PREDICTING ROTATION IN SUPERCELL THUNDERSTORMS WITH SUPERVISED MACHINE LEARNING

Master Thesis

by Louise Aubet

Handed in on January 20, 2022

Supervisor: Prof. Alexis Berne,
Co-supervisor: Monika Feldmann,
Environmental Remote Sensing Laboratory (LTE),
Ecole Polytechnique Fédérale de Lausanne

Acknowledgements

I would first like to thank Prof. Alexis Berne, my supervisor for this thesis. He has always taken the time to discuss the project progress and helped me to stay focused on the objective. I would like to offer my special thanks to Monika Feldmann, my co-supervisor, whose expertise in supercell thunderstorms was highly valuable. She provided the data used in this work, supported me in the formulation of the research questions, always had insightful feedback and was available at every stage of the project. Furthermore, I thank Meteoswiss members in Locarno-Monti as well as Prof. Tom Beucler, who provided relevant comments and suggestions that helped bring my work to a higher level. I acknowledge all the LTE members, for being great colleagues and for their cooperation. Finally, I thank all of my family and friends who are always there to support me.

Louise Aubet

Abstract

This work describes the development of Random Forests and Convolutional Neural Networks (CNN) to predict occurrences of supercells, based on radar-derived thunderstorm intensity variables. They are trained with a multi-year dataset of supercell occurrences in Switzerland, output from the operational Thunderstorm Radar Tracking algorithm and the Mesocyclone Detection Algorithm (MDA) that identifies supercells among detected storm cells. The MDA algorithm relies on Doppler velocity data. This work aims to develop a machine learning algorithm that can predict rotation in thunderstorms using the environmental variables only. There is no existing machine learning algorithm for this task. A Random Forest is used to predict the probability of rotation for each timestep and is compared to a Logistic Regression. A CNN is used to predict the probability of rotation of storms, taking into account the entire track. It is compared to a classic Artificial Neural Network (ANN) and a meteorological baseline that is using only the RANK feature which evaluates the severity of the storm. These two models could classify thunderstorms very fast and without the need for Doppler velocity data, which could offer the opportunity to extend mesocyclone detection further back in time, where the capability of the MDA algorithm may be limited. The Random Forest performs significantly better than the Logistic Regression, the performance is improved when focusing on severe thunderstorms only. Interpretation of the model through feature importance shows that shape and intensity features are the most important ones. The CNN performs better than the meteorological baseline but similarly to the ANN, suggesting that the convolutional step might not be necessary for this task. The analysis of the feature importance shows that the location and topography features play a bigger role in this case. For both models, the correlation observed between the features and the impact on the prediction tend to show that the models can extract meaningful information for the classification of the storms.

Contents

List of Figures

List of Tables

List of Acronyms

ANN Artificial Neural Network

ASL Above Sea Level

AUC Area Under the Curve

BS Brier Score

BSS Brier Skill Score

CNN Convolutional Neural Network

CPC Combiprecip

DT Decision Tree

ET15 15 dBZ Echo Top height

ET45 45 dBZ Echo Top height

FAR False Alarm Rate

FN False Negative

FP False Positive

GWT Gross Wetter Types

IQR Inter-Quantile Range

LIME Local Interpretable Model-agnostic Explanations

LR Logistic Regression

LSTM Long Short-Term Memory

MCC Matthews correlation coefficient

MDA Mesocyclone Detection Algorithm

MESHS Maximum Expected Severe Hail Size

POH Probability Of Hail

ReLU Rectified Linear Unit

RF Random Forest

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

SHAP SHapley Additive exPlanations

TN True Negative

TP True Positive

TPI Topographic Position Index

TRT Thunderstorms Radar Tracking

VIL Vertically Integrated Liquid

1 Introduction

1.1 Motivation

Severe weather phenomena can cause serious fatalities, injuries and damages. Although such events have a low annual frequency, especially for the most severe events that have the highest potential to create major disasters, they should not be underestimated (Doswell, 2003). Indeed, storms were the costliest natural hazard in Europe between 1980 and 2015 and more generally, losses caused by natural catastrophes have been multiplied by 3 within the last 35 years (Hoeppe, 2016). This increase is due to an intensification of urban sprawling over the last years but also an increased frequency of events. Indeed, the frequency of severe thunderstorms is likely to increase in Europe due to climate change (Rädler et al., 2019, Pucik et al., 2017). Because of this increasing hazard, it is crucial to accurately monitor the presence and evolution of severe convective storms. Supercell thunderstorms are particularly severe convection and are associated with various other phenomena, such as heavy rain, hail, lightning, strong wind gusts, and sometimes even tornadoes (Houze, 2014). In the Alpine area, tornadoes are rather rare but more frequently, supercell thunderstorms can trigger strong hail, flooding, strong winds and mudslides (Taszarek et al., 2019). Classification of supercell thunderstorms using machine learning could be useful for long-term climatology studies and could bring more information for real-time forecasting.

1.2 Theory

A storm cell is an air mass that contains an updraft and a downdraft in a convective movement, and that moves and reacts as a single entity. It is the smallest unit of a storm. To form, a thunderstorm needs warm and moist surface air below cold air, which is unstable, and a lifting force (Donald Ahrens and Henson, 2015). The typical life-cycle of an ordinary storm cell is composed of three stages (Lohmann et al., 2016), illustrated in Figure 1 :

1. Growth stage (or cumulus stage), Figure 1, part (a): Warm and moist air is rising, creating an updraft. When rising, it cools down and condenses into liquid drops of water, forming a cumulus cloud. At first, the cloud cannot grow high because the cloud droplets evaporate as dry air around is entrained. After evaporation the air is moister, so rising air can go higher and the cumulus grows in height.

