicate uncertainty, as the error bars in scoring show the upper and lower bound of the 95% confidence interval as obtained from 1000 block bootstrapping samples.

Verification scores of both lightning intensity thresholds show that typically ELR is not stably gaining information from giving it more potential predictors; the BSS of all potential predictor sets are within each other's 95% uncertainty bars. This is likely caused by the large set of potential predictors: with a big potential predictor set the initial cross-validation leads to different predictor selections than the final cross-validation frequently. The initial cross-validation provides the number of predictors selected for the final model and this number can be sensitive to the predictor selection. Therefore the connection between the initial and final cross-validation weakens when there are many potential predictors, leading to



(a) 50 discharges per 5 minutes



(b) 250 discharges per 5 minutes

Figure 17: Comparison of Brier skill score as a function of lead time for all methods for the indicated intensity threshold, with confidence intervals based on 1000 block bootstrapping samples indicated by error bars.

non-optimal settings, which can cause that models are more prone to overfitting. In principle, this should not happen, but small datasets are particularly vulnerable. Note that the initial cross-validation on which the number of predictors is selected contains only around $\frac{4}{9}$ of the datasets containing 381-699 samples for training for each valid time (Table 3). In QRF this would not be such a big problem, as extra information in a potential predictor set may rarely be used by hardly being selected for splitting and the skill likely goes to nearly constant values with large potential predictor sets (see Figure 15).

Among the QRF models it can be seen that QRF4 is typically worse than the other models, and its confidence interval sometimes is clearly lower than confidence interval of the other models at 50 discharges per 5 minutes. This means that it can be considered as not competitive to the other methods. Usually, QRF15 is close to QRF40 and QRF91. In general, the mean BSS of the QRF40 and QRF15 methods at lead time of +33 to +39 hours for the threshold of 250 discharges/5 min are at the bottom of the BSS confidence interval of the ELR15, ELR40 and ELR91 methods. This is probably because QRF cannot predict intensities that have not occurred in the (relatively small) training set, whereas ELR allows for some extrapolation (Equation 2). This happens with +33 to +39 hour forecasts at 250 discharges per 5 minutes, because it is above $q_{0.95}$ (see Table 3).

Since conditional probabilities are studied here and only thunderstorm cases are present in the dataset

for testing and training, results are in general not as stable as for thunderstorm occurrence due to the small sample sizes. This leads to somewhat more varying results between different lead times, but illustrates that it is not yet appropriate to draw definitive conclusions between the 15, 40 and 90 potential predictor sets for lightning intensity forecasts, especially because homogeneity issues might be relevant for the choice of the preferred method (see Section 4.4).

5.4.3 Reliability of QRF and ELR

Another important verification method is the reliability diagram, which shows whether a certain forecasting probability bin (for example 10-20% probability) is indeed associated with an observed predictand probability frequency that resembles this probability bin (see Section 3.4.1). The reliability diagrams of +39 to +45 hour forecasts of ELR40 and QRF40 are shown as illustrative example in Figure 18.

The QRF and ELR reliability diagrams show differential behaviour: the ELR method (Figure 18a, right) shows often (almost) exponentially decaying histograms of relative issuing frequency as a function of forecast probability for the high quantiles shown ($q_{0.90}$) in this example. On the other hand, QRF (Figure 18b, right) shows relative frequencies that are still high for the lowest probability bin of 0-10%, but frequently linearly decaying or constant issuing frequency for the next few probability bins, namely 10-20% and 20-30%. Therefore, the two reliability diagrams in this example show that QRF is more likely to issue forecasts with higher probabilities than ELR and the diagrams also show that QRF typically keeps this reliability up to higher probabilities. This can be seen, because the red lines of ELR with 1 predictor and QRF with 40 predictors are close to the optimal 1:1-line, for QRF up to the 60-70% probability bin and for ELR up to the 40-50% probability bin in Figure 18. Only one of the points below these probability bins are deviating clearly from the 1:1-line. Similar behaviour is found for some other lead times, of which reliability diagrams are shown in Appendix C.

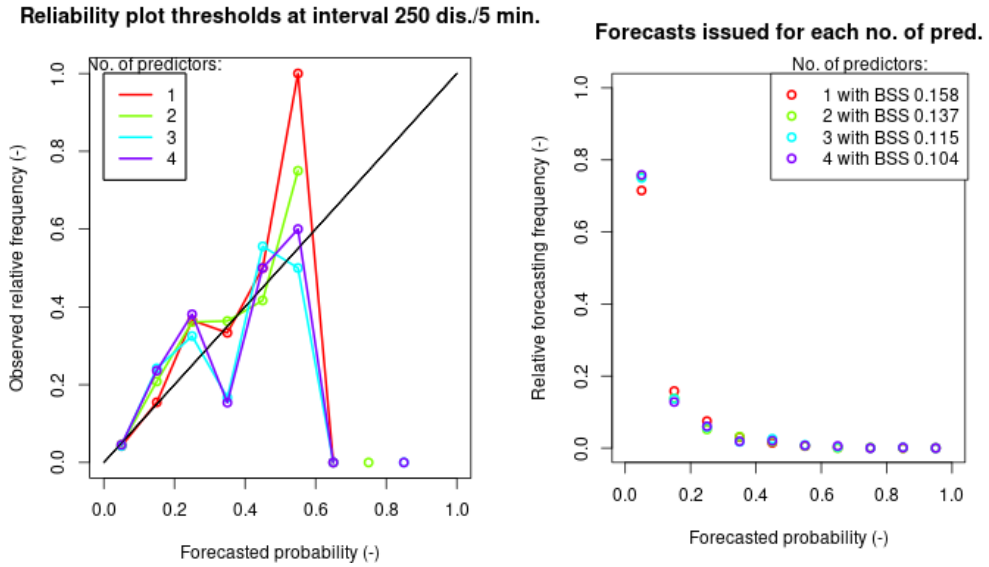
An important reason for the identified behaviour is that ELR makes assumptions on the shape of the PDF of the predictand. For ELR, the exponential distribution of issued probabilities will typically be the consequence of the shape of PDFs of transformed predictand and predictor(s) conditional on thunderstorm observations; the transformed predictand distribution clearly has a long tail (Figure 11) and for many continuous predictors this will hold as well. QRF may discriminate higher and lower probabilities empirically based on combinations of predictor values and since there are many trees that are random, this empirical part is combined with smoothing.

Since there is no clear signal that QRF or ELR typically performs better than the other based on BSS, higher forecast probabilities and better reliability with $\approx 50\%$ probabilities of QRF can be paid off by slightly larger forecasting errors when high lightning intensities are unlikely (say: 5% probability), which occurs more frequently for $q_{0.90}$ of the lightning intensities. For practical application and operational use, the reliability advantage of QRF is more favourable even though forecast quality might in general not improve, as warnings are not likely to be issued with very low probabilities. For operational and practical use, good reliability is desirable in combination with ability of the model to distinguish high probabilities of exceeding a high intensity threshold well from near-zero probabilities.

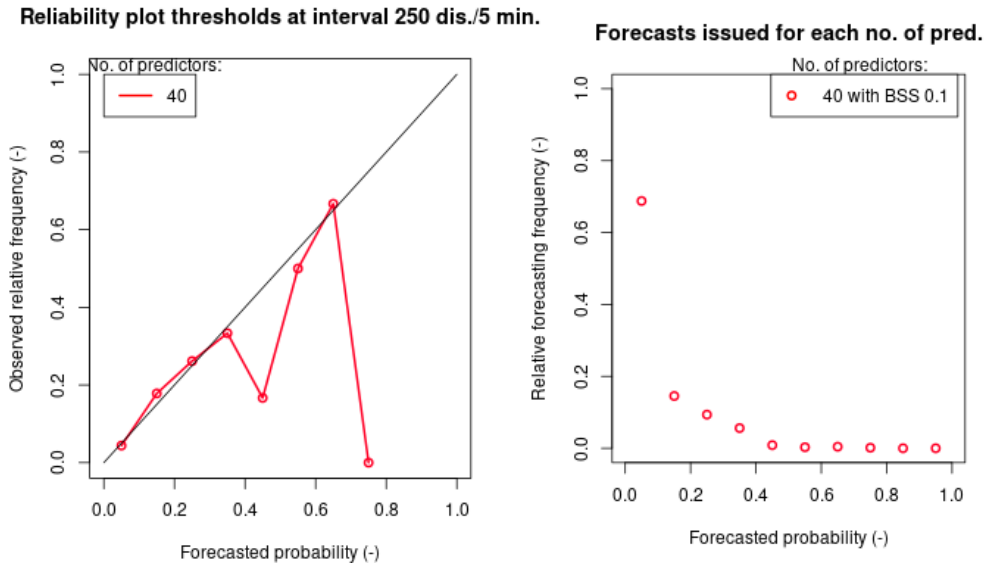
5.4.4 Continuous ranked probability skill score

In Table 4, the continuous ranked probability skill scores of ELR40 and QRF40 models can be found (see Section 3.4.1). It shows that the unweighted CRPSS for ELR40 ranges between -0.05 and +0.05. This means it is not more skillful in predicting the full lightning intensity distribution than climatology. However, QRF40 is successful in doing so: the skill scores vary between 0.10 and 0.19. The clear reason for this is that by intending to predict high intensities optimally, ELR is explicitly trained on the upper half of the observed lightning intensity distribution, whereas QRF is not trained on specific thresholds, but on the full empirical distribution. In the verification, ELR therefore does worse at low lightning intensities than QRF, as the whole lightning intensity distribution is verified with the unweighted CRPSS.

For verification purposes, the weighted CRPSS (wCRPSS) has also been calculated, with a weight of 0.1 for lightning intensities up to 25 discharges per 5 minutes and 1.0 for lightning intensities above



(a) ELR Harmonie00z +39 to +45 hours (selected: 1 predictor)



(b) QRF Harmonie00z +39 to +45 hours

Figure 18: Reliability diagrams of ELR40 and QRF40 forecasts, with both relative frequency of an event (LHS) and relative forecasting frequencies per forecast probability bin (RHS) for each lead time at lightning intensity of 250 discharges per 5 minutes ($q_{0.90}$). Note that for ELR, final models made with 1 to 4 predictors are all validated, but the model with 1 predictor (red) was selected in this case with the initial cross-validation verification.

Table 4: Unweighted and weighted continuous ranked probability skill score for the ELR40 and QRF40 models per lead time.

Lead time (h)	CRPSS		Weighted CRPSS	
	ELR40 (-)	QRF40 (-)	ELR40 (-)	QRF40 (-)
03 - 09	0.045	0.147	0.223	0.285
09 - 15	0.019	0.154	0.162	0.237
15 - 21	-0.033	0.190	0.158	0.230
21 - 27	-0.051	0.128	0.151	0.205
27 - 33	0.033	0.140	0.213	0.252
33 - 39	-0.007	0.098	0.129	0.184
39 - 45	0.002	0.136	0.134	0.176

25 discharges per 5 minutes (Section 3.4.1). The wCRPSS reflects the forecast skill at high intensities specifically: its value for QRF40 is systematically 0.04 to 0.08 higher than ELR40. This is very likely related to the identified higher frequency that high probabilities are issued by QRF than by ELR, as discussed in Section 5.4.3. In practice, ELR tends to infrequently issue relatively high probabilities (for example 40%) of exceeding a high intensity threshold, whereas QRF is more frequently able to forecast the higher probabilities of exceeding a high intensity threshold. In such cases when a high lightning intensity is indeed observed, ELR has the measured intensity systematically in the higher region of the CDF (low probability to be exceeded) compared to QRF, having the measured intensity typically more toward the centre of the CDF. This leads lower wCRPSS of ELR compared to QRF, whereas the BSS appears to be hardly affected. BSS can be hardly affected if ELR gives slightly lower probabilities than QRF in most verification samples when a threshold is not exceeded. Therefore, the CRPSS can add information about how good forecasts are and the systematically better skill of QRF in the wCRPSS indicates that QRF would likely be preferential in operational use, even though the BSS does not indicate systematic differences between QRF and ELR.

For the ELR91, QRF15 and QRF91 predictor sets, the behaviour of both weighted and unweighted CRPSS as a function of lead time closely resembles the pattern of the models with 40 potential predictors for the same method. In all cases, *QRF40* performs the best and QRF is clearly better than ELR, but with small differences between different potential predictor sets. For thunderstorm occurrence forecasts QRF40 is also optimal together with QRF91 (Section 5.3) and it is used in the next chapter to conduct more experiments.

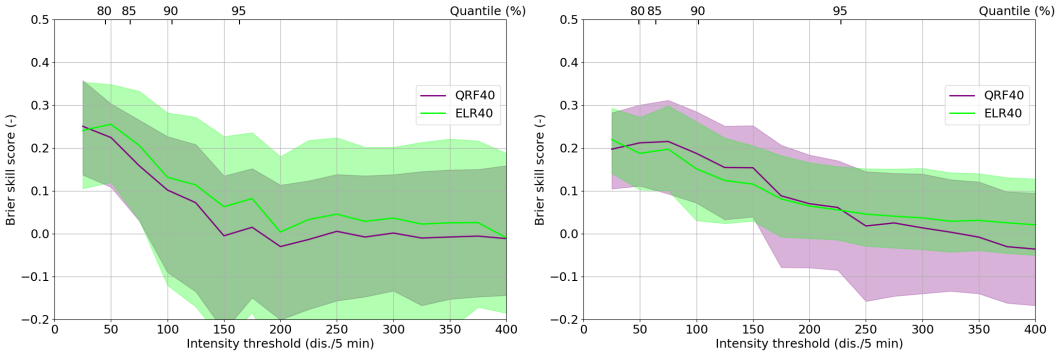
5.4.5 BSS as function of threshold: QRF40 and ELR40

Verification with BSS has been done between 25 and 400 discharges per 5 minutes with steps of 25 discharges per 5 minutes. Figure 19 shows that the BSS generally decreases with increasing lightning intensity, since the peak in BSS is typically observed with intensities of 25 or 50 and sometimes 75 discharges per 5 minutes. This is consistent with expectations as the observations still have a balanced distribution with many records above and many below the threshold leading to better sampling/more samples (see Chapter 4 and Table 3). At +21 to +27 hours ahead however, ELR40 has its peak around 200 discharges per 5 minutes, whereas QRF40 shows a small decrease in BSS with increasing lightning intensity threshold. The maximum BSS of ELR40 at 200 discharges per 5 minutes is not significant though, but none of the lead times show a significant difference between ELR40 and QRF40.

When comparing QRF40 and ELR40, QRF seems to perform better in terms of BSS for +9 to +21 hour forecasts in the skillful threshold range. For +21 to +27 hours the best method based on BSS changes as a function of threshold. For +3 to +9 and +27 to +45 hours, ELR seemws to perform better. These results are more or less consistent with Table 4: when the wCRPSS difference is maximum between ELR40 and QRF40 (0.07), QRF40 shows the best BSS for thresholds that have been verified; for lead times with smaller differences between QRF40 and ELR40 (0.04-0.06), ELR40 has the best BSS for verified thresholds. When forecasting +21 to +27 hours ahead with a wCRPSS difference of 0.05, the best method

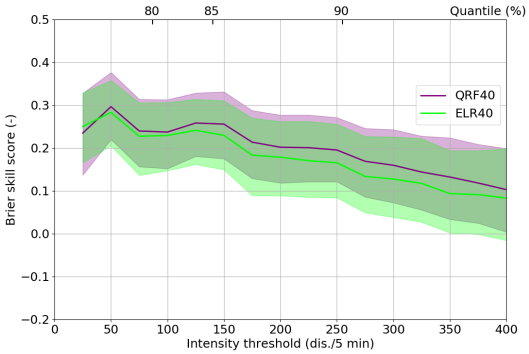
changes as a function of lightning intensity threshold. It should be noted that wCRPSS also takes into account thresholds for which BSS is not shown in Figure 19, such that BSS and wCRPSS do not need to be fully consistent.

If Figure 19 is interpreted in context of the unconditional climatology, both QRF and ELR are skillful at 95% confidence up to $q_{0.993}$ to $q_{0.995}$ in the first 33 hours, depending on the lead time and method. At longer lead times the quantiles decrease to $q_{0.989}$, except ELR at +39 to +45 hours ($q_{0.992}$).

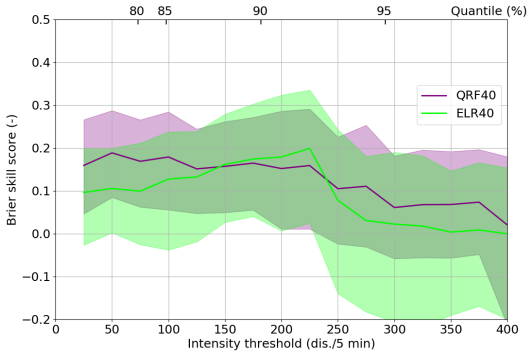


(a) +3 to +9 hours

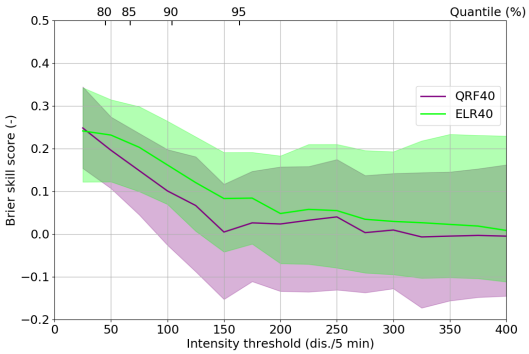
(b) +9 to +15 hours



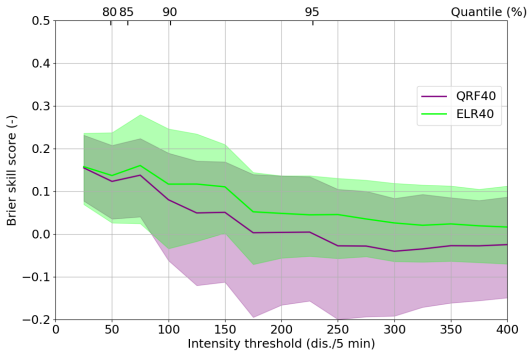
(c) +15 to +21 hours



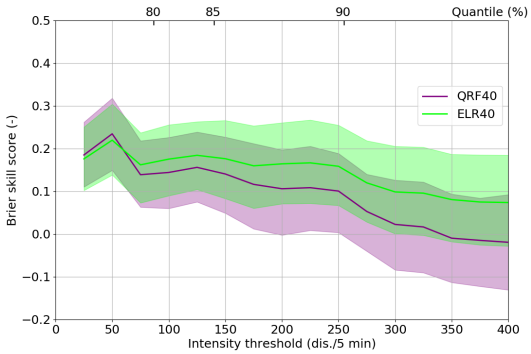
(d) +21 to +27 hours



(e) +27 to +33 hours



(f) +33 to +39 hours



(g) +39 to +45 hours

Figure 19: BSS of QRF40 and ELR40 as a function of lightning intensity for seven lead times. The four highest training quantiles are also given at the top (if within axis limits).