2. Mature stage, Figure 1, part (b): The air is more humid, the cloud grows bigger and higher. At high altitude above the 0°C isotherm, the cloud particles are able to become larger and heavier and they eventually fall as rain, snow or hail. The air is, therefore, drier, colder and heavier so it goes down and a downdraft appears. When the downdraft hits the ground, it spreads out in all directions and creates a gust front. This is also the stage of the life-cycle during which the highest amount of lightning, thunder, showers and hail is observed. This stage lasts for about 30min.

3. Dissipation stage, Figure 1, part (c): The downdraft dominates over the updraft, light precipitation can still occur.

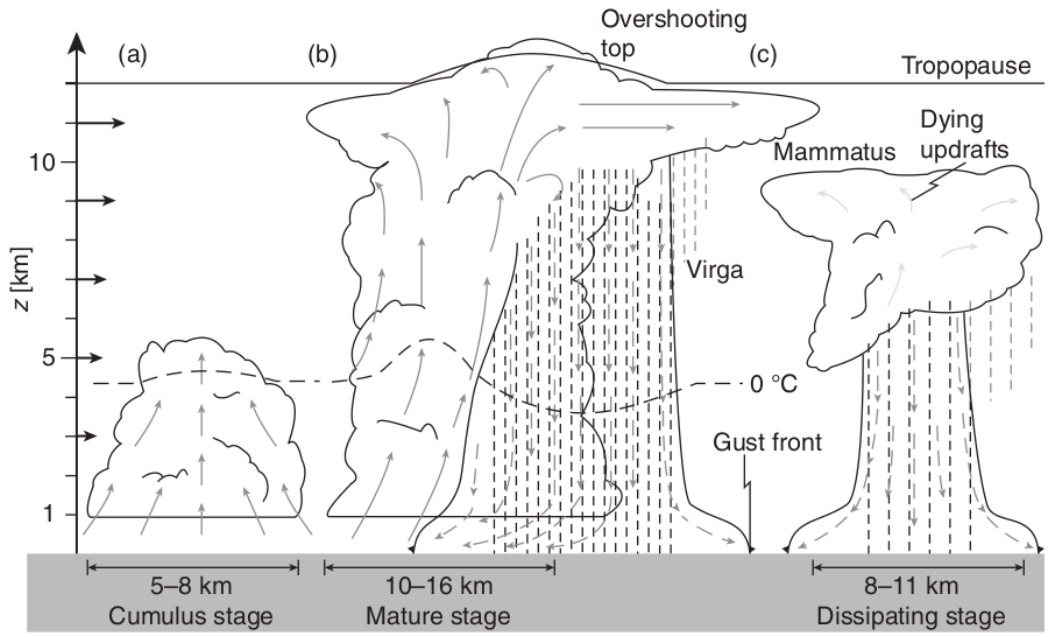

Figure 1 – Diagram showing the life-cycle of an ordinary storm cell. Image from Lohmann et al., 2016.

There are different types of thunderstorms :

- Single-cell thunderstorm (or ordinary cell, pulse storm): It is a thunderstorm producing severe weather for a short period, described by the typical life-cycle in Figure 1. It occurs when there is weak vertical wind shear.
- Multi-cell thunderstorm: It is a thunderstorm composed of multiple storm cells at different stages of the life-cycle described in Figure 1. It can therefore last longer in time. It usually occurs when there is a moderate to strong vertical wind shear.
- Supercell thunderstorm: It is a thunderstorm characterised by the presence of a mesocyclone, as illustrated in Figure 2. A mesocyclone is a deep persistently rotating updraft around a vertical axis. Supercell thunderstorms are more violent and destructive than the other types of thunderstorms. They occur less frequently because they need a stronger and deeper wind shear to develop.

A mesocyclone is a persistently rotating updraft that develops inside a supercell thunderstorm. It typically has a vortex diameter of around 3 to 9 km and a tornado can develop within it near the ground. At the top of the mesocyclone, the supercell thunderstorm we can often see the overshooting top of the cloud that is formed of rising air intruding into the stratosphere. The mesocyclone is detectable with a Doppler weather radar, as it presents as a rotation dipole with velocity seen as positive on one side and negative on the other side (Fabry, 2015). These radar images then allow forecasters to identify supercell thunderstorms. They have a typical life-cycle that differs from single-cell thunderstorms' one and that is more complex. Apart from moist air at the surface and conditionally unstable atmosphere, wind shear is the key element for supercells as it induces rotation. In a simplified example of supercell thunderstorms' life-cycle, strong vertical wind shear in the lower troposphere generates stronger wind speed at higher altitudes and therefore the creation of horizontal vorticity. When this horizontal vorticity is combined with a local updraft below the cloud, it is tilted into two rotating updrafts, with a cyclonic and an anti-cyclonic rotation. The formation of rain creates a downdraft in between the two updrafts and initiates the storm split. The supercell splits in two, one side of the storm propagates to the left and the other to the right. They are called left, respectively right, mover.

Each mover is considered a supercell. For the right mover, the updraft has a cyclonic rotation and the downdraft has an anti-cyclonic rotation. Often, one mover is favoured over the other and the other one dies. In the resulting supercell, the rotating updraft is persistent and can last for hours as the outflow of cold air from the downdrafts never crosses the updraft. Indeed, as it can be seen in Figure 2, they are horizontally separated. This allows potentially very large hailstones to form. The forward flank downdraft generates heavy precipitations, a gust front and has the higher reflectivity. The rear flank downdraft generates fewer precipitations and often appears wrapped around the mesocyclone, exhibiting a hook shape, also called hook-echo. A schematic diagram is shown in Figure 3. This hook-echo near the surface surrounds a small weak echo hole in which a tornado can develop.

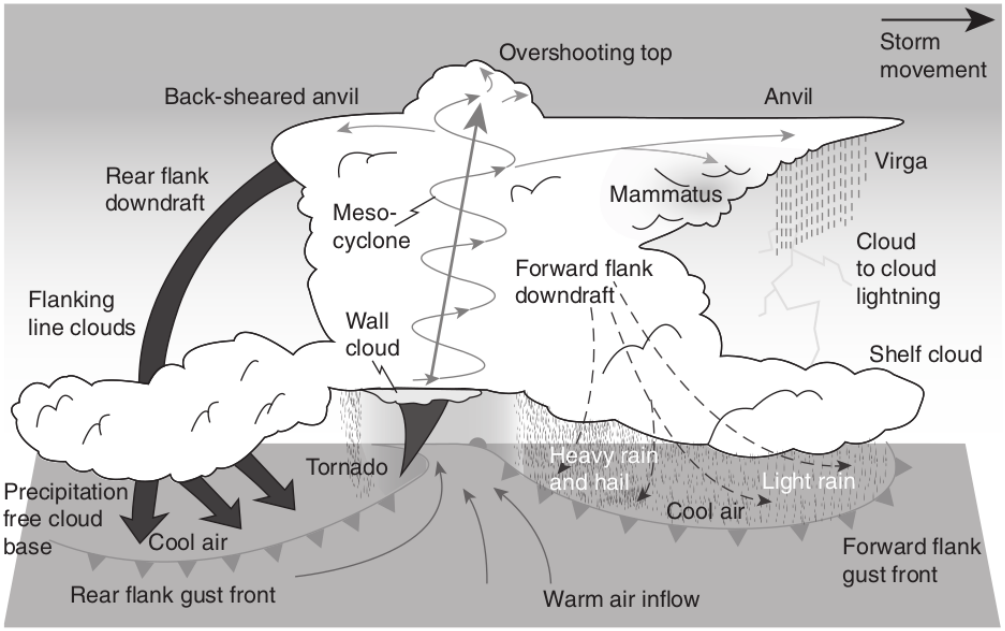

Figure 2 – Diagram showing the structure of a supercell thunderstorm. Image from Lohmann et al., 2016.

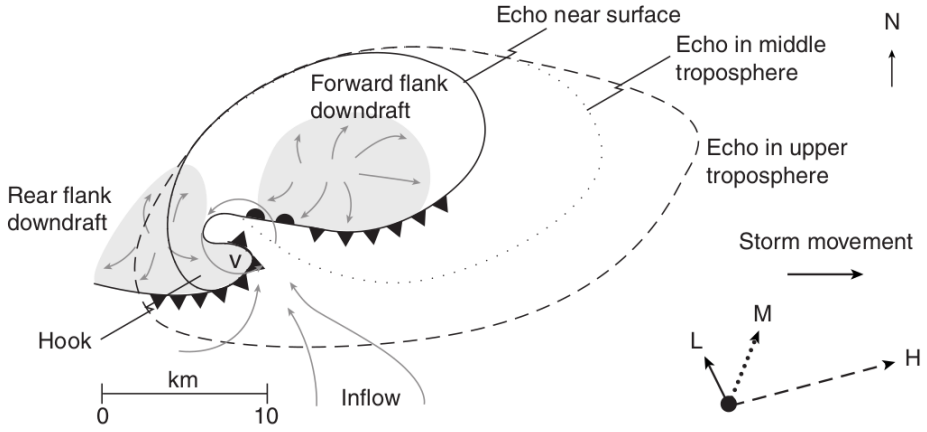

Figure 3 – Diagram showing radar echoes at different altitudes and a typical hook-echo near the surface in a supercell thunderstorm. Image from Lohmann et al., 2016.

1.3 Literature review

Studies on severe weather phenomena provide an overview of the climatology of such events. Taszarek et al., 2019 summarise that the annual peak for thunderstorm activity is in July and August over northern, eastern, and central Europe. A 15-years hail streak climatology in the Alps shows that synoptic weather situation is strongly correlated to hail streak frequency as most hail streaks occur during weather regimes with winds from the South-West and the West. In addition, the highest densities of events are found in the Prealps and the Jura. Only a few events are located over the Alps. A strong correlation is observed between the spatial density of hail occurrence and thunderstorm occurrence, which means that the areas with a high occurrence of hail are also areas with a high occurrence of thunderstorms (Nisi et al., 2018).

A study on supercell thunderstorms more particularly provides an overview of the occurrence and characteristics of mesocyclones in Germany, based on 3-year statistics from mesocyclone occurrences (Wapler et al., 2016). A typical annual and diurnal cycle is observed, as approximately 75% of mesocyclones occur during the months of June, July and August. The peak for the diurnal cycle is around 17 UTC, similarly to the peak of lightning stroke occurrences. Furthermore, mesocyclones are linked to hail as half of the hail events in Germany are associated with a nearby mesocyclone. Some storm attributes such as depth, maximum shear, vertically integrated liquid or EchoTop height, are found to be good indicators for the occurrence of hail. A study on mesocyclones in Germany by Wapler, 2021 compares mesocyclonic and non-mesocyclonic storms. It presents a deeper analysis of the physical characteristics of storms over the entire life-cycle, using a dataset covering several years, from 2013 to 2017. The study reveals that supercells tend to be larger, last longer, grow faster and move faster than other types of thunderstorms. These differences are presented as statistically significant with an increase of 30min in duration and 15 km/h in velocity. They also have on average higher reflectivity and higher lightning rates. For stronger mesocyclones, these discrepancies are more prominent. Finally, in terms of weather types, it shows that both supercells and non-supercells are mostly coming from the South-West and the West. Feldmann et al., 2021 describe the implementation of a Mesocyclone Detection Algorithm (MDA). Using data from the operational radar network of Switzerland, this algorithm creates a dataset with labelled mesocyclone occurrences from 2016 to 2020, allowing the characterisation of spatio-temporal distribution of mesocyclones. They observe that the areas with the highest densities of mesocyclones are the Southern Prealps, followed by the Northern Prealps. The study also shows the effect of terrain and more particularly the fact that altitude is detrimental to supercell intensity and frequency, as it can be seen that vorticity and rotational velocity decrease with altitude. The formulated hypothesis is that the terrain is in this case starting to interfere with the Planetary Boundary Layer. Furthermore, the occurrence of mesocyclones is shown to be strongly correlated with synoptic weather situations as approximately 80% of all mesocyclone detections are coming from South-West and West directions. These results are similar to Nisi et al., 2018 observations on hailstorms for both terrain and weather situations. They are also similar to Wapler, 2021 observations on mesocyclones. Concerning the diurnal cycle, mesocyclones are the most frequent in the late afternoon and early evening. Mesocyclones diurnal distribution is very similar to thunderstorms diurnal distribution with a peak at 16 UTC.