6 Role of specific predictors, predictor groups and lightning detection system

In this chapter, the value of individual predictors and physical groups of predictors for the probabilistic thunderstorm forecasts is investigated, to understand their role in the thunderstorm forecasts. Three experiments are carried out with the 40 potential predictor sets, namely leaving out vertically integrated buoyancy measures of parcels (no_CAPE), cloud content in the NWP model (no_mph) and lastly the precipitable water-instability combinations (no_PWinst), which are new in this study. With these experiments it can be understood what happens with forecast models if a group of predictors is removed from the potential predictor set. Results are analysed both from verification perspective and the role predictors have in fits, namely their selection in case of LR and their importance in case of QRF. The experiments have been summarised in Table 2. The verification of the elementary 15 potential predictor set (Section 2.2.6) is also displayed in figures as comparison and the chapter starts with experiments on thunderstorm occurrence forecasts.

Furthermore, we investigate the impact of modifying the lightning detection dataset on the verification scores (Section 6.5). In the last paragraph (Section 6.6) LR models are compared among lead times to give more insight in post-processed thunderstorm forecasts as function of lead time and valid time.

6.1 Comparison of the potential predictor sets: thunderstorm occurrence

With the no_CAPE experiment, 8 potential predictors are omitted, namely both surface based and most unstable CAPE, CIN and combinations of MUCAPE and MUCIN and MUCAPE-microphysics combinations; see Appendix A. Furthermore, when we do the no_PWinst experiment, six predictors are removed from the set of 40 potential predictors. Lastly, 4 potential predictors are removed with the no_mph experiment: snow, graupel and their combinations with MUCAPE.

Figure 20 shows the results when these three combinations of potential predictors are removed. For thunderstorm occurrence predictions, it can be seen that removing the PWinst predictors has little influence on the BSS. However, removing microphysics or CAPE results in a slight reduction in the skill of the QRF forecasts; these results are not tested for significance, but the signal is consistent between lead times. Only the +15 to +21 hours forecast is not affected when the microphysics is removed from the potential predictor set. In the QRF experiments, the removal of CAPE seems affect the verification scores most negatively, based on these seven runs. Except for the +33 to +39 hours forecast, the 15 potential predictor set performs worse than any of the three removal experiments. Additionally, the skill scores of QRF at +39 to +45 hours are very close together.

The results for logistic regression are less consistent and only the removal of microphysics has a consistently negative effect on all lead times, with on average also the largest negative effect. None of these effects are significant, but the signal persists through different lead times. For the +9 to +15 hours forecast, the LR40 and LR no_mph perform the worst of all based on average BSS. Furthermore, for lead times longer than 9 hours, the QRF40, QRF no_CAPE and QRF no_mph all perform better than all of the LR-experiments, with in the +3 to +9 hour forecast approximately equal performance between the best LR-runs and QRF no_CAPE and QRF no_mph runs.

6.2 Predictor importances: thunderstorm occurrence

6.2.1 QRF: importance as a function of lead time

In order to investigate the importance of a predictor in QRF, the time series of the predictor of interest is randomly reordered, which removes its predictive capacity. For both the ordinary and reordered time series, the prediction accuracy of QRF is evaluated on independent samples. By analysing the prediction accuracy with both ordinary and reordered sample order, an permutation importance measure is calculated (as mentioned in Section 3.2.3). The procedure is done for all predictors separately. When many correlated predictors are present, part of the benefits of one predictor will be absorbed by other predictors, such that

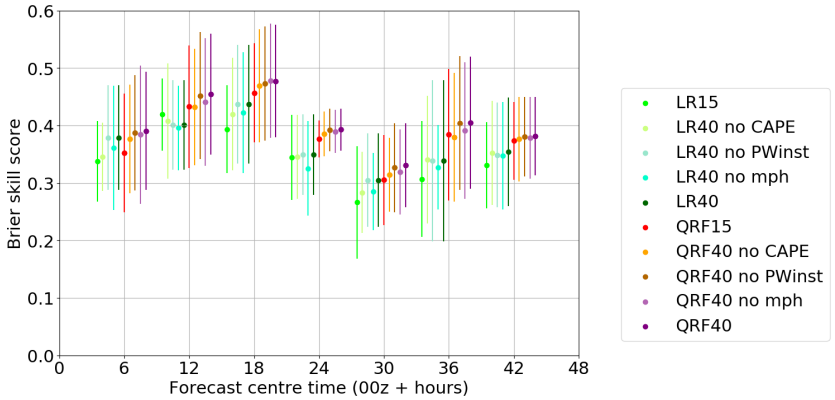


Figure 20: Comparison of Brier skill score for thunderstorm occurrence as a function of lead times for the standard LR40 and QRF40 methods and the two methods with no_CAPE, no_mph and no_PWinst potential predictor sets. In the figure $\mu_{reg} \pm \sigma_{reg}$ is indicated by error bars. Note that in some cases, two potential predictor sets lead to the same fit for LR. The LR15 and QRF15 predictions from the previous chapter are also included for convenience.

two very strongly correlated variables (r approaches 1) will share all of their importance measure if they are both present in the predictor set, but one of them gets all of the importance when the other would be eliminated. This has some effects on the importance measure [Gregorutti et al., 2016]; the permutation importance measure is the best importance measure that can be used in the QRF fitting package [Wright and Ziegler, 2017] used according to [Gregorutti et al., 2016]. For a random variable the permutation importance has an expected value of 0 with some spread and is to some extent depending on the random components in each QRF, whereas large values are obtained for important predictors.

When studying the importance measures of only QRF40 as a function of lead time, a few other interesting patterns pop-up. It can be seen in Figure 21 that importance of predictors for forecasts valid in the night and morning (21 UTC to 9 UTC) is very similar among all lead times. During daytime (9 to 21 UTC), when surface based convection occurs more frequently than in the night and morning, the importance of LNB, SBCAPE and its combination with SBCIN increase strongly. SBCAPE is approximately as important as Modified Jefferson, MUCAPE and the MUCAPE-graupel combination between 9 and 15 UTC. Additionally, a pattern with increasing importance of graupel, K-index, maximum Fateev and minimum Adedokun2 Index during daytime periods can be seen. This feature is very clearly present and therefore it seems a robust result. Besides this, it can be seen that $\Theta_{w,850,max}$ is in all cases more important on the second day than on the first day. Therefore the crosses are to the right of the circles for this predictor. For most predictors, consistent signals like this do not arise, but for Jefferson index (increasing on day 2) and MUCAPE-MUCIN as combined predictor (decreasing on day 2), this pattern is also found. It may be chance that some of these patterns arise though, as differences between day 1 and 2 are relatively small.

The general pattern in Figure 21 is that MUCAPE as integrated measure of thermodynamic instability is the most important predictor and Modified Jefferson is competitive as bulk approximation. Furthermore, the microphysics and other bulk instability measures such as Jefferson index (closely correlated to Modified Jefferson) and K-index (as well as SBCAPE and LNB during daytime) do clearly provide information on thunderstorm occurrence and are ranked high. A more intermediate rank is taken by $\Theta_{w,850,max}$, but it gives another type of information than predictors that neighbour it with intermediate-to-high ranking. Other predictors informing QRF about moisture, wind profile, circulation (MSLP, $\frac{dp}{dt}$) and (solely) CIN do not seem to be essential for improving thunderstorm forecasts. They end up with low importance in Figure 21. In summary, the depth and magnitude of instability and occurrence of showers is found very important and the potential wet bulb temperature is informative for thunderstorm occurrence forecasts with QRF.

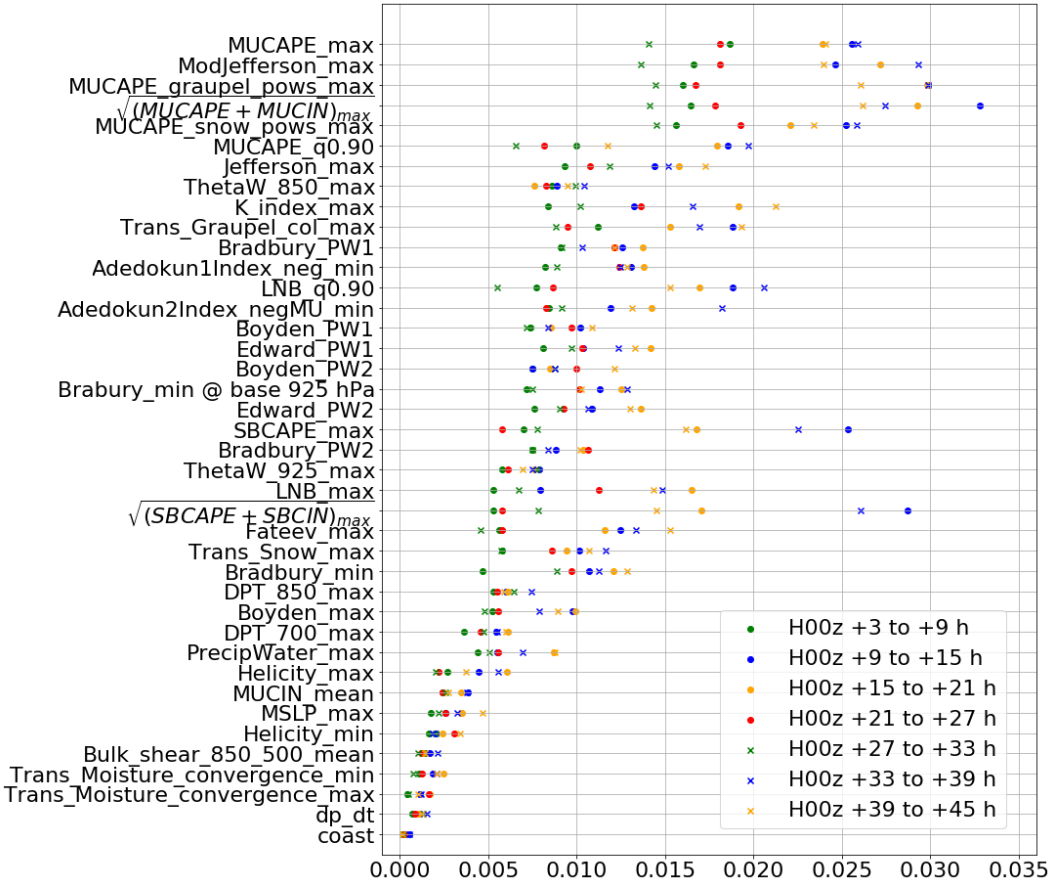


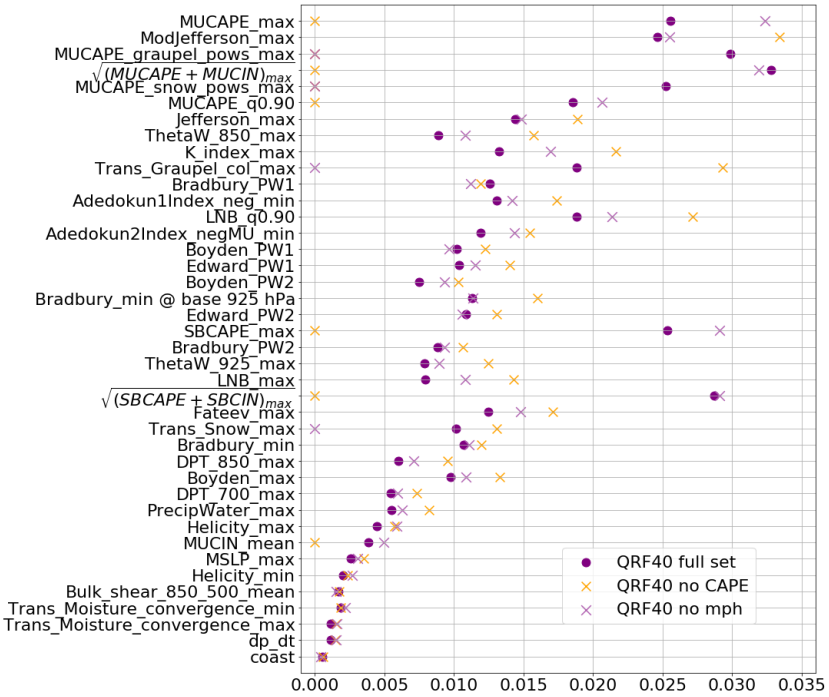
Figure 21: The permutation importance measure of QRF40 fits for seven lead times with thunderstorm occurrence forecasts, averaged over three final cross-validations. The colour of a symbol indicates the time of the day; circles indicate that the centre time of the forecast lies in the first 24 hours and crosses indicate a centre time on the second day.

6.2.2 QRF: importance change in no_CAPE and no_mph experiments

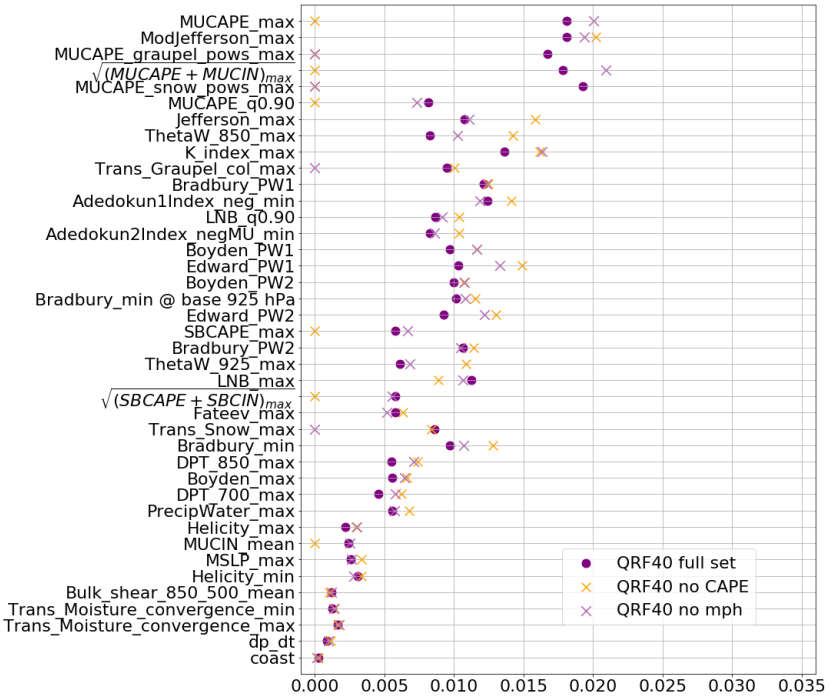
The permutation importance measure can give an indication as to the alternative information that QRF uses, when one or more predictors are removed. Some general signals as they are found in Figure 22 are now discussed. A clear feature in this figure at +9 to +15 hours is the strong increase in the importance of column graupel and LNB (both $q_{0.90}$ and maximum) in the no_CAPE experiment compared to the full QRF40 models. Furthermore, $\Theta_{w,850,max}$, maximum K-index, maximum Modified Jefferson and Fateev show strong increases. At lead times of +15 to +21 and +33 to +39 hours (not shown), the signal is the same, although the increase in K-index is sometimes replaced by a variant of Bradbury index, which is strongly correlated to K-index.

In the night and morning (up to +9 hours lead time and +21 to +33 hours lead time), CAPE is generally replaced by indicators of moisture and warm air in the lower atmosphere, with some bulk (potential or conditional) instability indicators over the 850 to 500 hPa layer also becoming more important. An example is shown in Figure 22. That means, instead of graupel, Fateev and Modified Jefferson, the potential wet bulb temperature at 850 and 925 hPa are more important in combination with Bradbury and/or K-index and Jefferson index.

The removal of graupel and snow is solved by the models with using MUCAPE and/or its combination with CIN more intensively, $\Theta_{w,850,max}$ and varying instability/PW combinations are also used more intensively during nighttime (Figure 22). Additionally, Bradbury based at 925 hPa, Adedokun2 Index, LNB,



(a) +9 to +15 hours



(b) +21 to +27 hours

Figure 22: The permutation importance measure of QRF models averaged over three final cross-validations; a linear correction for the number of potential predictors in a potential predictor set is applied. Some predictors have zero importance, because they are left out in that fit.

Table 5: First selected predictor in LR for three experiments, per valid time based on three-fold final cross validation with seven lead times. Sorting models by valid time means that the predictors on forecasting day one and two are merged in the same row. Other predictors used for LR40 can be found in Appendix D.

Valid time (UTC)	**Predictor**	**Selection frequencies (1^{st} pred.)**		
		LR40	**LR40 no_CAPE**	**LR40 no_mph**
03-09	ModJefferson_max	1	6	1
	MUCAPE_graupel_pows_max	2	0	0
	$\sqrt{MUCAPE + MUCIN}_{max}$	3	0	5
09-15	MUCAPE_graupel_pows_max	3	0	0
	MUCAPE_snow_pows_max	2	0	0
	$\sqrt{MUCAPE + MUCIN}_{max}$	1	0	6
	ModJefferson_max	0	5	0
	Trans_Graupel_col_max	0	1	0
15-21	ModJefferson_max	5	6	5
	$\sqrt{MUCAPE + MUCIN}_{max}$	1	0	1
21-03	ModJefferson_max	2	3	2
	$\sqrt{MUCAPE + MUCIN}_{max}$	1	0	1

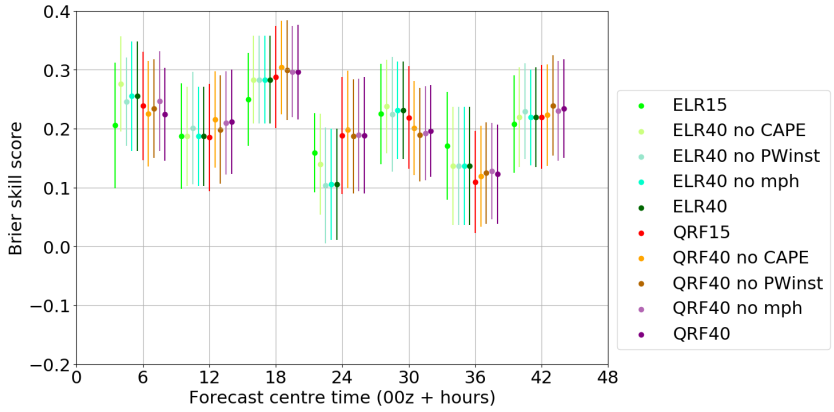
Fateev, $\Theta_{w,925,max}$, SBCAPE and K-index are predictors that are more informative to QRF for certain lead times.

In summary, there is a little consistency among the lead times in how some of the information contained in CAPE is replaced in QRF models: $\Theta_{w,850,max}$ and either maximum of Jefferson or Modified Jefferson (which are strongly correlated) have a more important role when CAPE is not available as predictor, with scattered signals among the roles of other predictors. For graupel the signal is even more obscure if several lead times are compared.