Machine learning is a topic that is becoming increasingly researched in the field of atmospheric science, to improve weather and climate predictions. It uses a different approach, without any explicit physical law, learning directly from the data to extract patterns and structures while optimising an error metric. Nowadays, with the increasing available computational speed,

large labelled datasets and more advanced methods, machine learning models could be able to compete with numerical weather models (Chantry et al., 2021). Supervised learning methods like Random Forests have demonstrated good performance in classification (Jergensen et al., 2020, Ruiz-Gazen and Villa, 2007 Gagne et al., 2017, Burke et al., 2020) and prediction tasks (Mostajabi et al., 2019, Wolfensberger et al., 2021, Kabir et al., 2019). Jergensen et al., 2020 use Random Forests and Gradient Boosted Forests, compared to other ML models, to classify storms as supercells, Quasi-Linear Convective Systems or others. The Gradient Boosted Forest performs the best in this case. Feature importance and interpretability of the models are investigated. Feature importance is measured using permutation importance and shape features are found to be the most important, compare to radar features or sounding features. Furthermore, partial dependence plots illustrate the effect of each feature on the outcome. Ruiz-Gazen and Villa, 2007 compare Random Forest and Logistic Regression for the classification of cloud systems, in two categories: convective or non-convective. As the dataset is highly unbalanced, the majority class is downsampled to reach a ratio of 0.2 between the two classes. In a study by Wolfensberger et al., 2021, a Random Forest regression algorithm is used for estimating the amount of precipitation over Switzerland. Compared to the current algorithm providing rainfall estimations from radar data, the algorithm has a smaller error and bias.

Convolutional Neural Networks (CNN) are another widely used supervised machine learning tool. 2D-CNNs are mostly used for image classification tasks. They allow using directly the images, without going through pre-processing steps such as the computation of various statistics, which would result in a loss of information. CNNs have been used in atmospheric sciences for the classification and prediction of thunderstorms and tornadoes using radar images. Molina et al., 2020 use CNN for the classification of strongly rotating storms. The model is trained with data from the current climate and tested on data from a future warmer climate. It performs similarly in both climates because it can extract relevant information. A study from Lagerquist et al., 2020 shows that a CNN is capable of predicting tornado occurrences during the next hour, using radar images from different sources. The model performs very good, comparably to an operational machine learning model for severe weather prediction. Gagne II et al., 2019 demonstrate that CNN can also be used to predict severe hailstorms. It performs better than other machine learning methods. Through interpretation of the model, key elements for the learning are identified, such as strong lapse rates, i.e. the rate at which temperature changes vertically, or wind shear. More recently, 1D-CNN have been used for time series classification. Ismail Fawaz et al., 2019 provide an overview on CNNs, and more particularly on the topic of time series classification. Kiranyaz et al., 2021 focus on 1D-CNN and their different applications. A study by Coffer et al., 2020 uses a 1D-CNN to predict tornadoes. As the dataset is highly unbalanced, random oversampling or undersampling are applied and both lead to better performance as the accuracy of the minority class is increasing.

1.4 Research questions

In that context, the objective of this project is to answer the following main research questions:

- How can supervised machine learning methods help identify supercell thunderstorms among other types of thunderstorms?
- Which features are identified as the most important for correct classification?
- What are the reasons for incorrect classifications?

The Thunderstorm Radar Tracking algorithm detects thunderstorm cells from radar images and the Mesocyclone Detection Algorithm identifies the supercell thunderstorms among them using Doppler velocity data. However, Doppler velocity data is not available in high enough quality before the year 2016. Machine learning models requiring environmental variables only, could be helpful to classify mesocyclones further back in time. As there is no existing machine learning algorithm for this task, this project aims to provide proof of concept and feasibility. If the model is robust enough, it can be extrapolated and applied to earlier years. This would allow to extend the mesocyclone database, which would be very useful for climatology studies. Furthermore, it can provide additional hazard assessment information on thunderstorms in the absence of a mesocyclone detection algorithm or help fill in probable detection in areas, where radar data coverage is poor. Additionally, the feature importance and predictive quality of the input variables hold information about rotation in thunderstorms.

1.5 Structure

This thesis is organised as follows. Section 2 provides a detailed description of the dataset. Section 3 describes how the data is prepared, how the models are built and evaluated. Section 4 contains results and discussion. Section 5 provides a summary and some concluding remarks.

2 Experimental data

The data is originally coming from the Swiss Rad4Alp network (Germann et al., 2016). This weather radar network is composed of five dual-polarisation Doppler weather radar stations that allow to record precipitation and storms in real-time over the whole of Switzerland and also in some areas outside the borders. Figure 4 shows the altitude Above Sea Level (ASL) [m] in the Alpine region, Switzerland's borders, as well as the location and elevation of the five Doppler radars.

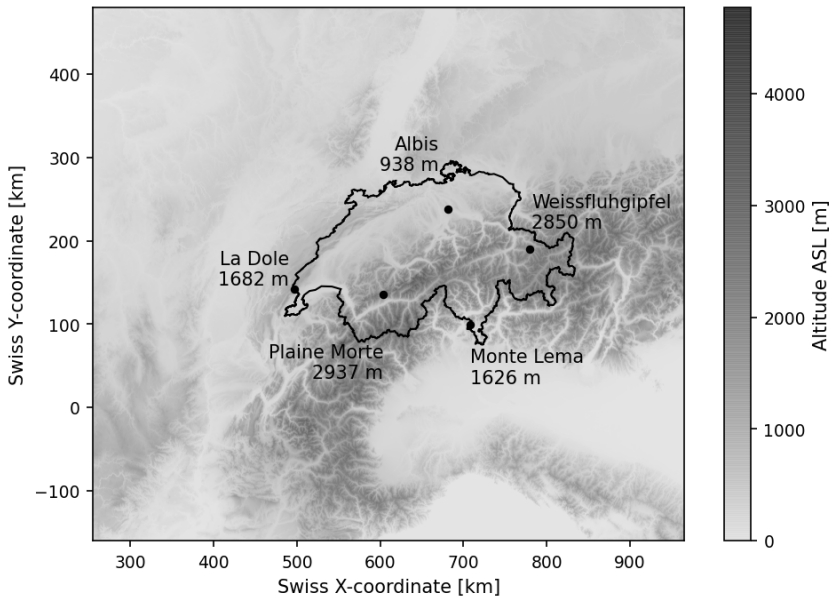

Figure 4 – Map showing the altitude ASL [m] in the Alpine region and Switzerland's borders. The black dots indicate the location of the five Doppler radars. Their respective names and altitudes are given.

The TRT (Thunderstorm Radar Tracking) algorithm (Hering et al., 2004), the current Meteoswiss thunderstorm tracking algorithm, is used to detect thunderstorm tracks and thunderstorm cells from radar images showing reflectivity values. Thunderstorm cells are identified with adaptive reflectivity thresholds. The algorithm provides metrics such as :

- general variables: storm ID, date and time features.
- position: longitude and latitude.
- thunderstorm intensity variables: area, zonal and meridional propagation velocities, height of EchoTop 45 dBZ (ET45), composite reflectivity, precipitation, Probability Of Hail (POH), Maximum Expected Severe Hail Size (MESHS), flash count, RANK.

The RANK is a heuristic variable describing thunderstorm severity, computed as follows (Hering et al., 2008) :

$$\text{RANK} = \text{round}\left[(2 \cdot \text{VIL} + 2 \cdot \text{ET45m} + 1 \cdot \text{dBZmax} + 2 \cdot \text{area57dBZ})\ /7\right] \in [0, 40] \qquad (1)$$

with VIL being the Vertically Integrated Liquid [kg/m^2], ET45m being the maximum height of EchoTop 45 dBZ [km], dBZmax being the maximum cell reflectivity [dBZ] and area57dBZ being the area above 57 dBZ [km^2]. The RANK allows aggregate data from radar-based attributes to assess the potential danger of the storm. Table I summarises the different severity classes of thunderstorms depending on the value of the RANK.

Table I – RANK and severity of thunderstorms

RANK	**Severity**
0-12	Very weak
12-15	Weak
15-25	Moderate
25-35	Severe
35-40	Very severe

The Mesocyclone Detection Algorithm (MDA, Feldmann et al., 2021) is used to identify mesocyclones among thunderstorms tracked by TRT. For every timestep, the algorithm provides a label indicating if a mesocyclone is detected at this timestep or not. For detected mesocyclone, it indicates the direction of rotation. A positive label corresponds to cyclonic rotation and a negative label corresponds to anticyclonic rotation. The algorithm also outputs rotational properties such as rotational velocity, vorticity, diameter, rank, altitude. The statistics of these variables computed over the storm cell are given. The rank of every detected 2D area is defined as below (Feldmann et al., 2021) :

$$\text{rank} = 5 \cdot \text{mean}\left[(\text{rvel} - \text{rvel\_min})/\text{rvel\_min};\, (\text{vort} - \text{vort\_min})/(4 \cdot \text{vort\_min})\right] \qquad (2)$$

with rvel being the rotational velocity [m/s] and vort the vorticity [s^{-1}]. The scale goes from 0 to 5, with 5 being approximately the 98^{th} percentile of cases in Switzerland. The final rank statistic values are the percentiles of all detections combined. The rvel_min and vort_min values are range dependent and thus different for each object contributing to the total rank.

In supplement, meteorological and topographical variables are added to the dataset. 4 features, named Gross Wetter Types (GWT), contain information about 4 synoptic weather classifications from MeteoSwiss (Weusthoff, 2011):

- GWT8: weather classification with 8 classes for the 8 wind directions (South, South-West, West, North-West, North, North-East, East and South-East), based on geopotential height at 500 hPa.
- GWT10: weather classification with 10 classes, 8 wind directions and 2 classes for low or high pressure, based on geopotential height at 500 hPa.
- GWT18: weather classification with 18 classes, 2×8 wind directions with an additional indication on cyclonic or anticyclonic rotation and 2 classes for low or high pressure, based on geopotential height in 500 hPa.
- GWT26: weather classification with 26 classes, with 3×8 wind directions with an additional indication on indifferent, cyclonic or anticyclonic rotation and 2 classes for low or high pressure, based on geopotential height in 500 hPa.