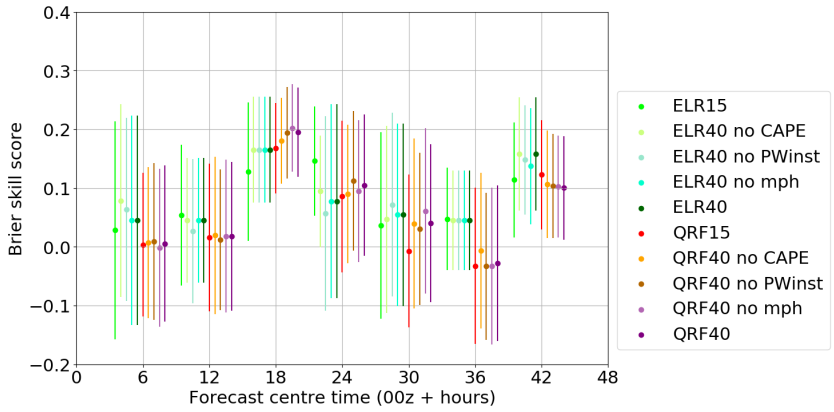
6.2.3 LR predictors

In Table 5 the first predictors selected during the forward predictor selection of LR40 and two removal experiments are shown. Four predictors dominate the table: maximum Modified Jefferson, $\sqrt{MUCAPE + MUCIN}_{max}$ and the two MUCAPE-microphysics combinations. Modified Jefferson is the most important predictor from the LR40 between 15 and 3 UTC and the CAPE composites between 3 and 15 UTC. When all CAPE predictors are removed, it is very clear that Modified Jefferson is the most favourable predictor to replace MUCAPE composites, with 20 selections as first predictor out of 21. Transformed graupel is only selected once as first LR predictor. With the no_mph experiment, it can seen that the models have a tendency towards more consistency, with the most unstable CAPE-CIN combination dominating between 3 and 15 UTC and Modified Jefferson doing so between 15 and 03 UTC. The dispersion of first LR predictors means that the three types of predictors are approximately similar in their ability to discriminate between non-thunderstorm and thunderstorm conditions. It is consistent with Figure 21, except for that untransformed $MUCAPE_{max}$ does not fit well in a logistic regression curve, which is because its relation to thunderstorm occurrence does not resemble a logistic curve. Therefore composite predictors such as $\sqrt{MUCAPE + MUCIN}_{max}$ have been created and the composite predictors with microphysics or CIN are preferred.

The second, third and fourth predictors of LR40 are shown in Appendix D. Besides the predictors that also appear as first predictors (and are already discussed), we see that Boyden index and Jefferson index appear twice as second selected. The Jefferson index complements LR models with similar information as Modified Jefferson, as their definition difference is only the inclusion of the dew point depression at 700 hPa and both are strongly correlated. Boyden is different from the other instability indices because it only assesses a lapse rate in the layer below 700 hPa compared to 700 hPa via thickness, whereas other instability indices describe mainly conditional instability between layers of 850 and 500 hPa. Furthermore, pressure, moisture convergence (divergence) and helicity are used complementary to instability indices in



(a) 50 discharges per 5 minutes



(b) 250 discharges per 5 minutes

Figure 23: Comparison of Brier skill score as a function of lead times for the standard LR40 and QRF40 methods and the two methods with no_CAPE, no_mph and no_PWinst potential predictor sets. The 95% confidence for indicated intensity is based on 1000 bootstrapping samples is indicated by error bars. Note that in some cases, two potential predictor sets lead to the same fit for ELR. The ELR15 and QRF15 predictions from the previous chapter are also included for convenience.

LR40 models in one or a few cases; they give information about forcing, low level moisture and organisation of convective cells (helicity), but do not appear very frequently in LR-equations.

In summary, like with QRF, instability indices are by far the most important for thunderstorm prediction in the LR models closely followed by microphysical predictors and then by a little information about forcing that is used. This order is largely consistent with QRF importance measures, although information about low level air mass ($\Theta_{w,850,max}$) is used by QRF (Figure 22), but not by LR.

6.3 Comparison of the potential predictor sets: lightning intensity

The impact on skill scores for experiments with QRF verified on lightning intensity forecasts is generally small when removing a certain type of predictor and acts in both directions (Figure 23). Typically, the microphysics has small impact and CAPE typically the strongest.

For ELR it can be said that the microphysics scheme typically does not influence forecast skill, except for the last lead time at +39 to +45 hours. This is because microphysical predictors are usually absent in the models; the lightning intensity forecast is conditional on thunderstorm occurrence. When thunderstorms occur, graupel and snow should be present in Harmonie, if Harmonie simulations are correct. The graupel and snow content in Harmonie should therefore not be very informative for conditional thunderstorm

Table 6: First ELR predictor for three experiments, per valid time based on three-fold final cross validation with seven lead times. Sorting models by valid time means that the predictors on forecasting day one and two are merged in the same row. Second predictors used for ELR40 can be found in Appendix D.

Valid time (UTC)	**Predictor**	**Selection frequency** ($1^{st}pred.$)	
		ELR40	**ELR40 no_PWinst**
03-09	Bradbury_PW1	6	0
	Bradbury_min	0	2
	DPT_850_max	0	1
	K_index_max	0	2
	ThetaW_850_max	0	1
09-15	Bradbury_925_min	1	1
	Bradbury_min	1	1
	K_index_max	3	3
	Adedokun1Index_neg_min	1	1
15-21	Bradbury_PW1	1	0
	K_index_max	5	6
21-03	Boyden_PW1	1	0
	K_index_max	1	1
	MUCAPE_q0.90	1	1
	Boyden_max	0	1

intensity predictions. Removal of CAPE and PW-instability combinations, has small impacts in both directions. For no_CAPE and no_mph, this small impact is expected, because these predictors turn out to be not selected by ELR, but for no_PWinst, this is at first not expected. This is because PW-instability combinations are important for conditional thunderstorm intensity predictions, which is about to be clarified in the next section, Section 6.4, and Table 6.

6.4 Predictor importances: lightning intensity

6.4.1 ELR predictors

The important predictors for lightning intensity forecasts for ELR40 and ELR no_PWinst are shown in Table 6. For 03-09 UTC valid times, there is clearly a preferred predictor in Bradbury_PW1. At other valid times, K-index is preferred over other predictors, but Bradbury and its versions with precipitable water are also included, just like Boyden and its combinations with precipitable water and Adedokun1 index. Most unstable CAPE is used once at nighttime. These variables are generally strongly correlated in the conditional dataset, the part of the dataset where thunderstorms are actually observed. To give an example: in the whole Harmonie dataset, the correlation coefficient between maximum K-index and minimum Bradbury index is -0.83, but for the thundery cases, which are used in the intensity prediction, it decreases to -0.89. These predictors correlate with transformed lightning intensity with comparable magnitudes, namely 0.51 and -0.50. Both will therefore be similarly good predictors of lightning intensity and they can replace each other. With high values for K-index or low values for Bradbury index, large variation in lightning intensity occurs, whereas low lightning intensities occur at high values of Bradbury index.

Physically, it is reasonable that Bradbury index and K-index relate well, especially in moist conditions (which is typically the case in Dutch thunderstorm environments): K-index adds temperature and dewpoint at 850 hPa; it subtracts the 700 hPa dew point depression from this and the 500 hPa temperature. Bradbury index is equal to the potential wet bulb temperature difference between 850 hPa and 500 hPa. In moist environments, dew point temperature and wet bulb temperature go to the temperature with increasing moisture; the wet bulb temperature is in the middle between the two others. After conversion of wet bulb temperature to potential wet bulb temperature, qualitative effects do not change. Moreover, in drier mid-troposphere, both the dew point depression and potential wet bulb temperature at 500 hPa

will go down.

More of the potential predictors used have strong correlations among each other due to the fact that many use 850 and 500 hPa temperature. In addition, many instability indices correlate with transformed lightning intensity at a magnitude similar to maximum Bradbury and maximum K-index: conditional on thunderstorms, the correlations vary typically between 0.45 and 0.53 (sometimes of negative sign).

When the PW-instability combinations are omitted, they are replaced partly by their original instability indicator and partly by other covariates, such as dew point and Θ_w at 850 hPa during the morning hours and K-index for all valid times.

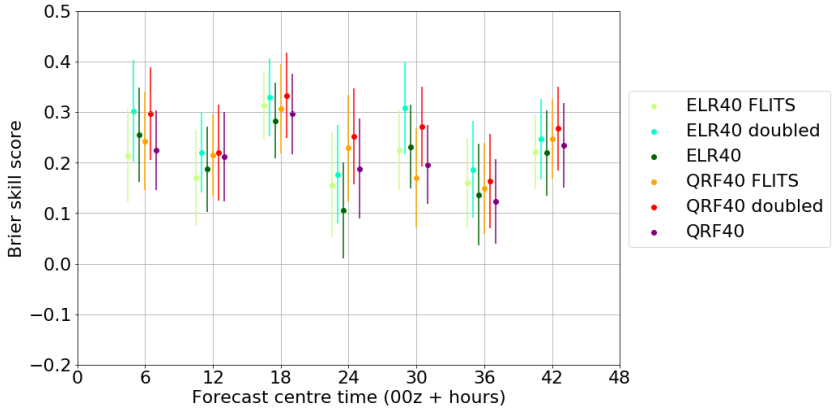
6.4.2 QRF40 predictors

In QRF40 fits, we mainly find the same important predictors as those that are first predictor in ELR40 (Table 6); see Appendix E. In addition to those important for ELR, predictors that turn out to be important at many or all lead times, are maxima of precipitable water and LNB. Furthermore, they are the PW2 versions of the PW-instability combinations and Bradbury based at 925 hPa. During the daytime, the maximum of MUCAPE is also important. For 15 to 21 UTC valid time, a few additional important predictors are found. However, the result is much more dispersed among predictors with this predictand than with the thunderstorm occurrence predictand, which indicates that all information is to some extend useful. Nonetheless, information about forcing and organisation of convection is ranked low, with coast, moisture convergence, pressure and its tendency, shear, helicity and CIN on the low end of the importance measure. Additionally, snow and Fateev have low importance values. All together, it means that steepness of lapse rates between 850 and 500 hPa and to a lesser extent instability depth and total or low layer (up to about 3 km/700 hPa) moisture content are most informative when making lightning intensity forecasts. A detailed figure of the importance measures can be found in Appendix E.

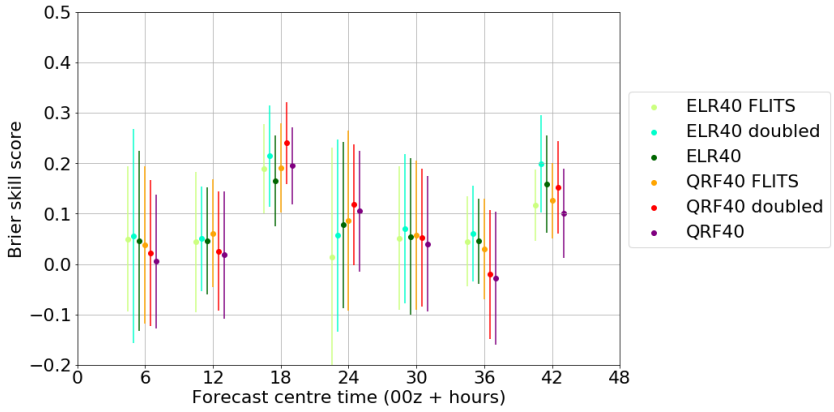
6.5 Sensitivity of intensity forecasts to lightning detection perturbations

Since the lightning intensity observations are inhomogeneous (Section 4.4), two sets of adjusted lightning detections in 2015 are used to gain some insight in what the variation in expected value of skill scores caused by inhomogeneity among and within years might be. The random cross-validation strategy (Section 3.3) is used for training and verification of the forecasts with adjusted truth. Although the uncertainty range would normally indicate the estimate for variation of the verification skill score, it is not necessary that the best score estimate for the homogeneous truth would fall with 95% certainty within the uncertainty range of the reference dataset with undisturbed KLDN detections. To estimate some uncertainty related to inhomogeneity of lighting detections, the detection dataset is modified in two ways.

Note that only the detection data of 2015 have been adjusted. In contrary, the documentation of Météorage [personal communication - Stéphane Pedeboy] shows that significant effects on detections will also have occurred due to the installation of extra sensors in De Kooy and Eelde in June 2017. The main interests are to what extent expected skill may be vanishing or growing due to inhomogeneity, to show whether the ranking of statistical methods can be affected, and whether there is large variability in the intensity threshold up to which forecasts are skillful. It is important to note that conditional quantiles of lightning intensity are influenced by the "truth" set (see Table 3), so a change in skillful quantiles are also of interest. Results for the two thresholds of 50 and 250 discharges per 5 minutes are shown in Figure 24. Improvements of best estimate for BSS can be seen among all lead times if doubled detections in 2015 are assumed as "truth". The BSS increases for this detection set by up to almost 0.1 at the 50 discharges per 5 minutes threshold for +3 to +9 hours (QRF), +21 to +27 hours (ELR) and +27 to +33 hours (QRF and ELR). In the latter case the best estimate of the doubled detections for 2015, the BSS is at the top of the 95% confidence interval of the reference detections. Note that for the most interesting lead time, +15 to +21 hours and +39 to +45 hours (highest lightning intensities, see Table 3 in Section 4.4), the difference among the generated detection sets is small with well overlapping confidence intervals. For the FLITS detections, the BSS is always closer to the reference detection set than for doubled detections and with strongest deviations between +21 and +27 hours for ELR (around 0.06 of BSS-value). This is where



(a) 50 discharges per 5 minutes



(b) 250 discharges per 5 minutes

Figure 24: Comparison of Brier skill score as a function of lead times for the reference LR40 and QRF40 fits and those with modified truths with FLITS detections in 2015 ("FLITS") and doubled KLDN detections in 2015 ("doubled"). In the figure the 95% confidence for indicated intensity is based on 1000 bootstrapping samples is indicated by error bars.

the least skillful forecast at threshold of 50 discharges per 5 minutes is made by ELR40 for the reference detections. BSS goes both up and down frequently with FLITS detections for 2015, with QRF giving mostly increases. This would indicate that especially QRF suffers from inhomogeneity in the detections, since the best estimate for BSS goes up with respect to the reference 13 out of 14 times.

At 250 discharges per 5 minutes, Figure 24 shows some minor changes in BSS, up to about 0.05. The doubled 2015 detections lead to mostly higher best estimates of BSS than reference detections, the exception is ELR at +21 to +27 hours. FLITS gives mostly higher BSS with QRF40 than the reference set. But for +21 to +27 hours, the detection set with FLITS detections in 2015 is notably worse than the reference set and the doubled set also performs worse. Usually, the differences between QRF and ELR are more similar within one lead time than those between the different detection sets for different lead times, but among different lead times there is no clearly preferred method.

In both figures, the most skillful statistical method varies among detection sets and lead times. This points out that the ranking can change just due to uncertainty in the observation.

The figures showing the uncertainty in BSS as function of threshold for each lead time are in Appendix F. For the +3 to +9 hours forecast 95% confidence, skill persists up to 150 discharges per 5 minutes in the QRF40 experiment with doubled intensities in 2015, whereas this is only up to 75 discharges per 5 minutes for the standard QRF40 and ELR40. It means that significant impact is seen at this lead time

Table 7: Summary of LR models for the no CAPE run with only one predictor as presented in Section 6.2.3, with intersection, predictor and coefficient of the model as function of forecast time. Given are model coefficients for three final cross-validations and their average. Additionally, the 10% and 50% thunderstorm probability predictor values are given and the difference between these two. Note that firstly all predictors refer to their spatial and temporal maxima, so that this part is omitted from their name.

Forecast time (h)	**First predictor** x	**Coefficient**	$x(P = 0.1)$	$x(P = 0.5)$	Δx
03-09	ModJefferson	0.864	32.79	35.33	2.54
	ModJefferson	0.920	32.80	35.19	2.39
	ModJefferson	0.766	32.63	35.50	2.87
Average 03-09	ModJefferson	0.850	32.74	35.34	2.60
09-15	ModJefferson	0.661	31.77	35.09	3.33
	ModJefferson	0.734	32.04	35.03	2.99
	Trans. col. graupel	5.009	0.98	1.42	0.44
Average 09-15 (2 models!)	ModJefferson	0.698	31.91	35.06	3.16
15-21	ModJefferson	0.723	32.00	35.04	3.04
	ModJefferson	0.741	32.19	35.15	2.96
	ModJefferson	0.822	32.49	35.16	2.67
Average 15-21	ModJefferson	0.762	32.23	35.12	2.89
21-27	ModJefferson	0.797	32.77	35.53	2.76
	ModJefferson	0.804	32.63	35.36	2.73
	ModJefferson	0.708	32.86	35.97	3.10
Average 21-27	ModJefferson	0.800	32.75	35.62	2.86
27-33	ModJefferson	0.739	32.49	35.46	2.98
	ModJefferson	0.757	32.57	35.47	2.90
	ModJefferson	0.691	32.36	35.54	3.18
Average 27-33	ModJefferson	0.729	32.47	35.49	3.02
33-39	ModJefferson	0.569	31.49	35.35	3.86
	ModJefferson	0.604	31.62	35.25	3.64
	ModJefferson	0.551	31.51	35.50	3.99
Average 33-39	ModJefferson	0.574	31.54	35.37	3.83
39-45	ModJefferson	0.572	31.74	35.58	3.84
	ModJefferson	0.623	32.01	35.53	3.53
	ModJefferson	0.664	32.18	35.49	3.31
Average 39-45	ModJefferson	0.620	31.98	35.53	3.56

when replacing the observed truth in 2015 with an alternative truth, because the skillful range in quantiles will persist longer as well: up to above $q_{0.90}$ instead of otherwise below that. The extension of the skillful range is a general feature when doubled 2015 intensities are used. The extended skillful range is especially clear for both alternative experiments at +33 to +39 hours. However, at +15 to +21 hours, all methods have skill until the highest threshold verified and shown, except the ELR40 with reference detections.

6.6 Analysis of LR models as function of lead time

Decreasing sharpness due to increasing errors in forecasts of an important predictor in time can be demonstrated from model coefficients of LR models where that predictor is included. Sharpness of a predictor can be seen as the tightness of the part of its distribution associated with the gradient in predictand outcome, so in our case thunderstorm occurrence probability gradient. Hereto we shortly look at the no_CAPE models (Table 5), because it selects 20 times out of 21 maximum Modified Jefferson as first predictor. This can be interpreted as a high predictor consistency among lead times. However, among lead times that are not exactly 24 hours apart, strict conclusions cannot be drawn, as the predictor values for which models should lead to a constant thunderstorm probability (for example 50%) may vary during the diurnal cycle and furthermore sharpness of associated predictor can vary with the diurnal cycle.

Table 7 shows the model composition of LR models with the no CAPE experiment. What can be seen clearly see is that the model coefficients decrease when the lead time increases by 24 hours: from +3 to +9 hour forecasts to +27 to +33 hour forecasts the average decrease is 0.12 or 14% per day and from +15 to +21 hour forecasts to +39 to +45 hour forecasts it is 0.14 or 19% per day, averaged over the three final cross-validations. Comparing +9 to +15 hour forecasts with +33 to +39 hour forecasts, there are models for only two final cross-validations, since one of them picks graupel as best predictor. They have an average coefficient decrease of 0.11 or 16% per day.

This increase in model coefficients with increasing lead time indicates the widening distribution of Modified Jefferson values that are connected thunderstorm occurrence (probabilities): the difference between predictor values associated with 10% probability and those with 50% probability increase over 24 hours lead time, as shown in the last column of Table 7. The predictor value associated with 50% probability increases with time. In other words, larger instability and smaller 700 hPa dew point depression are required when the lead time increases by 24 hours, to issue the 50% probability of thunderstorms. Therefore the decrease in sharpness of forecasts with time can be seen in Table 7. Lastly, the threshold of maximum Modified Jefferson above which thunderstorms are more likely to occur than not, is almost constant among lead times: 35.0 to 35.6 with one outlier of 36.0. Note that Modified Jefferson does not include a substraction of the constant 8 in this dataset, which is commonly done.