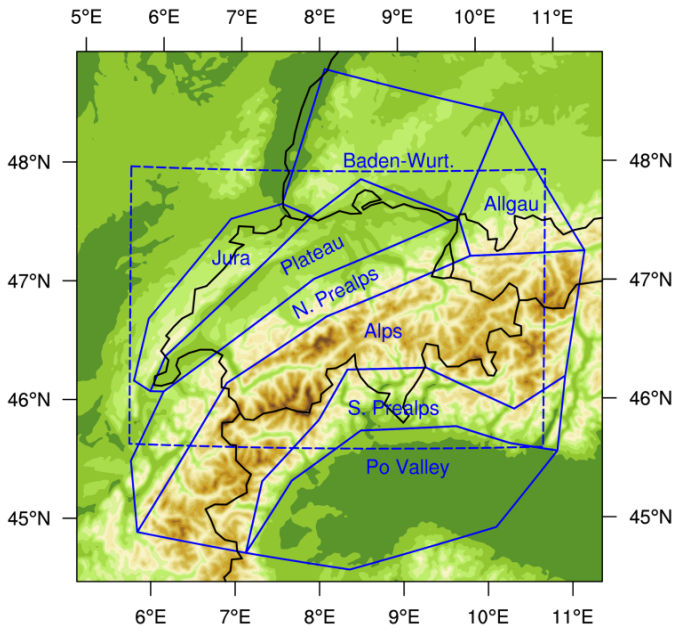

Figure 5 – Map illustrating the different sub-domains used in the dataset in solid blue lines, over the Alps. Terrain elevation is also shown. Image from Raupach et al., 2021

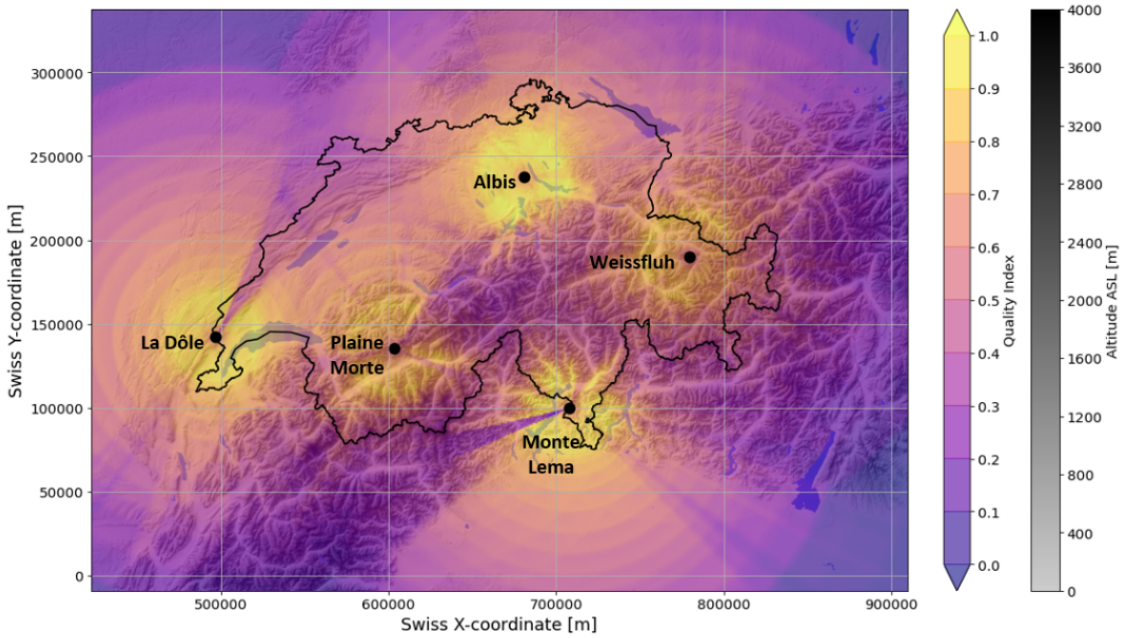

Figure 6 – Map showing the relative quality index [-] over the Alps. The altitude ASL [m] is also displayed. Image from Feldmann et al., 2021.

Furthermore, the digital elevation model Swisstopo (Swisstopo) provides elevation and other topographical data (aspect, slope, Topographic Position Index, altitude) inside Switzerland boundaries. It is extended by data outside the borders of Switzerland (Jarvis et al., 2008). The Topographic Position Index (TPI) compares the elevation of the cell to the mean elevation of the neighbourhood of the cell (Weiss, 2001). A positive TPI value characterises a ridge, i.e. a location that is higher than the average of the neighbourhood. On the opposite, a negative TPI value characterises a valley, i.e. a location that is lower than the neighbourhood. The area is divided in 8 sub-domains : Po Valley, Southern Prealps, Alps, Northern Prealps, Plateau, Jura, Baden-Württemberg (BW) and Allgau. They are illustrated in Figure 5. Their definition is taken from a study by Raupach et al., 2021. Each data point has a label corresponding to the sub-domains it belongs to. Data points that don't belong to any of these domains are labelled with 0. Finally, the relative quality index computed in Feldmann et al., 2021 qualitatively describes observational uncertainties in the Rad4Alp network and is illustrated in Figure 6. It is computed from the theoretical visibility, the minimum and maximum altitude of measurements, spatial resolution and numerical noise, with values from 0 to 1. It is very important to take into account as orography results in beam blockage and a decrease in visibility over the domain.

These preceding steps lead to a multi-year dataset of supercell occurrences in Switzerland. The dataset consists of 6 years of measurements from 2016 to 2021 during the convective season, i.e. from April to October. The timesteps are 5min apart. There are 255 variables in total with 26 TRT variables, 148 rotational variables (74 for a positive rotation and 74 for a negative rotation) and 81 environmental variables. The list of these features and their definition is given in Table XV in Appendix. Table II summarises the number of detections, tracks and active days for thunderstorms and mesocyclones. There are only 6′786 mesocyclones among the 2′525′317 data points in total so the dataset is extremely unbalanced. The two minority classes represent about 0.18% of the dataset each. Table 7 presents the ratio of the number of mesocyclone detections divided by the number of thunderstorm detections in a grid with a 1km resolution. The ratio is only defined in grid cells in which at least one mesocyclonic timestep is detected. This explains why the colourmap is sparse.

Table II – Summary of the number of detections of thunderstorms and mesocyclones in the dataset. For one timestep, there are can be both cyclonic and anti-cyclonic detections, which is why the number of cyclonic and anti-cyclonic detections don't add up to the number of mesocyclonic detections.

	Number of detections	**Number of storm tracks**	**Number of active days**
Mesocyclones	6′786	590	94
• Cyclonic detections	4′644	447	89
• Anti-cyclonic detections	4′704	441	88
Thunderstorms	2′518′600	594′060	216
Total	2′525′317	594′066	216

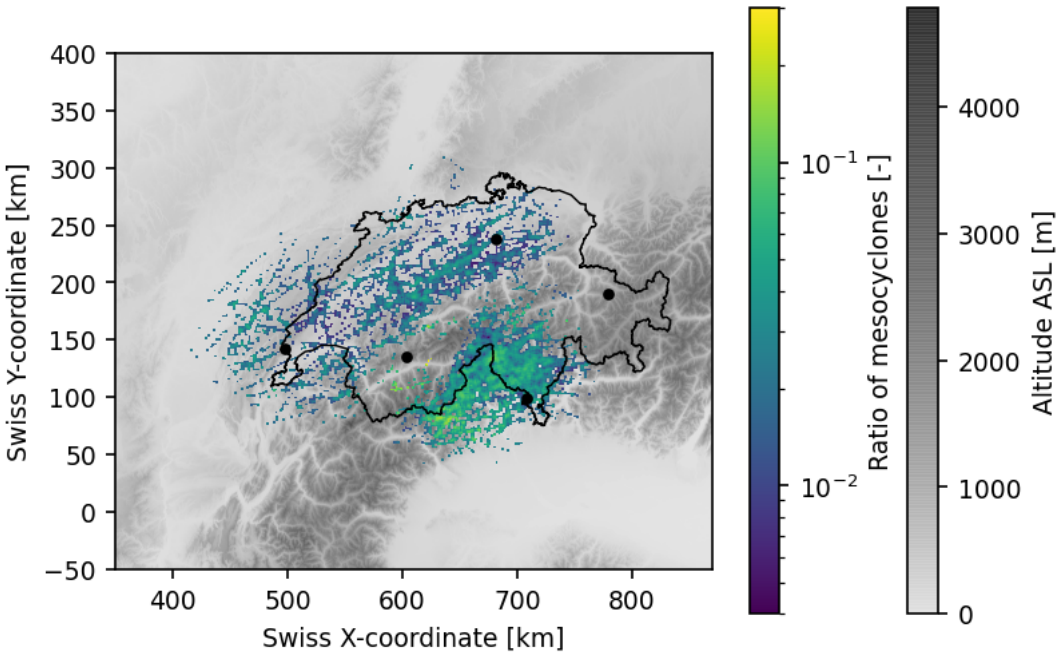

Figure 7 – Ratio of detections of mesocyclones per detections of thunderstorm in a 1km resolution grid.

3 Methods

In this section, we describe the model building process for the two parts of the project. The first step is the construction of a classification model to classify every timestep as mesocyclonic or non-mesocyclonic, using a Random Forest. The second step is the construction of a Convolutional Neural Network to classify every storm track as containing at least one mesocyclone or not.

3.1 Classification of timesteps

In this first part, the aim is the classification of timesteps in two classes, with or without mesocyclone, using a Random Forest. Let $\mathbf{X}$ be the feature matrix and $\mathbf{y}$ the vector of labels:

$$\mathbf{X} \in \mathbb{R}^{n \times p}, \ \mathbf{y} \in \{0, 1\}^n \tag{3}$$

with n the number of data points and p the number of features. The goal is to estimate the probability $P(Y = 1|X)$.

3.1.1 Data preparation

Some transformations are applied to the dataset as pre-processing steps:

1. Positive and negative rotation variables are combined into one single rotation variable. There are therefore two classes: timestep with rotation and without rotation. By doing so, the size of the minority class is slightly increased.
2. Any non-numeric cells and missing values are replaced by the value -9999. This is necessary as we use the scikit-learn implementation (Pedregosa et al., 2011) of the machine learning models and they don't support missing or non-numeric values.
3. Region labels and synoptic weather classifications are encoded to one-hot arrays.
4. Data points with a relative quality index below 0.3 are discarded. To discard data points that have high uncertainty and that can disturb the learning process of the model, the relative quality index is used as a measure of the uncertainty. A threshold can be set to a particular value of the quality index. Figure 8 illustrates the distribution of the ratio of the number of mesocyclones divided by the number of thunderstorms over a 1km grid, smoothed by computing the median over bins. Below 0.3, the ratio is decreasing suggesting detection issues. Timesteps with median quality index < 0.3 are discarded. This also allows keeping a large part of the Swiss domain including the central Alps that have a low relative quality index $(0.4 - 0.5)$ as it can be seen in Figure 6. Finally, the relative quality index statistics are removed from the dataset. They are not given as features to the model as they don't represent environmental data.

Table III summarises the number of mesocyclonic, non-mesocyclonic and total timesteps, after these pre-processing steps. The minority class containing the mesocyclonic timesteps represents 0.62% of the total dataset.