7 Case studies

The following two cases studies are to demonstrate potential weaknesses of statistical post-processing models to predict thunderstorm occurrence well. Very low (non-zero) thunderstorm probabilities do not imply that thunderstorms cannot occur given the meteorological situation, as the forecast probability should match the observed frequency in a reliable forecasting system. The forecast probabilities issued with both statistical methods are reliable (Chapter 5). The thunderstorm probability should relate well to and be connected to meteorological variables and atmospheric dynamics/meteorological synopsis surrounding the environment of the potential thunderstorm occurrence. In principle, limitation of (synoptic) meteorological information to one, two or a few predictors could lead to imperfect estimation of probabilities due to lacking information in the assessment of the model, when the potential information is available and useful, even if in a verification with a large dataset shows that a forecasting model is generally reliable. Whether forecasts are really unreliable in certain subregions of predictor space is hard to demonstrate if these regions are poorly sampled. Two thundery cases where probability estimates were very low are studied in detail.

The previous statement implicitly favours the use of QRF with many variables. It should explicitly be stated that this does not mean that with 40 or 91 variables, all information is captured and used well by QRF. First of all errors in NWP model forecasts can deviate from the eventual outcome, affecting estimates of thunderstorm probability after post-processing. This could give very low probabilities in both LR and QRF with thunderstorms occurring. This section shows that LR is more vulnerable to misinterpreting the meteorological situation, while both methods could make errors when the model output is not realised in the real atmosphere. It is specifically noted that the cases studied here were those cases in the dataset that are at first seemingly unsuitable for thunderstorms based on the important predictors that are found in Chapter 6, with low values of maxima of MUCAPE and Modified Jefferson index.

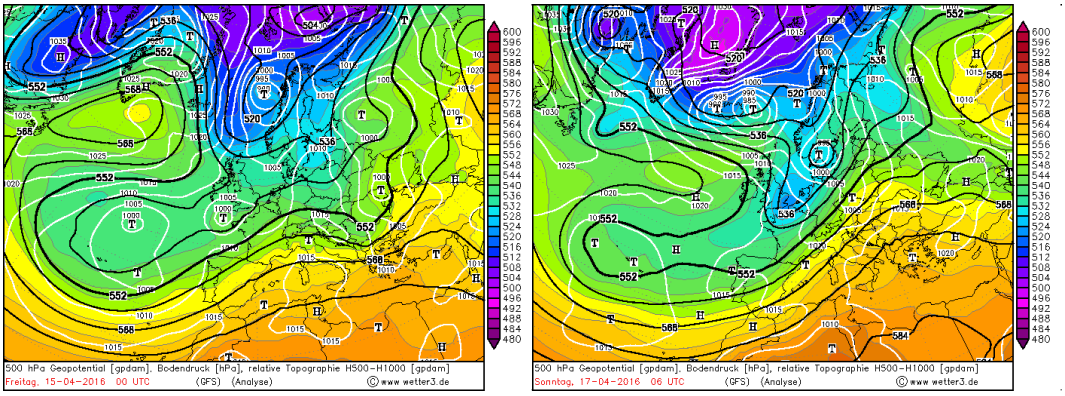
7.1 Cold air advection over North Sea and Benelux on April 17^{th} 2016

On April 14^{th} and 15^{th} a mobile ridge of high pressure southwest of Iceland moves southeastward with low pressure over the Norwegian Sea, causing a cold air mass east of Greenland to flow southward towards Scotland. Interaction with a northeastward moving baroclinic wave/low over the Channel region causes this cold air mass in the lower and mid troposphere to bend eastward towards The Netherlands in the wake of this baroclinic low pressure system on April 16^{th} and 17^{th} (Figure 25). The sea water has already warmed due to the three previous weeks which had been warmer than normal [KNMI, 2016]. Both the warm sea surface temperature and warm land temperatures lead to a conditionally unstable lower troposphere. On the front (eastern) end of the eastward moving trough, an extensive area with showers is triggered with the entrance of an area with upward motions in the lower few kilometres of the troposphere. This area passes The Netherlands from northwest to southeast during daytime. Detections of thunderstorms occur in the central regions from south to north between 09 and 15 UTC and in the central and western of the southernmost regions between 15 and 21 UTC, with in total 23 detections.

At +39 to +45 hours, LR40 gives an extremely low probability of 0.0002 for the southwestern region (10), even though four predictors are selected and used, including maximum Modified Jefferson. This is not strange, as the maximum Modified Jefferson is forecast to be 20.4 °C and maximum MUCAPE with entrainment correction forecast is 72 J/kg, with LNB's up to 3400 m. The other regions have MUCAPE values below 150 J/kg as well and the model has no graupel. Nonetheless QRF40 predicts 2% thunderstorm

Table 8: Thunderstorm occurrence probabilities issued for region 10 for April 17^{th} 2016 by QRF40 and LR40, valid between 15 and 21 UTC.

Model	**Probability (-)**	
	LT +15 to +21h	**LT +39 to +45h**
LR40	0.0019	0.0002
QRF40	0.0181	0.0186



(a) April 15^{th} $00UTC$ (b) April 17^{th} $00UTC$

Figure 25: Reanalysis of the April 2016 case by NCEP/GFS. Shown are MSLP (white contours), geopotential height of 500 hPa (black lines) and layer thickness between 500 and 1000 hPa (colours). Retrieved from [wetter3.de, nd].

probability, based on the set of 40 potential predictors.

On the day itself, low thunderstorm occurrence probabilities are still issued by the LR40 models for region 10. At lead time of +15 to +21 hours, this is caused by low values of maximum Modified Jefferson index: 25.4 °C. Maximum of MUCAPE and LNB are also low, at 113 J/kg and 3777 m. Surrounding regions have higher values for both of these important predictors in the Harmonie forecast with a maximum LNB up to about 4700 m, but maximum MUCAPE is only up to 200 J/kg in 11 regions and 245 J/kg the other region (8, the central southern region). Modified Jefferson is peaking the highest in region 8 at 30.2 °C. The neighbouring region 11 where discharges have also been detected had probabilities of 2% according to LR and 11% according to QRF40. Maximum probabilities over all KOUW-regions are 7% in region 8 with LR40 and 23% with QRF40 in the same region.

When we dive further into the best available observed vertical atmospheric profile, obtained 200 km east of the region of interest and three hours before the thunderstorms happened, a cold profile in the lower troposphere (Figure 26) can be seen that is close to neutral with respect to the state of an adiabatically lifted near-surface parcel, up to slightly above 4000 meters height. The profile is moist up to 2500 meters height. The roughly 4 km is in good agreement with the maximum level of neutral buoyancy reforecast data per region. Temperature at these levels is between -20 and -25 °C. Just ahead of the trough the atmosphere will be enriched with low level moisture, partly due to upward motions and near-surface convergence of low level winds. This could lead to conditionally unstable profiles and with small convective inhibition also to showers that likely reach just over 4 kilometres height in the vertical with top temperatures almost as low as -25 °C. Based on the experiments by [Takahashi, 1978], this condition in combination with sufficient local graupel and ice concentrations would lead to potential lightning occurrence. Based on this assessment, the probability issued by LR40 for +15 to +21 hours is very low. Situations like this one, with a low LNB_{max} of 4 or 5 km and the detection of a thunderstorm happen more often in the winter half year in The Netherlands and less frequently in the summer half year, because the convective cloud tops are usually not cold enough for lightning initiation.

Furtheremore, Modified Jefferson is strongly limited as useful predictor in this case. As can be seen in Figure 26, the much warmer air at 500 hPa would surpress high Modified Jefferson values and MUCAPE is neither high. This causes problems for logistic regression technique, which preferentially selects these predictors or predictors derived from them as discriminator for thunderstorms (Table 5). QRF can still use combination of many other variables to find out that thunderstorms could occur: importantly, when thunderstorm occurrence probability versus maximum level of neutral buoyancy and maximum Modified Jefferson index is plotted (as in Figure 10), the plot shows that the Modified Jefferson maximum values of Table 7 are always good indications for thunderstorm occurrence probability. When LNB maxima

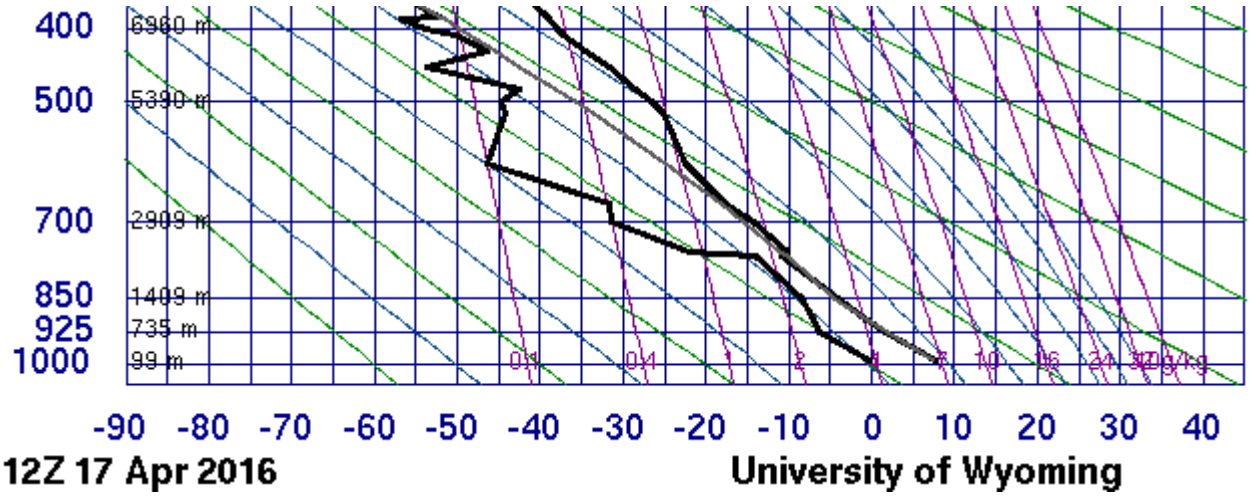


Figure 26: Radio sounding of Essen (WMO 10410) of April 17^{th} 2016, which is the nearest available. Green lines show dry adiabats, blue lines saturated adiabats and purple lines are isolines water vapour mixing ratio. The black lines show observed temperature (RHS) and dew point profile (LHS), with a grey line indicating the behaviour of a near-surface parcel after it would be adiabatically lifted. Retrieved from [University of Wyoming, nd].

are between 4000 metres and 5500 metres (5500-6000 metres is approximately at 500 hPa) it gives low empirical thunderstorm probabilities of 5% when maximum Modified Jefferson is approximately 24-32, whereas otherwise its value should be at least 32 for such thunderstorm probabilities.

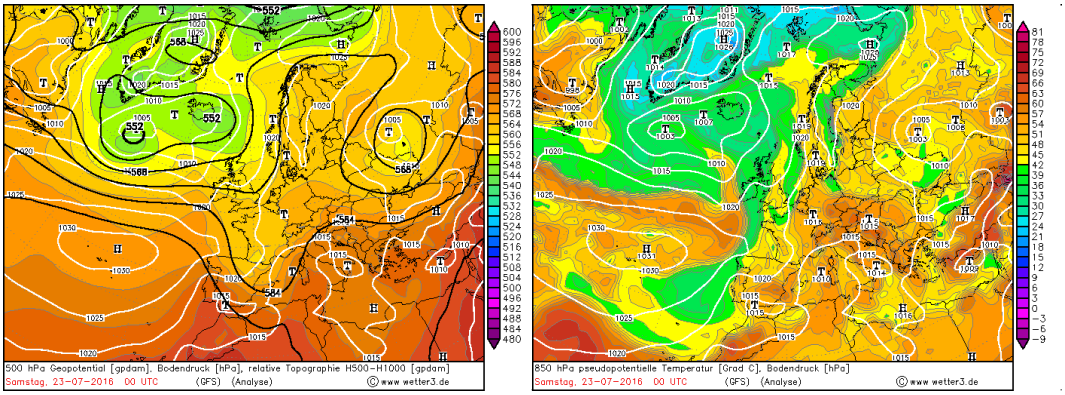
7.2 Thunderstorms in the night of July 22^{nd} to 23^{rd} 2016

On July 22^{nd} and 23^{rd} a warm air mass is located over the Benelux and Germany, previously brought in this region by a southerly flow (see Figure 27). There is a small low pressure system in the lower troposphere, oriented from WNW-ESE into southern Germany. Some low level upward motion has likely been caused by the diurnal cycle and some additional upward motion occurs in the surroundings of the low pressure system where the low level flow is convergent, though the two components are probably not unrelated. Over the west of The Netherlands lies a stationary frontal zone (large $\Theta_{w,850}$ gradient) with mild air mass west of it.

In the conditionally unstable air, thunderstorms develop during nighttime over southern regions near the Dutch-Belgian border. This is not surprising, since Harmonie forecast of 24 hours earlier produces maximum MUCAPE-values in the southeastern regions of 200-300 j/kg and maximum LNB of 6-8 km. However, in KOUW-region 7 the forecast MUCAPE is only 24 J/kg at maximum and the maximum Modified Jefferson is only 26.1 °C. Lastly, LNB does not reach above 3011 metres. Intuitively, the conditions in region 7 are not suitable for a thunderstorm, but in many other regions conditions were suitable. This is reflected by both LR40 and QRF40 predicted probabilities: 8 to 41% in the LR40 model for regions 8-12 and 18 to 50% according to QRF40. Note that there is only one forecast available, because the nearest forecast covers +21 to +27 hours and Harmonie runs only 48 hours ahead. Regions 7, 8, 10 and 11 turn out to get thunderstorms during this night, with 7 lightning detections in region 7 and 843 in total. For region 7, probabilities issued are 0.09% by LR40 and 0.99% by QRF40. The small but non-marginal thunderstorm probability is reflected by other predictors than those earlier in this section, such as the minimum Adedokun2 Index in that region: -0.6 °C. Near zero and positive values of this predictor indicate favourable thunderstorm conditions (see also [Haklander and Van Delden, 2003]).

7.3 Concluding summary of two cases

Some predictors can be very sensitive to minor forecasting errors in both space and time and atmospheric non-linearities. As illustrated with Section 7.1, many convective indices may miss relevant information in



(a) see Figure 25 for meaning of colours/lines

(b) Pseudo-$\Theta_{e,850}$ (indication for moisture/temperature at 850 hPa like $\Theta_{w,850}$) in colours and MSLP in white contours

Figure 27: Reanalysis of July 23rd 2016 00z by NCEP/GFS. Retrieved from [wetter3.de, nd].

specific cases due to their common use of 500 hPa temperature and if an inversion or very stable (and potentially neither unstable) layer is slightly below this level, many predictors could miss the information indicating a potential thunderstorm occurrence. On the other hand, CAPE calculations can be sensitive to the layer (for example layer depth) and entrainment assumptions, if entrainment is applied. Furthermore, graupel and snow inform a post-processing model about precipitation and cloud contents and can be useful in particular cases like Adedokun2 Index is in Section 7.3, but these may not or barely pop-up as affecting the large sample scores.

Although the use of the 500 hPa values of variables as temperature or Θ_w is identified as a limitation (sometimes), the 500 hPa level is likely close to an optimum: by making a Modified Jefferson index with 600 hPa temperature for example, almost all of the individual thunderstorm cases will have high values for this index. The case of Section 25 would have a large value for this index, but additional non-thunderstorm cases with an inversion or very stable layer between 500 and 600 hPa can be included as suspected potential thunderstorm events based on the new index with 600 hPa temperature, while the cloud top temperature is in fact too warm for thunderstorm occurrence. Based on Figure 10, most thunderstorm cases happen when LNB_{max} is above 500 hPa (5.5-6.0 km) and the exploitation of this level is probably close to an optimum for thunderstorm occurrence forecasts. Some comprehensive testing of a (good) convective index with various pressure levels (such as 600, 550, 500, 450 hPa) may provide more insight. In addition, MUCAPE from specific temperature layer(s) associated with lightning intiation by [Takahashi, 1978] can provide extra insight. Such optimisation of individual predictors derived with Harmonie output is beyond the scope of this study.

The two case studies show that the combination of many predictors gives additional information, which sometimes may prevent seemingly conservative estimates of thunderstorm probabilities. In addition, it helps improve representation of thunderstorm cases in the training dataset for forecasting models, which can be achieved with a longer time series.

8 Discussion

8.1 Performance of QRF compared to (E)LR

8.1.1 Thunderstorm occurrence forecasts

Quantile regression forests (QRF) is explored as a technique to forecast (severe) thunderstorms over The Netherlands by comparing them to (extended) logistic regression (LR) models and in that sense findings complement those by [Whan and Schmeits, 2018]. QRF is found to be valuable in the context of probabilistic thunderstorm forecasts and is particularly favourable, though it is found that careful pre-selection of predictors upon physical arguments may help leading to improved forecasts that are not obtained when applying the stepwise elimination algorithm that is used often in statistical studies (see [Gregorutti et al., 2016]). This is demonstrated with 40 and 90 potential predictors leading to better BSS than 228 potential predictors in an initial cross-validation experiment (Figure 14). Though a saturation of forecasting skill with increasing number of predictors when applying QRF is identified, our experiments do not always show that using all predictors gives the best result. As far as we are aware, aforementioned two findings are not shown in any meteorological context with QRF. Despite that these general features are identified in the results, they cannot be made robust with 95% confidence intervals, because uncertainty in BSS is too large. The saturation behaviour is largely consistent with results obtained in the statistical literature by [Gregorutti et al., 2016]; their results are more robust though, because they apply various statistical methods for variable selection/elimination and estimate confidence intervals during eliminations of predictors.

It is important to note that for thunderstorm occurrence LR can objectively give similarly good results as QRF when a dataset of few potential predictors is used for both methods; this is shown with a dataset of 4 potential predictors and based on best estimate of BSS (Figure 16). Additionally, LR does not stably profit from increasing the number of potential predictors; a reason is that with the initial cross-validation only a few LR predictors will effectively be selected, even if the potential predictor set is extended with many predictors. Both findings are consistent with [Whan and Schmeits, 2018] when they compare use of one predictor (Harmonie precipitation) with a potential predictor set of 41 predictors for probabilistic precipitation forecasts using ELR, QRF and a third method. A reason for enhanced benefits of many predictors in QRF is that random subsamples are used to build many random trees in QRF and not just one mathematical function with all the training data at once as with LR, likely increasing robustness of the QRF model and also smoothing probabilistic forecasts in predictor space with high weight of near neighbours in predictor space (especially near neighbours in important predictors).

The saturation of skill scores we find will likely occur in any forecasting problem with appropriate use of the information contained in the data due to the chaotic nature of the atmosphere that results in uncertainty in future atmospheric state, even if conditions that will lead to thunderstorms would be understood perfectly. The latter is very unlikely to be reached.

8.1.2 Lightning intensity forecasts

For lightning intensity predictions, QRF does not outperform ELR based on BSS (Figure 17), but CRPSS indicates better performance by QRF than by ELR (Table 4, in accordance with reliability diagrams (Figure 18 and Appendix C). It is found that QRF sometimes is able to issue high probabilities and still be reliable, where the probabilities issued by ELR frequently show an exponential decay leading to medium and high probabilities not or hardly being issued. On the other hand, ELR has the ability to profit from the imposed transformation of lightning intensities by extrapolating to higher intensities than present in the training set, but may simultaneously suffer from assumptions on the forecast distribution. The assumption is that ELR always issues a PDF of the same shape in transformed lightning intensity space, which is a reason that high intensities are hardly getting medium or high probabilities. The advantage of QRF is that it is purely empirical and does not assume a shape for forecast intensity PDF. A similar reliability feature can also be found in the reliability diagrams of 20 mm per hour precipitation forecasts for several lead times in [Whan and Schmeits, 2018] when comparing ELR and QRF. Note that the lightning

intensity probabilities are conditional, in contrary to those by [Whan and Schmeits, 2018]; unconditional probabilities are only equal to conditional probabilities when the thunderstorm occurrence probability is 100%.

If the unconditional quantiles (includes non-thunderstorm cases) up to which we can make skillful lightning intensity forecasts are computed based on 95% uncertainty in the BSS, it ranges up to about $q_{0.994}$ in the first 33 hours for both QRF and ELR (Section 5.4.5), which varies from 100 discharges per 5 minutes between 03 and 15 UTC to almost 400 discharges per 5 minutes between 15 and 21 UTC. This quantile is generally similar or slightly higher than in the study by [Whan and Schmeits, 2018] in the first 24 hours, where unconditional predictions were compared to climatology using a similar sample size. However, the upper range of skillful quantiles is not directly comparable: as it has been demonstrated, accuracy and homogeneity of the predictand dataset (lightning detections) is important for the upper skillful threshold. It would be interesting to conduct an experiment where conditional probability may extend the skillful range of thresholds compared to an equivalent non-conditional probability forecast when using the same potential predictor set and predictand: for example with rainfall forecasts as in [Whan and Schmeits, 2018] by training conditional high intensity forecasts with a subset of a dataset with samples exceeding a precipitation threshold of 2 or 3 mm/h (eliminating most stratiform precipitation but not the convective cases that are of main interest), using for example ELR and QRF. Another useful experiment would vary the composition of the training set, such that one selects only a subset to make the relative frequency that a high threshold is exceeded higher in the training set.

8.2 Exploitation of complementary predictors for deep convection forecasts

8.2.1 Potential of complementary predictors for thunderstorm occurrence

Shortcomings of individual thunderstorm predictors have been explained and it has been demonstrated how use of many complementary predictors from NWP output is beneficial for improving short term (+3 to +45 hours) forecasts of deep convection, when attempting to discriminate thundery conditions from non-thundery conditions, even if some are strongly correlated. BSS increases when adding meaningful predictors. A detailed discussion on individual predictors is found in Section 8.3.

Many preceding studies have also shown that use of complementary predictors is beneficial when aiming to forecast deep convection, by [Doswell et al., 1996] by describing the ingredient based approach thoroughly and additionally by (many) others, such as [Schmeits et al., 2005, 2008], [Van Zomeren and Van Delden, 2007], [Púčik et al., 2015], [Ahijevych et al., 2016] and [Whan and Schmeits, 2018]; the latter two show how combining many predictors can be beneficial with (Q)RF. The statistical post-processing approach can also be used fruitfully to gain physical and forecasting insights (we have tried to illustrate this). Therefore, it is encouraged that operational forecasters, severe weather researchers and post-processing researchers collaborate to understand extremes from different perspectives to share physical, practical/operational and statistical insights.

8.2.2 Potential of complementary predictors for lightning intensity

For lightning intensity forecasts, the indications that many predictors can add complementary information are not strong. Limitations that may have obscured differences between methods and potential predictor sets in BSS, are the small dataset, with only 381 to 699 cases per valid time (Table 3) and inhomogeneity in lightning detections (Figure 12). Since the results can be strongly affected by the two limitations, the results that have been obtained on lightning intensity predictions are not definitively conclusive from an objective point of view. Furthermore, collecting a homogeneous lightning detection set in time on the national or continental spatial scale and decadal time scale will serve science, but observation techniques are usually progressing along with technology.

To circumvent the issue of inhomogeneity, two different lightning detection datasets have been used to investigate some uncertainty related to inhomogeneity. The results from these so-called inhomogeneity experiments show that ordering of ELR and QRF in lightning intensity forecasts may depend on it (Figure 24 and Section 6.5), which is a main reason that results are not conclusive. Furthermore, the BSS increases

typically a bit when the detections of 2015 are modified to likely an improved representation of 2015 from the climatological point of view. But lightning detection datasets are not well comparable and some individual cases could strongly deviate in detected lightning intensity between different detection datasets, making detections in essence inconvertible. Especially for QRF the improvement of BSS is a consistent feature after modifying the KLDN detections in 2015; one would probably expect improved scores compared to the reference KLDN dataset for a homogeneous lightning detection set based on these experiments, in particular for QRF. The QRF method likely profits in particular when lightning detections are more representative, because it strongly depends on representativeness of the empirical distribution in a training sample, whereas ELR assumes a distribution around the training samples for which other regions in predictor space can also be used with some extrapolation.

8.2.3 Removal experiments

While it has been shown that neither QRF nor (E)LR are highly dependent on a single type of potential predictor, good indications of consistent but small BSS improvements with CAPE and microphysics predictors are demonstrated for thunderstorm occurrence forecasts with QRF. This feature is only stably present with microphysical predictors for LR models (Figure 20). Investigation of thunderstorm occurrence in predictor planes shows that an extra predictor typically gives additional information to refine probabilistic forecasts, but the predictors together build a stable framework such that any extra predictor indeed should on average not give more than some refinement of the forecasts and have relatively larger implications in just a few cases (a small region in predictor space) within the large verification set. Since CAPE-predictors and microphysical predictors have been shown to be very important for both QRF and LR, the small refinements and slight BSS improvements for thunderstorm occurrence forecasts made by adding these very important predictor types suggest that we may be close to optimal thunderstorm occurrence forecasts that are possible with the available NWP output and that most of the variability is indeed probabilistic and resulting from forecast uncertainty, but not or hardly from missing predictor information. The role of CAPE-predictors is to some extent taken over mainly by maxima in Modified Jefferson and $\Theta_{w,850}$ in both LR and QRF models at all lead times, while other signals of CAPE and graupel/snow replacements are noisy and diffuse.

8.3 Predictors

8.3.1 Thunderstorm occurrence

Selection frequencies (LR) and predictor importance (QRF) can give useful information about the physical interpretation of predictor relevance. QRF shows that SBCAPE and its combination with SBCIN are much more important during daytime (09 to 21 UTC) than during nighttime, consistently with physical arguments, namely that convection is often driven by warm surface/boundary layer during daytime (Figure 21). Additionally, maximum Modified Jefferson is a very good thunderstorm occurrence discriminator in The Netherlands based on both LR and QRF. Part of the strong discriminative performance may be due to that Modified Jefferson takes into account both air mass instability ($1.6\Theta_{w,925} - T_{500}$) and moisture via the dew point depression at 700 hPa. The dew point depression term will typically go to zero locally when a deep convective cloud (potential lightning producer) is present in the weather model, which can lead to a local maximum in Modified Jefferson in the shower that would have been smoothed in a hydrostatic model, where all convective clouds are parameterised. In other words, in non-hydrostatic models Modified Jefferson may be relatively more favourable than in hydrostatic models.

Consistently with the no_CAPE and no_mph experiments both LR and QRF show that maximum column graupel and maximum MUCAPE as well as their product are very helpful for thunderstorm occurrence predictions. Moreover, indices of atmospheric instability in general appear to be very important for thunderstorm forecasts. The initiation of showers and air mass/moisture are less important; forcing is even lower on the list. Selection frequency and predictor importance also help to identify which predictors may substitute each other (Section 8.2.3).

In the previously made and currently operational statistical post-processing model for thunderstorm

occurrence [Schmeits et al., 2008] Jefferson appeared to be more favourable, which might be because Hirlam is hydrostatic (see previous paragraph). Despite this, Jefferson and Modified Jefferson have a strong correlation of 0.87 for their regional maxima and 0.83 for their mean values in the reforecasting dataset. The most important predictor was ECMWF convective precipitation and its predictive value is very likely replaced by column integrated graupel and snow (or their combinations with MUCAPE) in the new statistical post-processing models for thunderstorms, while CAPE (as MUCAPE) is still frequently selected in the new models; their predictive value is consistent with expectations.

We are aware of one other NWP post-processing study for thunderstorm forecasts above Europe (excluding The Netherlands) and none in a mild and humid region outside Europe. [Simon et al., 2018] use ECMWF output for probabilistic thunderstorm occurrence forecasts over the Eastern Alps for 12-18 UTC up to 5 days ahead. Their potential predictor set was strongly deviating from ours and their terrain that is largely mountainous as well. Their most influential predictor is mean relative humidity at 700 hPa, followed by temperature change at 700 hPa from 12 to 18 UTC and square root of convective precipitation. Although importance of 700 hPa relative humidity (relative humidity above 70% as favourable condition) connects to our finding that Modified Jefferson is informative, their predictor selections are generally not so comparable. CAPE is among their 9 eventually applied predictors.

The selection of MUCAPE and Modified Jefferson maximum as important predictors is consistent with regional maximum hourly precipitation QRF and ELR forecasts made by [Whan and Schmeits, 2018] based on Harmonie output from an older model version, but on the other hand maximum Fateev is far less important and as one would expect the wind speed at 500 hPa too (giving information about how long a cell will reside at a place). Instead of column graupel or column snow they use cumulative precipitation to include the presence of showers in Harmonie, which is of course more appropriate when forecasting regional maximum precipitation.

Other thunderstorm occurrence models or forecasting evaluations are typically based on radio sounding data. [Haklander and Van Delden, 2003] compared 6-hourly radio soundings from De Bilt with thunderstorm occurrence in a 100 km radius from De Bilt and have found that Lifted Index performs as best discriminator for whether a thunderstorm occurs. From the common predictors that we have studied, Boyden Index, Adedokun2 Index and MUCAPE are also good discriminators in their study. Modified Jefferson is called Jefferson evaluation (with almost the same definition, but Θ_w at 850 hPa instead of 925 hPa as we use) and has intermediate performance. Several aspects make it hard to compare both studies: proximity soundings may be as far as 100 km and 6 hours away from a thunderstorm, which gives a very limited description of the state of the atmosphere in and around the thunderstorm. Mainly instability and larger scale moisture would correlate on larger spatial scales. The Harmonie reforecasting dataset is much more detailed at 2.5 by 2.5 km resolution and hourly time steps, such that local extremes in both the pre-convective and convective environment are included, whereas sounding information is very incomplete.

8.3.2 Lightning intensity

With predictor selection frequencies in ELR and perdictor importances in QRF, important conditional thunderstorm intensity predictors have been identified. Bradbury index (regional minimum) and K-index (regional maximum) are typically selected as good predictors for conditional lightning intensity forecasts (Table 6). They are strongly correlated, especially if a thunderstorm occurs. Additionally, combining Bradbury index and Boyden index with regional precipitable water values as was done using their regional maxima works well. The conditional lightning intensity forecasts are not depending on these precipitable water-instability combinations and neither on CAPE nor on microphysical predictors (Figure 23). Besides this it is found that warm air mass with high Θ_w in the lower troposphere and boundary layer and additionally large instability are favourable for high lightning intensities, which is consistent with the parameterisation by [Lopez, 2016] and consistent with the expectation that cases as the Spanish plume produce particularly high lightning intensities in The Netherlands. Evaluation of the models shows that other information appears to be hardly used for lightning intensity forecasts.

In addition, Bradbury index was also the most frequently selected predictor in the previous severe thun-

derstorm post-processing model based on Hirlam, ECMWF and the KNMI precipitation radar [Schmeits et al., 2008]; this suggests that hydrostacy of the model does not affect the optimal lightning intensity predictor. In a wider (forecasting/meteorological) perspective, it can be said that the mid-level lapse rate in Θ_w has a good relationship with lightning intensity.

Few other similar studies to predict high lightning intensities have been carried out, but the study by [Ahijevych et al., 2016] to predict mesoscale convective system (MCS) initiation is comparable to this study: firstly because they use random forest (RF), furthermore they apply predictor ranking like it is also done here (Chapter 6) and they show what happens to predictions when removing a specific type of information (see also Section 8.4). Important predictors they find (see also Chapter 1) are terrain information, precipitable water and radar reflectivity. In our study precipitable water is among the most informative predictors in QRF (Appendix E) and LR via its combination with the Bradbury Index, but they do not use many instability indices such as Bradbury index. On the other hand their predictand is column integrated liquid water, which is likely to be strongly related to PW.

8.3.3 Limitation of our predictors and dataset

An important limitation of Modified Jefferson and many other convective indices is that they assess only the 850 or 925 to 500 hPa temperature (and also often moisture availability; exception is the Boyden index), but no shallower layer. Showers of 4-5 km deep can also lead to thunder as shown in Figure 10 and Chapter 7). Sometimes this can lead to probabilistic estimates that are very likely poor, especially with LR. This can be improved by combining information from LNB and other predictors such as $\Theta_{w,850,max}$. But probably, using a more direct predictor associated with lightning could be beneficial, in particular convective cloud top temperature in the NWP models. This can involve masking convective regions using for example vertical velocity or MUCAPE of a layer with a specific temperature.

A predictor that might be useful to improve is moisture convergence. Regional statistics based on individual grid cells in Harmonie have been used, although larger spatial structures may contain more information than these 2.5 by 2.5 km grid cells; importance measures from QRF and selection by LR do not indicate it as important. [Van Zomeren and Van Delden, 2007] have found it as useful thunderstorm predictor at 100 km resolution, but nonetheless suggest improvement with a higher resolution moisture convergence field than they used. Moisture convergence might reveal valuable information on spatial scales between 2.5 and 100 km. Predictors such as graupel which indicate the actual presence of showers may diminish the additive value that moisture convergence could have. A second predictor that could have been improved to a potentially more valuable one is the shear, by covering the layer below 850 hPa as well. Lastly, it would be interesting to see whether latest probabilities from the previous run for the same valid time could be beneficial as potential predictor.

Some extension of this work can be made by comparing predictor planes with the winter Harmonie40 reforecast dataset for some important predictors found in this study, such as maximum values of Modified Jefferson, level of neutral buoyancy, MUCAPE, column graupel. Besides this deriving statistical post-processing models with QRF and/or LR for the winter could be interesting and also a specific study of aircraft induced lightning (AIL). It would be expected that such a winter lightning dataset gives similar information as Figure 10 if it would be extrapolated towards a lower $\Theta_{w,850,max}$ climatology. This means that one would expect thunderstorm probabilities increasing from near-zero with LNB_{max} of about 4000 m to high probabilities (0.5-1.0) when LNB_{max} increases to 7000-8000 metres. In the regions close to Schiphol Airport where AIL thunderstorms can occur frequently, the probability increase might be located at even lower LNB_{max} values and Modified Jefferson is expected to be a poorer predictor due to the potential ability of relatively shallow showers to produce lightning and especially AIL.

8.4 Application in nowcasting-forecasting continuum

Other studies that use statistical tools and machine learning to improve forecasts of deep convection on the very short term (up to 6 or 12 hours or even shorter) combine real time information, such as radar imagery, satellite images, NWP output and lightning detection. This has been discussed in Chapter 1.

One of the reasons that satellite and radar imagery is especially useful for the first hours, is that NWP output is typically only available every 3, 6 or 12 hours, such that NWP output can become outdated compared to satellite and radar imagery. Meteorologists "nowcast" by smartly combining observations and model output to extrapolate observations for the first hours. The current study uses NWP output that may become outdated for the first lead time, the +3 to +9 hours forecast (during its availability).

For the currently operational version of the KNMI post-processing model for thunderstorms, by [Schmeits et al., 2008], a smart "radar and lightning detection advection" scheme based on Hirlam output was applied. RF has been used for convective initiation and MCS initiation nowcasts in the United States of America by [Mecikalski et al., 2015] (see Chapter 1) and [Ahijevych et al., 2016]. The extensive set of experiments in the latter study demonstrates that especially extrapolated radar has added value in optimising their forecasts and the former concludes that satellite and NWP output are complementary forecasting convective initiation in the first hour. Therefore a very interesting extension of our study could be to expand the QRF models by including satellite and radar information together with the NWP output to search for potential improvement of the thunderstorm forecasts in the first hours. In the end, such a dataset could make it possible to improve predictions in the nowcasting-forecasting continuum. In some cases nowcasting might work 6 hours ahead while sometimes it might not work. If QRF can find out when it should use which information, the first day could be optimised with some nowcasting data. Additionally, ensemble information about showers and convection from few important predictors such as column graupel may be helpful, such that QRF can compare NWP ensemble predictors with observations. Such a study would require a very extensive potential predictor exploration and selection, because radar and satellite imagery and NWP ensemble output could produce many additional predictors.

Furthermore, it might be interesting to build extreme weather post-processing models on sub-European scale, using a larger dataset of high lightning intensities. A spatial limitation in such a study could be the extension of domains of compatible operational Harmonie models, or other non-hydrostatic NWP models.

9 Conclusions

Probabilistic models for thunderstorm occurrence and conditional lightning intensity with Harmonie derived predictors have been made and evaluated. Based on model verification, experiments, case studies and the discussion, we can draw the conclusions below.

Application of QRF for thunderstorm forecasts

It is found that quantile regression forests (QRF) is a beneficial technique to post-process numerical weather prediction model output for thunderstorm occurrence and intensity forecasts at the short term. Furthermore, it is demonstrated that large sets with complementary predictors are useful to improve thunderstorm forecasts, even if some predictors are strongly correlated. The stepwise elimination strategy applied to QRF, which is regularly applied in statistical studies, will not lead to an optimal predictor selection, but properly merging information from machine learning and physical understanding can improve statistical post-processing models.

Performance of QRF and (extended) logistic regression

Based on Brier skill score (BSS), QRF systematically performs better than logistic regression (LR) for probabilistic forecasts of thunderstorm occurrence. For lightning intensity, fitting probabilistic post-processing models is likely complicated by inhomogeneity in lightning detection dataset KLDN, although significant forecast skill does not critically depend on strict homogeneity for the lower verified thresholds. The latter holds for both extended LR and QRF.

Predictors

CAPE is not irreplaceable as thunderstorm occurrence predictor and direct predictors from the microphysics scheme neither are. Nonetheless, removing them as potential predictors for thunderstorm occurrence leads almost exclusively to degradation of BSS. New predictors containing precipitable water and an instability index that have been constructed (see Section 2.4.2) are neither essential for the thunderstorm forecasts. Among the various convective indices that have been developed, Modified Jefferson is a very skillful and easily computable predictor to isolate (potential) thunderstorm situations from non-thunderstorm situations over The Netherlands using non-hydrostatic NWP output; it has an important limitation when regional maximum LNB is slightly below 500 hPa. Similarly, extremes in K-index (maximum) and Bradbury index (minimum) derived from non-hydrostatic NWP forecasts are both good indicators for (conditional) lightning intensity. High lightning intensity forecasts almost exclusively depend on information from instability indices and indicators of a warm, moist air mass.

A Table of potential predictors

The table below provides a description of the potential predictor sets that have been used. The set of lowest number of potential predictors in which the specified potential predictor is included, is given in the last column of the table. This means that the 15 set consists of all potential predictors with 15 as indication and those with 4 as indication and the 40 set consists of those with 4, 15 and 40 as indication. For the selection of 25 extra predictors besides the elementary predictors (Section 2.2.6) when making the 40 potential predictor set, the QRF assessment of predictor permutation importance, selection frequency and elimination order have been used to analyse the model fits with the 91 potential predictor set. This selection step is partially arbitrary.

The "_" with subsequent statistical measure (max, min, mean, $q_{0.98}$, $q_{0.90}$, $q_{0.50}$, $q_{0.10}$ and $q_{0.02}$) refers to the respective statistical measure of Harmonie output over 6 hours and within the whole KOUW region of a variable that is computed for the predictor. When instead of these _delta is used, it refers to the difference between the maximum and minimum within 6 hours and one region. Furthermore, a name starting with Trans_ refers to standard variables on which a power transformation has been applied and _pows refers to a composite to which a transformation has been applied. Lastly, q0.98_q0.02 stands for the difference between $q_{0.98}$ and $q_{0.02}$ in a spatial box and a time bin.

Predictors investigated and pointed out as new potential predictors either due to more detailed representation of microphysics in Harmonie than in Hirlam and ECMWF (Section 2.1) or in the search of new potential information (Section 2.4.2) have been printed in bold in Table A-i.

Table A-i: Potential predictors in QRF91 and (E)LR91.

Predictor name	**Varable equation or full description**	**In set (min. size)**
MSLP_max	Mean sea level pressure (MSLP)	4
MUCAPE_max	$\int_{LFC}^{LNB} g\frac{T_{v,p}-T_{v,env}}{T_{v,env}}dz$ see Section 2.3	4
PrecipWater_max	$\int_{Surf}^{TOA} \rho q_v dz$	4
Trans_Graupel_col_max	$\left\{\int_{Surf}^{TOA} \rho q_{graupel} dz\right\}^{\frac{1}{5}}$	4
$\frac{dp}{dt}$	Region mean of 6 hour MSLP tendency	15
Bradbury_min	$\Theta_{w,850} - \Theta_{w,500}$	15
Bulk_shear_850_500_mean	$\sqrt{(u_{850}-u_{500})^2+(v_{850}-v_{500})^2}$	15
coast	Indicator that equals 1 in coastal KOUW-regions (1-5 & 7) and 0 in others (6 & 8-12)	15
Helicity_max	$\int_{Surf}^{z_{700}} \vec{k}\times(\vec{v}-\vec{c})\times\frac{d\vec{v}}{dz}dz$	15
Helicity_min		
LNB_max	Maximum level where $\Theta_{v,p}=\Theta_{v,env}$ see Section 2.3	15

MUCIN_mean	$\int^{LFC} g\frac{\Theta_{v,p}-\Theta_{v,env}}{\Theta_{v,env}}dz$ where buoyancy is negative; see Section 2.3	15
ThetaW_850_max	Potential wet bulb temperature 850 hPa ($\Theta_{w,850}$)	15
Trans_Moisture_convergence_max	$\left\{\int_{Surf}^{TOA}\left(\frac{\partial\rho u q_v}{\partial x}+\frac{\partial\rho v q_v}{\partial y}\right)dz\right\}^{\frac{1}{7}}$	15
Trans_Moisture_convergence_min		
$(\sqrt{MUCAPE+MUCIN})_{max}$	Square root of sum of convective available potential energy and convective inhibition based on most unstable parcel	40
$(\sqrt{SBCAPE+SBCIN})_{max}$	Square root of sum of convective available potential energy and convective inhibition based on surface based parcel	40
Adedokun1Index_neg_min	$\Theta_{s,500}-\Theta_{w,850}$ (850 hPa originating parcel, no entrainment)	40
Adedokun2Index_negMU_min	$\Theta_{s,500}-\Theta_{w,mu}$ (most unstable layer, no entrainment) In the original script, its negative value was provided.	40
Boyden_max	$0.1(z_{700}-z_{1000})-T_{700}-200$	40
Boyden_PW1	$(Boyden_{max}-85)log(PrecipWater_{max})$	40
Boyden_PW2	$(Boyden_{max}-85)PrecipWater_{max}$	40
Bradbury_925_min	$\Theta_{w,925}-\Theta_{w,500}$	40
Bradbury_PW1	$(Bradbury_{min}-14)log(PrecipWater_{max})$	40
Bradbury_PW2	$(Bradbury_{min}-14)PrecipWater_{max}$	40
DPT_700_max	Dew point temperature at 700 hPa ($T_{d,700}$)	40
DPT_850_max	Dew point temperature at 850 hPa ($T_{d,850}$)	40
Edward_PW1	$(\Theta_{w,925,max}-\Theta_{w,500,min})log(PrecipWater_{max})$	40
Edward_PW2	$(\Theta_{w,925,max}-\Theta_{w,500,min})PrecipWater_{max}$	40
Fateev_max	$T_{850}-T_{500}-DD_{850}-DD_{700}-DD_{600}-DD_{500}$	40
Jefferson_max	$1.6\Theta_{w,925}-T_{500}-11$	40
K_index_max	$T_{850}-T_{500}+T_{d,850}-DD_{700}$	40
LNB_q0.90	See LNB_max	40
ModJefferson_max	$1.6\Theta_{w,925}-T_{500}-\frac{1}{2}DD_{700}$ Note that Modified Jefferson does not include a substraction of the constant 8 in this dataset, which is commonly done.	40
MUCAPE_q0.90	See MUCAPE_max	40
MUCAPE_snow_pows_max	$\left\{\int_{LFC}^{LNB} g\frac{T_{v,p}-T_{v,env}}{T_{v,env}}dz\right\}^{\frac{1}{4}}\left\{\int_{Surf}^{TOA}\rho q_{snow}dz\right\}^{\frac{1}{6}}$	40
MUCAPE_graupel_pows_max	$\left\{\int_{LFC}^{LNB} g\frac{T_{v,p}-T_{v,env}}{T_{v,env}}dz\right\}^{\frac{1}{4}}\left\{\int_{Surf}^{TOA}\rho q_{graupel}dz\right\}^{\frac{1}{10}}$	40

SBCAPE_max	Parcel starts at T_{2m} and $T_{d,2m}$ furthermore see MUCAPE_max and Section 2.3	40
ThetaW_925_max	Potential wet bulb temperature 925 hPa	40
Trans_Snow_max	$\left\{\int_{Surf}^{TOA} \rho q_{snow} dz\right\}^{\frac{1}{3}}$	40
$\frac{dBoyden}{dt}$	Region mean of 6 hour tendency in Boyden index	91
Bulk_shear_850_500_max	see Bulk_shear_850_500_mean	91
Bulk_shear_850_700_max	$\sqrt{(u_{850} - u_{700})^2 + (v_{850} - v_{700})^2}$	91
Bulk_shear_850_700_mean		91
Cloud_base_max	Height of cloud base (lowest point where grid cell cloud cover exceeds $\frac{5}{8}$)	91
Cloud_layers_depth_delta	Difference between cloud base and cloud top	91
Cloud_layers_depth_max		91
Cloud_top_max	Height of cloud top (heighest point where grid cell cloud cover exceeds $\frac{5}{8}$)	91
Cross_Totals_max	$T_{d,850} - T_{500}$	91
MSLP_delta	See MSLP_max	91
DPT_500_max	Dew point temperature at 500 hPa ($T_{d,500}$)	91
DPT_600_max	Dew point temperature at 600 hPa ($T_{d,600}$)	91
LFC_mean	Level where $\Theta_{v,p} = \Theta_{v,env}$ below CAPE layer see Section 2.3	91
Lid_Strength_mean	Vertical maximum of $\Theta_{w,s} - \Theta_w$ for most unstable parcel, below 500 hPa see Section 2.3	91
LNB_mean	See LNB_max	91
mean_cloud_cover	Fraction of grid cells with defined cloud base within KOUW-region	91
ModJefferson_q0.90	See ModJefferson_max	91
Moisture_convergence_q0.98_q0.02	Difference between 0.98 and 0.02 quantile of moisture convergence (divergence) without power transformation of $\frac{1}{7}$ see Trans_Moisture_convergence_max	91
MUCAPE_water_pows_max	$\left\{\int_{LFC}^{LNB} g\frac{\Theta_{v,p} - \Theta_{v,env}}{\Theta_{v,env}} dz\right\}^{\frac{1}{4}} \left\{\int_{Surface}^{TOA} \rho q_l dz\right\}^{\frac{1}{10}}$	91
MUCAPE_q0.98	See MUCAPE_max	91
MUCIN_min	See MUCIN_mean	91
Rackliff_min	$\Theta_{w,925} - T_{500}$	91
Rain_acc_max	Accumulated hourly rainfall at ground level	91
SBCIN_max	Parcel starts at T_{2m} and $T_{d,2m}$ furthermore see SBCAPE_max and Section 2.3	91
sin_dd_dif_500_850_q0.90	Sine of difference in wind direction between 500 and 850 hPa	91
Storm_Travel_max	See [Whan and Schmeits, 2018]	91

SWEAT_max_pows0.2	Severe weather thread index and its power transform at power $\frac{1}{5}$ $SWEAT = 12T_{d,850} + 20(TotalsTotals - 49) + 3.88(\sqrt{u_{850}^2 + v_{850}^2}) + 3.88(\sqrt{u_{500}^2 + v_{500}^2}) + 125x$ where x is the sine of difference in wind direction between 500 and 850 hPa	91
SWEAT_mean		91
ThetaW_500_max	Potential wet bulb temperature 500 hPa	91
ThetaW_925_850_diff_max	$\Theta_{w,925} - \Theta_{w,850}$	91
ThetaWs_500_max	Saturated potential wet bulb temperature 500 hPa	91
Totals_Totals_max	$T_{850} + T_{d,850} - 2T_{500}$	91
TQ_max	$T_{850} + T_{d,850} - 1.7T_{700}$	91
Trans_Cloud_ice_max	$\left\{\int_{Surface}^{TOA} \rho q_s dz\right\}^{\frac{1}{2}}$	91
Trans_Cloud_water_q0.98	$\left\{\int_{Surface}^{TOA} \rho q_l dz\right\}^{\frac{1}{5}}$	91
Trans_Rain_max	$\left\{\int_{Surface}^{TOA} \rho q_{rain} dz\right\}^{\frac{1}{3}}$	91
U_500_delta	u_{500} (u-component of 500 hPa wind speed)	91
U_500_min		91
U_700_min	u_{700}	91
U_850_delta	u_{850}	91
U_850_min		91
V_500_delta	v_{500}	91
V_500_min		91
V_700_mean	v_{700}	91
V_850_delta	v_{850}	91
V_850_max		91
Vertical_Totals_max	$T_{850} - T_{500}$	91
WSPD_500_delta	Wind speed at 500 hPa	91
WSPD_500_mean		91
WSPD_850_delta	Wind speed at 850 hPa	91
WSPD_850_min		91

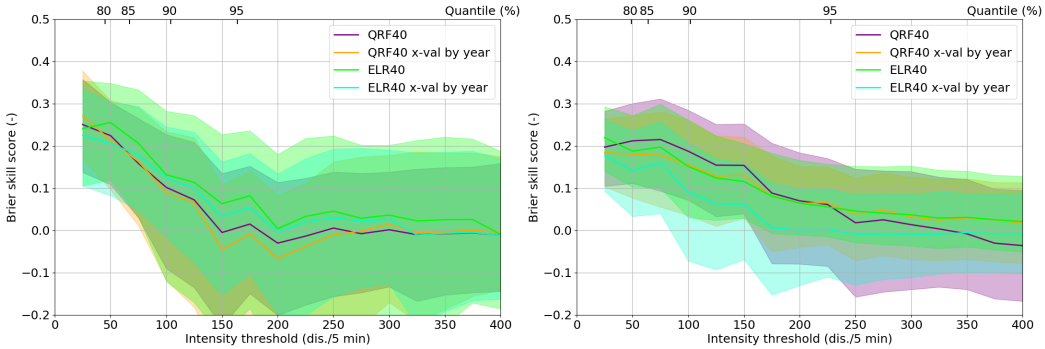
In the table above, some symbols used are not yet defined. Their definition is given in Table A-ii.

Table A-ii: Table explaining the variables that have not yet been explained in Table A-i

Symbol	**Value (if constant)**	**Unit**	**Description**
Θ_v		K, °C	Virtual potential temperature
$_p$			Parcel value
$_{env}$			Environmental value
g	9.81	m/s^2	Acceleration by gravity
z		m	Geopotential height
TOA			Top of the atmosphere in model
$Surf$		m	Surface of atmosphere in model
ρ		kg/m^3	Air density
q_v		g/kg	Specific humidity
$q_{graupel}$		g/kg	Specific mass ratio of graupel to moist air
$_{lvl}$		hPa	Value at given vertical pressure coordinate
u		m/s	Wind speed, zonal component
v		m/s	Wind speed, meridional component
$\vec{k}$			Unit vector in the vertical direction
$\vec{v}$		m/s	Wind speed as vector
$\vec{c}$		m/s	Assumed storm motion vector
T		K, °C	Temperature
DD		K, °C	Dew point depression, $T - T_d$
q_{snow}		g/kg	Specific mass ratio of snow to moist air
q_l		g/kg	Specific mass ratio of cloud liquid water to moist air
q_s		g/kg	Specific mass ratio of cloud ice water to moist air
q_{rain}		g/kg	Specific mass ratio of rain to moist air

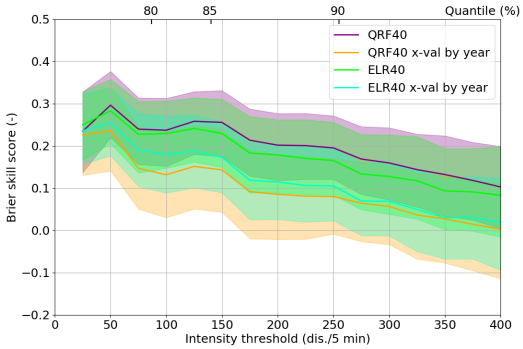
B Dependence on cross-validation strategy of lightning intensity forecasts verification score

The results for two different cross-validation strategies for lightning intensity forecasts are shown in Figure B-i. The two cross-validation strategies are firstly training on two years and testing on the third year ("x-val by year") and secondly generating three random subsets on two of which training is done from, with model testing on the third random sample. The random strategy is selected for lightning intensity forecasts. Typically, the scores are better when applying a random strategy. The random strategy is also preferred, because the KLDN detections are found to be inhomogeneous (Section 4.4). However, for the +21 to +27 hours forecasts, the verification scores with random cross-validation are not better for ELR40. Additionally, some of the verification scores are also very close to each other, such as low thresholds at +39 to +45 hour forecasts. Furthermore, a really big improvement is seen for QRF40 at +15 to +21 hour forecasts and +21 to +27 hour forecasts. At +15 to +21 hours, all shown thresholds verified are skillfully forecast with QRF40 and the random cross-validation strategy.

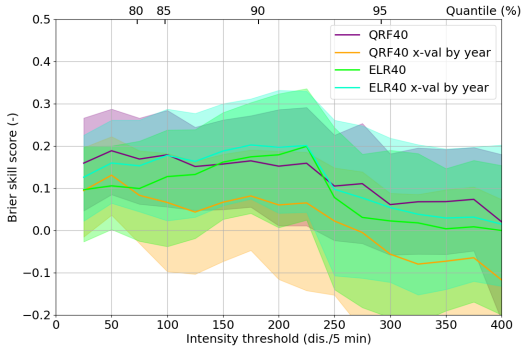


(a) Harmonie00z +3 to +9 hours

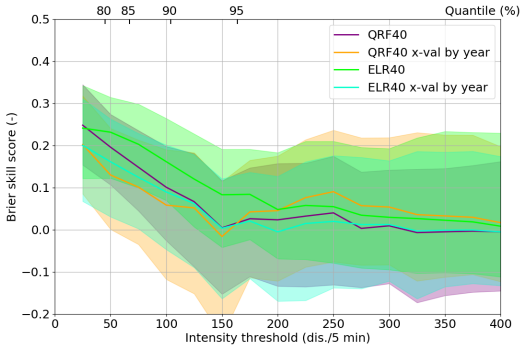
(b) Harmonie00z +9 to +15 hours



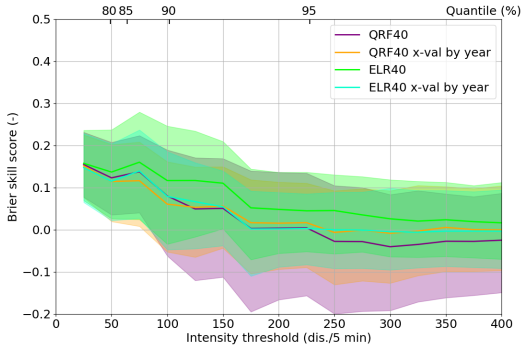
(c) Harmonie00z +15 to +21 hours



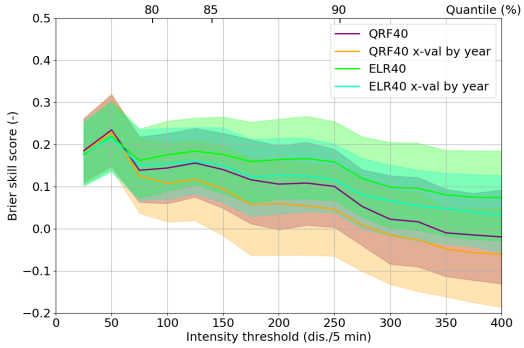
(d) Harmonie00z +21 to +27 hours



(e) Harmonie00z +27 to +33 hours



(f) Harmonie00z +33 to +39 hours



(g) Harmonie00z +39 to +45 hours

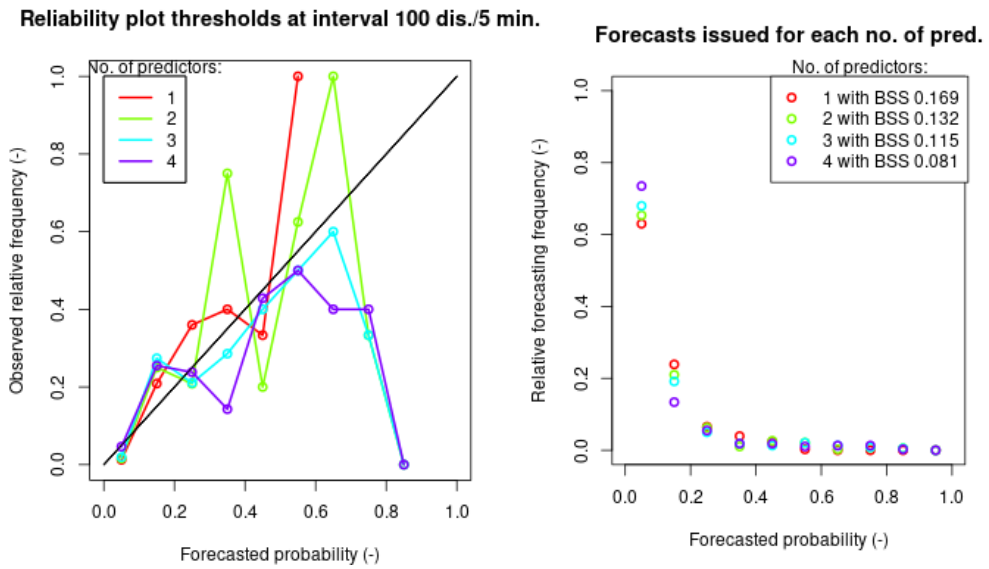
Figure B-i: Brier skill score of ELR40 and QRF40 as a function of lightning intensity with two cross-validation strategies: one with verification by year and one with three randomly generated verification sets. The four highest training quantiles are also given at the top (if within axis limits).

C Reliability diagrams of QRF40 & ELR40

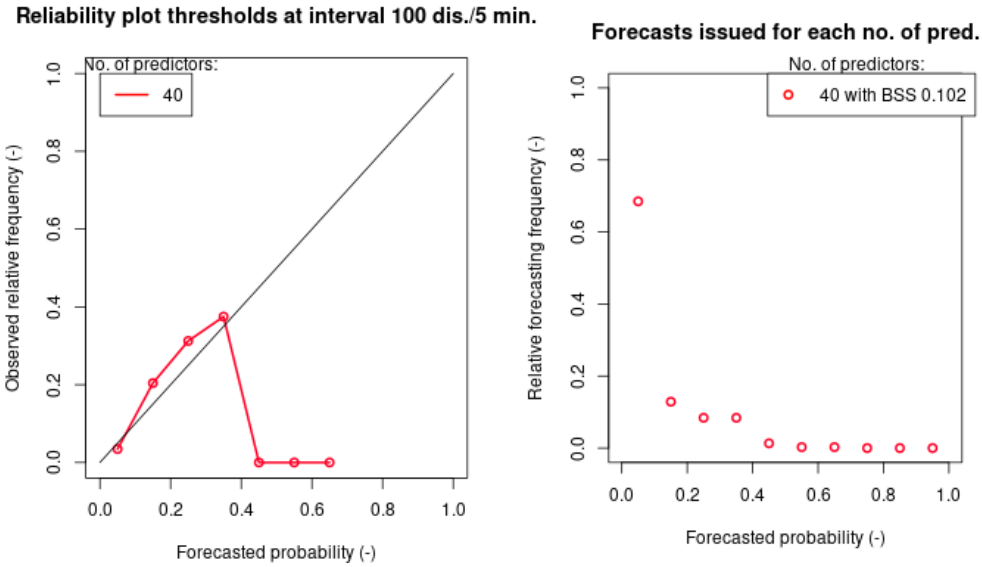
This appendix shows some additional reliability diagrams for QRF40 and ELR40 to illustrate the difference between probabilities issued by both methods and their reliability. In all cases we look at $q_{0.90}$ of the conditional lightning intensity distribution for at all lead times; in the unconditional lightning intensity distribution (cases without thunderstorm included), these translate to $q_{0.994}$ for 21 to 9 UTC and $q_{0.989}$ for 9 to 21 UTC, due to their relative thunderstorm occurrence frequency (see Section 4.1). The climatology in Table 3 shows that this quantile is close to 100 discharges per 5 minutes for night and morning, 250 between 15 and 21 UTC and 175 between 21 and 03 UTC.

Note that the practical number of predictors in the ELR40 models is always one or two, although for 1-4 predictor ELR models lines have all been included. Additionally, it is important to note that the initial cross-validation has not worked out well for example for +3 to +9 hour forecasts: the model with one predictor has a higher skill, so the model with two predictors seems to be overfitting, whereas the initial cross-validation is not indicating this overfitting. Nonetheless, it can be seen that probabilities of up to 30% are issued reliably with ELR and up to 40% for QRF at +3 to +9 hours lead time. This is the worst among all conditional $q_{0.90}$ forecasts. Although QRF has better reliability, the BSS of the model is 0.03 lower (Figure C-i a and b). The exponential decay (with increasing probability) of relative frequency per probability bin in ELR is clear at this lead time. The same applies to the +9 to +15 hour forecasts: a decay of relative frequency issued per probability bin with increasing probability can clearly be seen for ELR, but not for QRF (Figure C-i c and d). Simultaneously, the forecast probabilities are reliable up to the 60% bin for the +9 to +15 hour lead time, especially for QRF. Beyond this point, the probabilities are not reliable, in part because they are hardly issued. Based on the BSS, QRF is also preferred for +9 to +15 hours. At +21 to +27 hours, the difference in issued probabilities and reliability between ELR and QRF seems not so large.

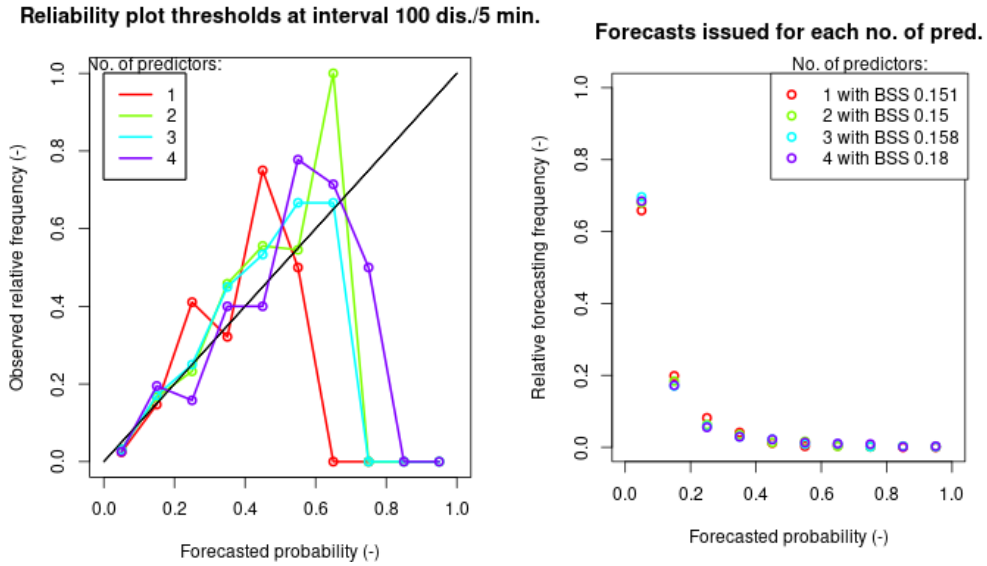
The Figure C-i that belongs to this appendix consists of subfigures a to l, which are on the next six pages, with one page for each valid time; the last valid time has been discussed in Chapter 5 as example. The common caption is found on the last of these six pages.



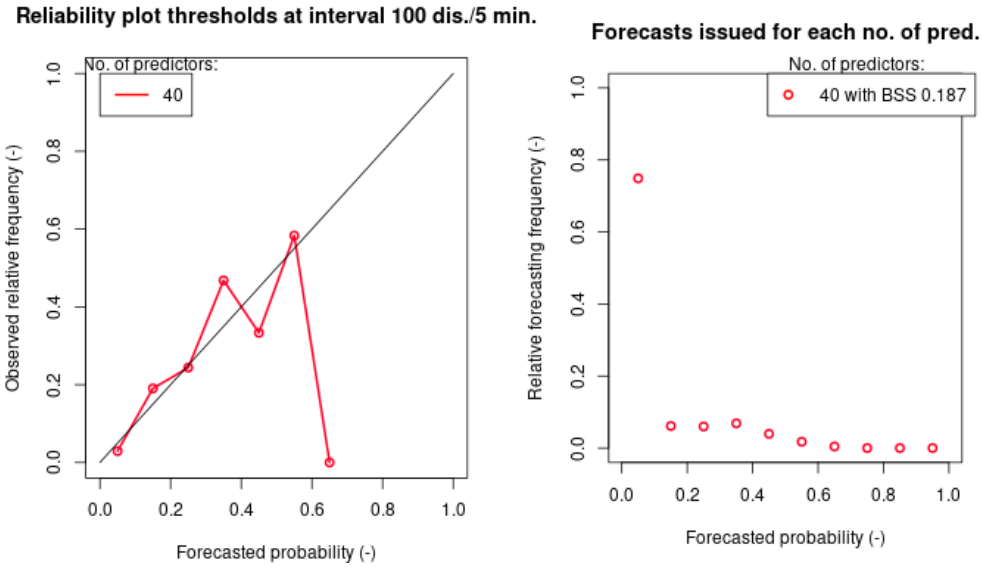
(a) ELR Harmonie00z +3 to +9 hours (selected: 2 predictors)



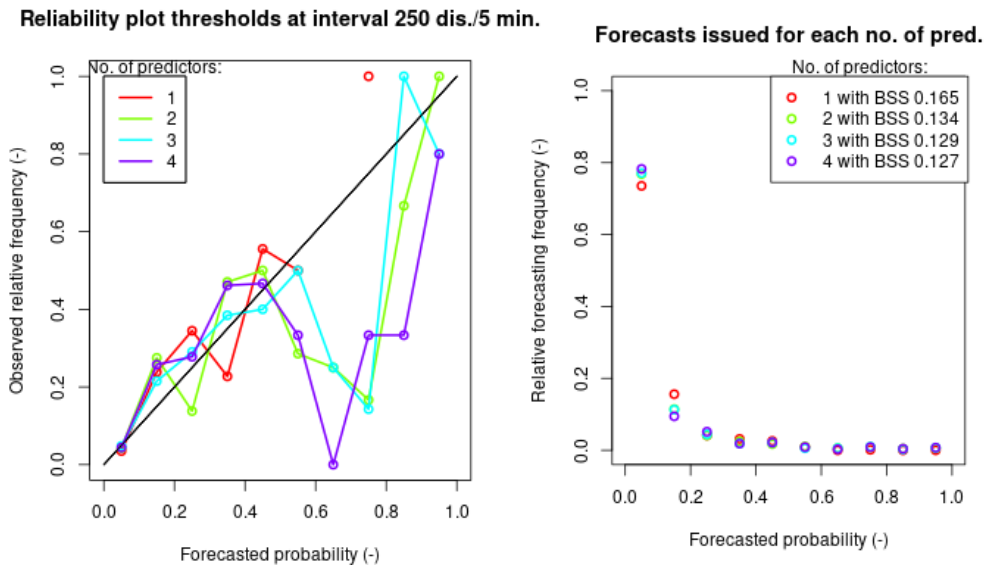
(b) QRF Harmonie00z +3 to +9 hours



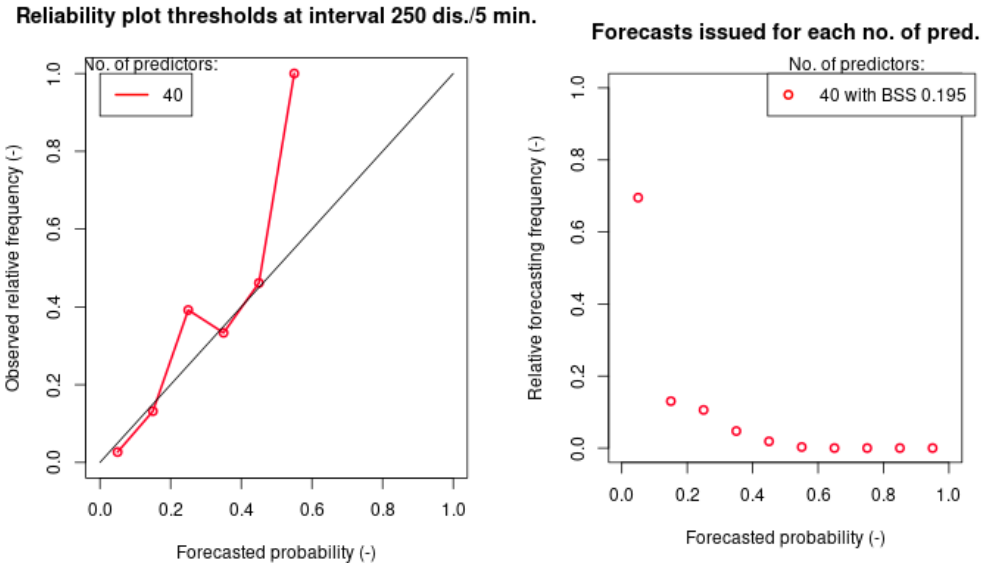
(c) ELR Harmonie00z +9 to +15 hours (selected: 1 predictor)



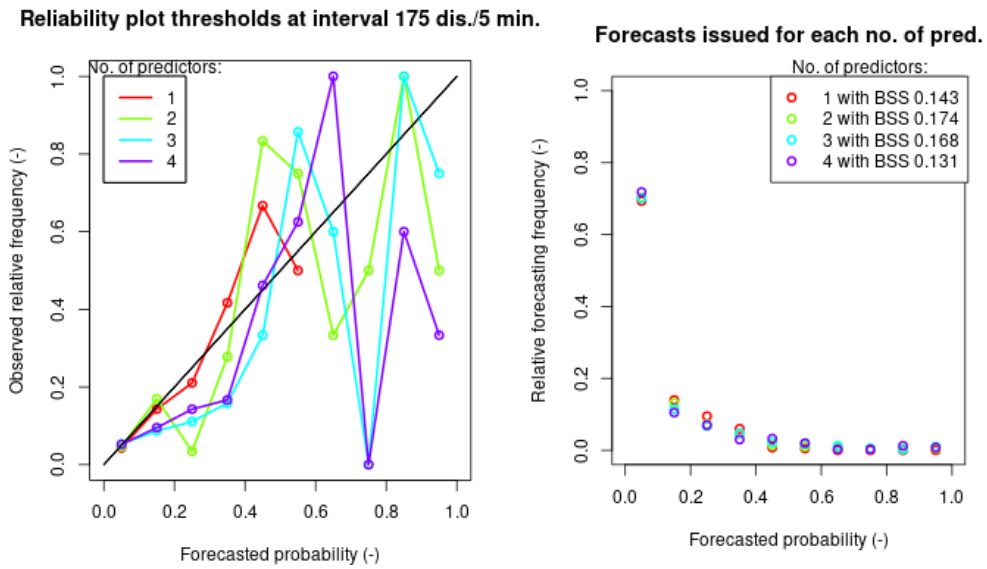
(d) QRF Harmonie00z +9 to +15 hours



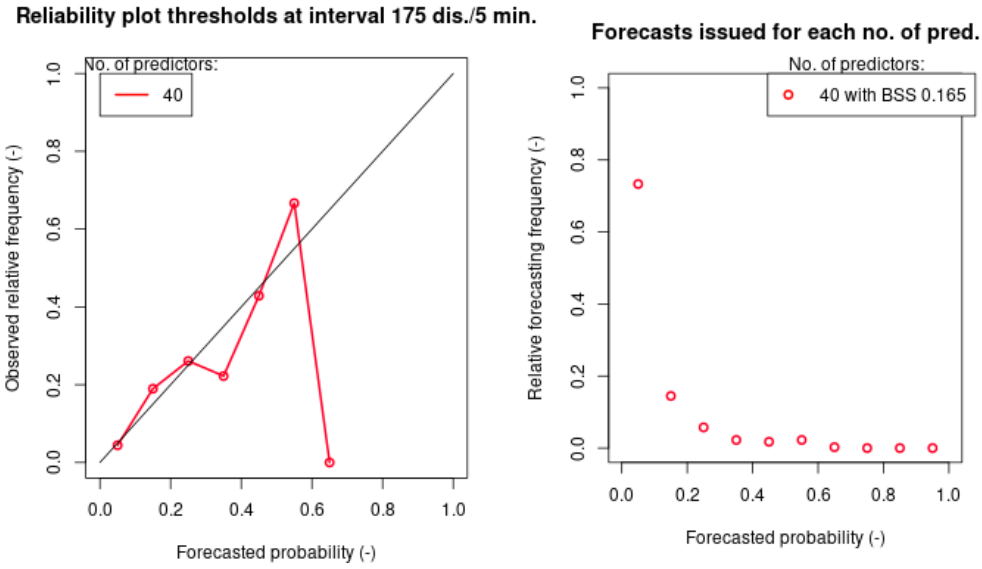
(e) ELR Harmonie00z +15 to +21 hours (selected: 1 predictors)



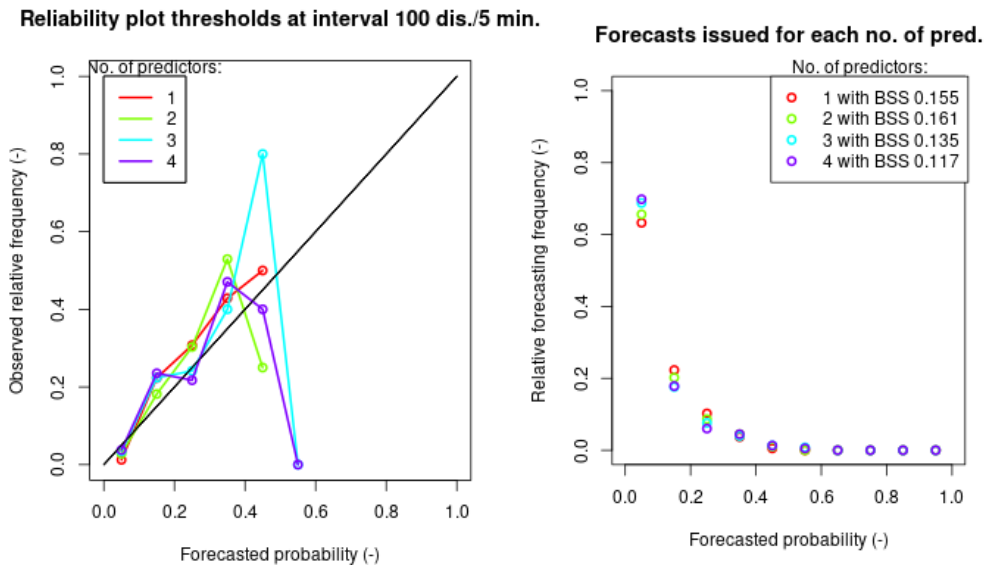
(f) QRF Harmonie00z +15 to +21 hours



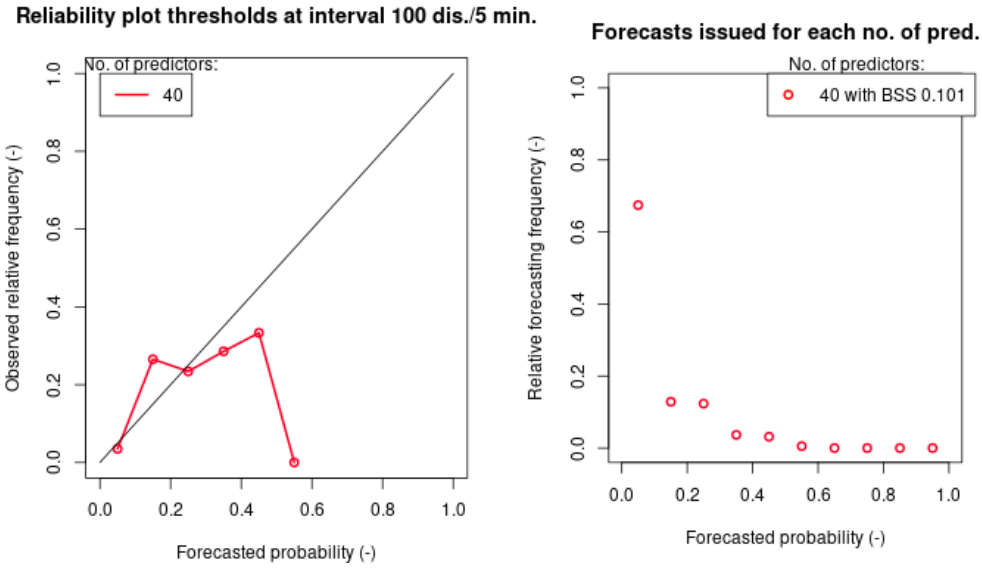
(g) ELR Harmonie00z +21 to +27 hours (selected: 2 predictors)



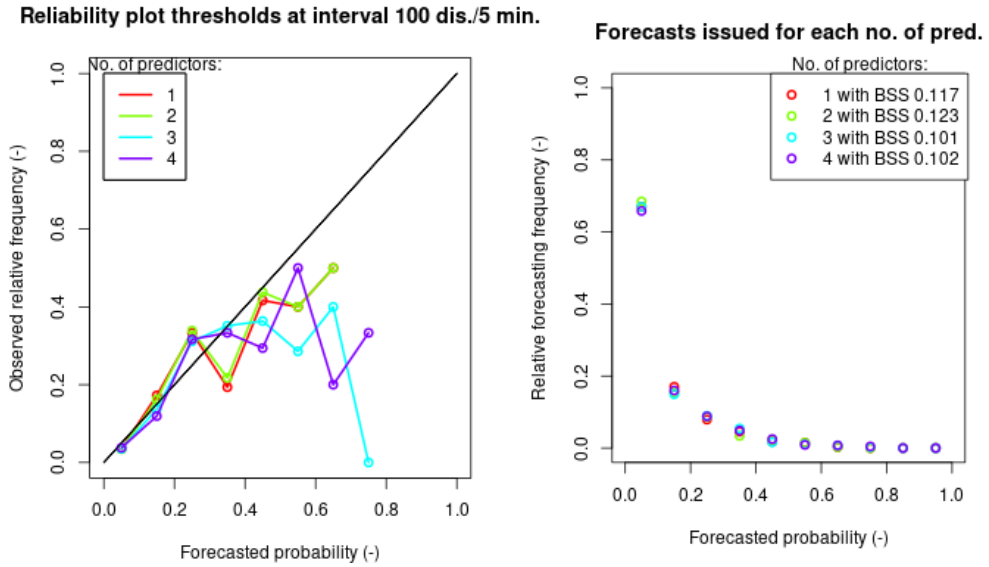
(h) QRF Harmonie00z +21 to +27 hours



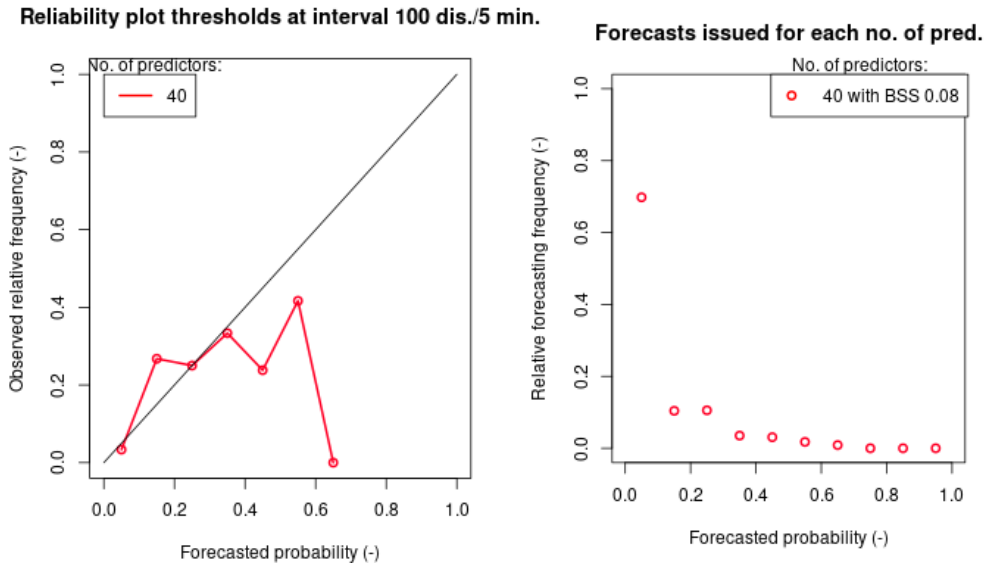
(i) ELR Harmonie00z +27 to +33 hours (selected: 2 predictors)



(j) QRF Harmonie00z +27 to +33 hours



(k) ELR Harmonie00z +33 to +39 hours (selected: 1 predictor)



(l) QRF Harmonie00z +33 to +39 hours

Figure C-i: Reliability diagrams of ELR40 (a, c, e, g, i, k) and QRF40 (b, d, f, h, j, l) forecasts, with both relative frequency of an event (LHS of each double figure) and relative forecasting frequencies per forecast probability bin (RHS of each double figure) for lead times up to +39 hours. Shown lightning intensities are closest to $q_{0.90}$ of the intensity distribution for that valid time.

D Table of additional predictors in (E)LR

The following tables are presented in addition to Tables 5 and 6. They indicate second, third and fourth predictors in the logistic regression models based on the 40 potential predictor set, conditional on their appearance in the final model and expected signs in front of the coefficient if appropriate (see Section 5.2).

Table D-i: Frequency table of predictor selection in LR40 per valid time, with second, third and fourth predictor given if included the valid time. Each valid time is grouped with a common background color. Empty cells indicate no selection (0 frequency).

Validtime	Predictor	Freq as 2nd	Freq as 3rd	Freq as 4th
03-09_UTC	Boyden_max	1		2
	Jefferson_max	2		
	ModJefferson_max	2		
	Trans_Graupel_col_max	1	1	1
09-15_UTC	ModJefferson_max	3		
15-21_UTC	ModJefferson_max	1		
	MUCAPE_graupel_pows_max	1		
	MUCAPE_snow_pows_max	3		
	Trans_Snow_max	1		
21-03_UTC	Boyden_max	1	1	
	LNB_q0.90	1		
	Trans_Graupel_col_max	1		
03-09_UTC	Helicity_max		2	1
	MUCAPE_q0.90		1	2
	Trans_Snow_max		1	
	$\sqrt{SBCAPE + SBCIN}_{max}$		1	
15-21_UTC	MSLP_max		2	1
	Trans_Moisture_convergence_min		1	
	Trans_Moisture_convergence_max			1
	$\sqrt{MUCAPE + MUCIN}_{max}$			1
21-03_UTC	Trans_Snow_max		2	

Table D-ii: Frequency table of second predictor selected per valid time in ELR40 models. Only two valid times have two predictors in ELR40.

Validtime	Predictor	Freq as 2nd
03-09_UTC	coast	2
	dp_dt	1
	MUCAPE_q0.90	1
	$\sqrt{SBCAPE + SBCIN}_{max}$	1
	MSLP_max	1
21-03_UTC	MSLP_max	1
	MUCAPE_q0.90	1
	Boyden_max	1

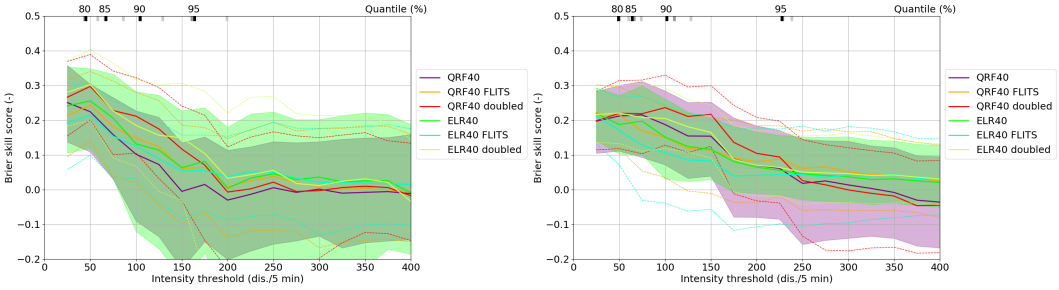
E Importance of predictors in QRF40 for lightning intensity forecasts

In Figure E-i, the permutation importance measure for lightning intensity forecasts is found per predictor and lead time. The figure shows that predictors indicate the magnitude of convective instability and air mass flowing in are typically the most important for conditional thunderstorm intensity predictions: Θ_w, convective indices, their combinations with precipitable water and dew points at 700 and 805 hPa always have relatively large importance (0.04 to 0.18). In addition, the "Edward" predictor with vertical and horizontal gradients in Θ_w included (both convective and baroclinic gradients), is among the more important predictors. Other predictors have sometimes or mostly lower importances. Note that the total of all importance is clearly larger for 15-21 UTC valid times.

Figure E-i: The permutation importance measure of QRF40 fits for seven lead times with lightning intensity forecasts, averaged over three final cross-validations. The colour of a symbol indicates the time of the day; circles indicate that the centre time of the forecast lies in the first 24 hours and crosses indicate a centre time on the second day.

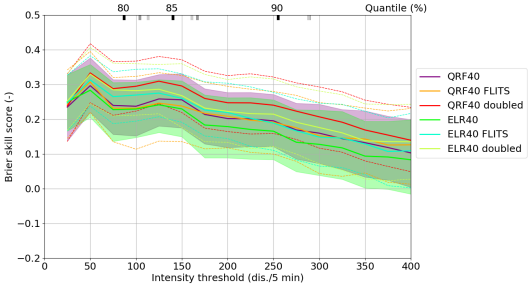
F Dependence on “truth”/“observations”

The BSS as function of lightning intensity threshold for three different ”truths” for lightning intensity forecasts is shown in Figure F-i. The figure shows that the range of skilfully forecast thresholds can shift as a function of the truth, both in terms of climatological quantiles and absolute lightning intensity threshold. Additionally, QRF profits most notably at many lightning intensity thresholds for +21 to +27 hour forecasts (Figure F-i d) when the truth is adjusted and regularly for some thresholds, for example lower thresholds in +3 to +9 hour forecasts (Figure F-i a).

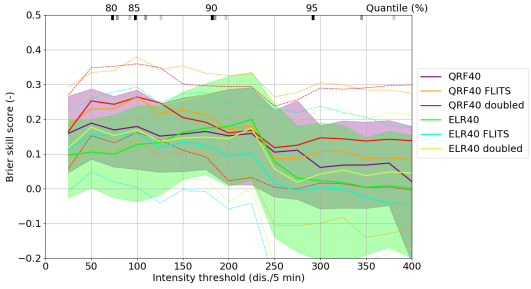


(a) Harmonie00z +3 to +9 hours

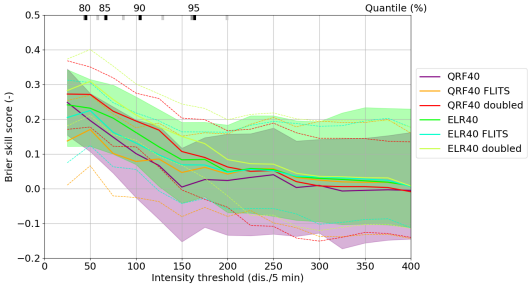
(b) Harmonie00z +9 to +15 hours



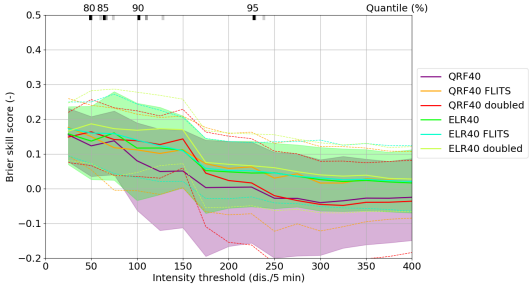
(c) Harmonie00z +15 to +21 hours



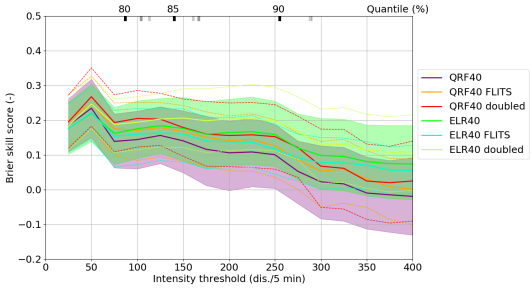
(d) Harmonie00z +21 to +27 hours



(e) Harmonie00z +27 to +33 hours



(f) Harmonie00z +33 to +39 hours



(g) Harmonie00z +39 to +45 hours

Figure F-i: BSS of QRF40 and ELR40 as a function of lightning intensity with the reference truth and two modified truths: 2015 FLITS detections and 2015 doubled KLDN detections, including uncertainty margins (shaded for reference truth and dashed lines for adjusted truths). The four highest training quantiles (if 400) are shown as ticks: black = reference, dark grey = doubled, lighter grey = FLITS).

References

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

[Bauer et al., 2015] Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55.

[Bengtsson et al., 2017] Bengtsson, L., Andrae, U., Aspelien, T., Batrak, Y., Calvo, Javier and, D. R. W., Gleesone, E., Hansen-Sass, B., Homleid, M., Mariano, M. H., Ivarsson Karl-Ivar, Lenderink, G., Niemelä, S., Pagh Nielsen, K., Onvlee, J., Rontu, L., Samuelsson, P., Santos Muñoz, D., Subias, A., Tijm, S., Toll, V., Yang, X., and Ødegaard Køltzow, M. (2017). The harmonie–arome model configuration in the aladin–hirlam nwp system. *Monthly Weather Review*, (145):1919–1935.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, (45):5–32.

[De Vos, 2015] De Vos, L. (2015). Intercomparison of the flits and kldn with atdnet lightning detection systems in the netherlands, with a case study on the potential of infrasound lightning detection. Technical report, Royal Dutch Meteorological Institute (KNMI).

[Doswell et al., 1996] Doswell, C. A., Brooks, H. E., and Maddox, R. A. (1996). Flash flood forecasting: An ingredients-based methodology. *Weather and Forecasting*, 11(4):560–581.

[Glahn and Lowry, 1972] Glahn, H. R. and Lowry, D. A. (1972). The use of model output statistics (mos) in objective weather forecasting. *Journal of Applied Meteorology*, 11(8):1203–1211.

[Gregorutti et al., 2016] Gregorutti, B., Michel, B., and Saint-Pierre, P. (2016). Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678.

[Haklander and Van Delden, 2003] Haklander, A. and Van Delden, A. J. (2003). Thunderstorm predictors and their forecast skill for the netherlands. *Atmospheric Research*, 67-68:273–299.

[Johns and Doswell, 1992] Johns, R. H. and Doswell, C. A. (1992). Severe local storms forecasting. *Weather and Forecasting*, 7(4):588–612.

[Jordan et al., 2018] Jordan, A., Krueger, F., and Lerch, S. (2018). Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*. forthcoming.

[KNMI, 2016] KNMI (2016). Maandoverzicht van het weer in nederland. Retrieved from: *https://cdn.knmi.nl/knmi/map/page/klimatologie/gegevens/mow/mow_201604.pdf*.

[Laboratory, 2015] Laboratory, N. R. A. (2015). *verification: Weather Forecast Verification Utilities*. R package version 1.42.

[Lewis and Gray, 2010] Lewis, M. W. and Gray, S. L. (2010). Categorisation of synoptic environments associated with mesoscale convective systems over the uk. *Atmospheric Research*, 97(1-2):194–213.

[Lilly, 1986] Lilly, D. K. (1986). The structure, energetics and propagation of rotating convective storms. part ii: Helicity and storm stabilization. *The Structure, Energetics and Propagation of Rotating Convective Storms. Part II: Helicity and Storm Stabilization: Journal of the Atmospheric Sciences: Vol 43, No 2*.

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

[Markowski and Richardson, 2010] Markowski, P. and Richardson, Y. (2010). *Mesoscale meteorology in midlatitudes*. Wiley-Blackwell.

[Mecikalski et al., 2015] Mecikalski, J. R., Williams, J. K., Jewett, C. P., Ahijevych, D., Leroy, A., and Walker, J. R. (2015). Probabilistic 0–1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *Journal of Applied Meteorology and Climatology*, 54(5):1039–1059.

[Meinshausen, 2006] Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, (7):983–999.

[Messner et al., 2014a] Messner, J. W., Mayr, G. J., Wilks, D. S., and Zeileis, A. (2014a). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review*, 142(8):3003–3014.

[Messner et al., 2014b] Messner, J. W., Mayr, G. J., Zeileis, A., and Wilks, D. S. (2014b). Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Monthly Weather Review*, 142(1):448–456.

[Munich Re, 2016] Munich Re, . (2016). Reinsurance: global risk solutions from munich re. Article editor: Faust, E. Retrieved from: *https://www.munichre.com/topics-online/en/climate-change-and-natural-disasters/natural- disasters/storms/severe-thunderstorms-europe-2015.html*.

[Noteboom, 2006] Noteboom, S. (2006). Processing, validatie, en analyse van bliksemdata uit het safir/flits systeem. Technical report, Royal Dutch Meteorological Institute (KNMI). In Dutch.

[Púčik et al., 2015] Púčik, T., Groenemeijer, P., Rýva, D., and Kolář, M. (2015). Proximity soundings of severe and nonsevere thunderstorms in central europe. *Monthly Weather Review*, 143(12):4805–4821.

[Schmeits et al., 2005] Schmeits, M. J., Kok, K. J., and Vogelezang, D. H. P. (2005). Probabilistic forecasting of (severe) thunderstorms in the netherlands using model output statistics. *Weather and Forecasting*, 20(2):134–148.

[Schmeits et al., 2008] Schmeits, M. J., Kok, K. J., Vogelezang, D. H. P., and Westrhenen, R. M. V. (2008). Probabilistic forecasts of (severe) thunderstorms for the purpose of issuing a weather alarm in the netherlands. *Weather and Forecasting*, 23(6):1253–1267.

[Siegert, 2017] Siegert, S. (2017). *SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate*. R package version 0.5-2.

[Simon et al., 2018] Simon, T., Fabsic, P., Mayr, G. J., Umlauf, N., and Zeileis, A. (2018). Probabilistic forecasting of thunderstorms in the eastern alps. *Monthly Weather Review*, 146(9):2999–3009.

[Taillardat et al., 2016] Taillardat, M., Mestre, O., Zamo, M., and Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393.

[Takahashi, 1978] Takahashi, T. (1978). Riming electrification as a charge generation mechanism in thunderstorms. *Journal of the Atmospheric Sciences*, 35(8):1536–1548.

[Taszarek et al., 2019] Taszarek, M., Allen, J., Púčik, T., Groenemeijer, P., Czernecki, B., Kolendowicz, L., Lagouvardos, K., Kotroni, V., and Schulz, W. (2019). A climatology of thunderstorms across europe from a synthesis of multiple data sources. *Journal of Climate*, 32(6):1813–1837.

[Thorarinsdottir and Schuhen, 2018] Thorarinsdottir, T. L. and Schuhen, N. (2018). Verification: Assessment of calibration and accuracy. *Statistical Postprocessing of Ensemble Forecasts*, page 155–186.

[University of Wyoming, nd] University of Wyoming, . (n.d.). Wyoming weather web. Data service maintained by: Oolman, L.D. Retrieved from: *http://weather.uwyo.edu/upperair/sounding.html*.

[Van Delden, 1998] Van Delden, A. (1998). The synoptic setting of a thundery low and associated prefrontal squall line in western europe. *Meteorology and Atmospheric Physics*, 65(1-2):113–131.

[Van Zomeren and Van Delden, 2007] Van Zomeren, J. and Van Delden, A. J. (2007). Vertically integrated moisture flux convergence as a predictor of thunderstorms. *Atmospheric Research*, 83(2-4):435–445.

[Venables and Ripley, 2002] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

[wetter3.de, nd] wetter3.de (n.d.). Archiv-version des animationstools. Website exploited by: Behrendt, R. and Mahlke, H. Retrieved from: *http://www1.wetter3.de/archiv_gfs_dt.html*. In German.

[Whan and Schmeits, 2018] Whan, K. and Schmeits, M. (2018). Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Monthly Weather Review*, 146(11):3651–3673.

[Wilks, 2009] Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution mos forecasts. *Meteorological Applications*, 16(3):361–368.

[Wilks, 2011] Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. Elsevier Science Publishing Co Inc.

[Williams et al., 1999] Williams, E., Boldi, B., Matlin, A., Weber, M., Hodanish, S., Sharp, D., Goodman, S., Raghavan, R., and Buechler, D. (1999). The behavior of total lightning activity in severe florida thunderstorms. *Atmospheric Research*, 51(3-4):245–265.

[Wright and Ziegler, 2017] Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.