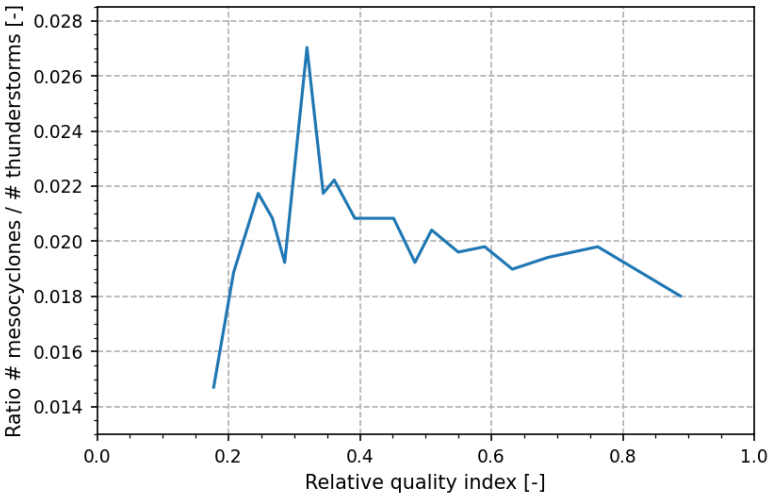

Figure 8 – Ratio [-] between number of mesocyclones and number of thunderstorms as a function of the relative quality index [-].

Table III – Summary of the number of mesocyclonic, non-mesocyclonic and total timesteps, after the pre-processing steps.

Category	**Number of timesteps**
Mesocyclones	6′740
Non-mesocyclones	1′080′552
Total	1′087′292

As mentioned in Section 1.2, supercell thunderstorms sometimes exhibit a hook shape around the centre of rotation. This information is not contained in the dataset. The contour of the cells is an output from the TRT algorithm and can be extracted to compute a shape predictor. One idea for a shape predictor is convexity. It is defined by the ratio of the perimeter of the convex hull divided by the perimeter of the actual contour, as defined below :

$$\text{Convexity} = \frac{\text{P(ConvexHull)}}{\text{P(Contour)}} \tag{4}$$

An example of a convex hull is given in Figure 9, as an orange dashed line around the original polygon in blue. Convexity ranges between 0 and 1. If it equals 1, the shape is completely convex. If the convexity is small, the shape is concave and the assumption is that the cell is more likely to belong to a supercell thunderstorm. To smooth the contour and make the computation more robust, the elliptic Fourier coefficients (Kuhl and Giardina, 1981) of the contour are computed and then used to reconstruct a smoother contour. The perimeters are computed using the smoothed contour. This new feature is added to the dataset.

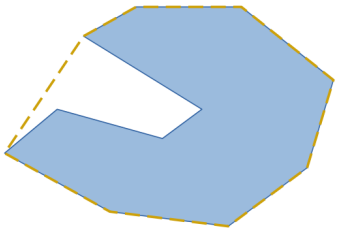

Figure 9 – Example of convex hull, represented by the orange dashed line around the original polygon in blue.

3.1.2 Model building: Random Forest Classifier

A Random Forest (RF, Breiman, 2001) is a non-parametric ensemble learning method for classification or regression. It combines two recent learning methods: Decision Tree (DT) and bagging (bootstrap and aggregation). In this work, we use the binary classification configuration and it is implemented with scikit-learn (Pedregosa et al., 2011).

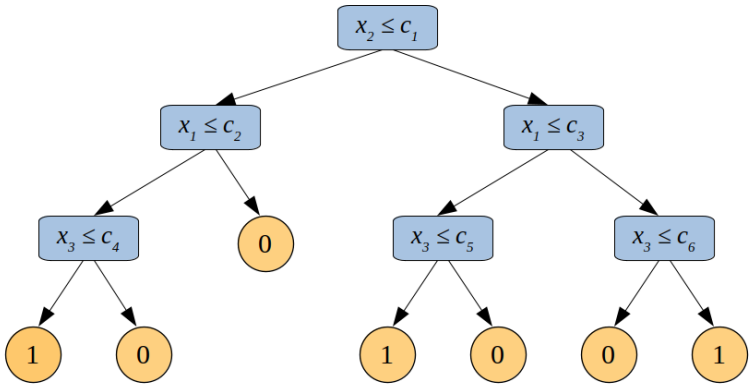

Figure 10 – Diagram showing an example of DT, where the blue rectangles are the nodes, the arrows are the branches and the orange disks are the leaves.

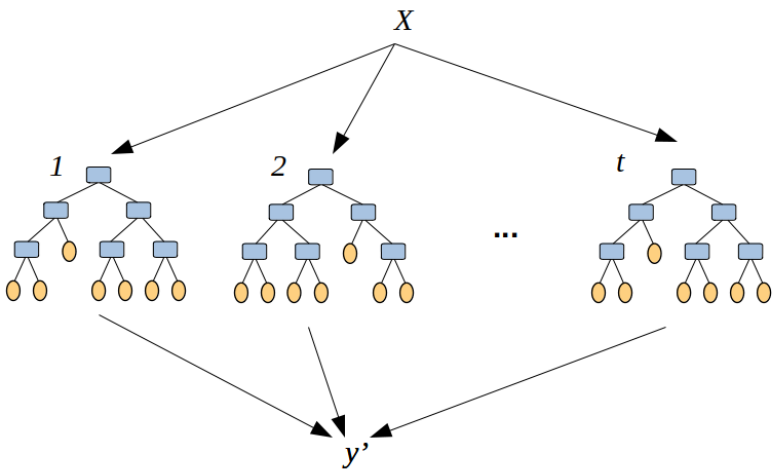

Figure 11 – Diagram showing an example of RF, combining t trees.

A DT is a flow-chart-like tree structure to support decision making. Each internal node is a test on a single feature, using a chosen threshold, as shown in Figure 10. A branch contains the answer to the test: 'yes' or 'no'. The leaves, at the bottom of the tree, are the class labels. The goal is to define the right tests at the right place in the tree, to perfectly separate the class labels. At each node, we need to select the feature that maximises information gain or minimises entropy, i.e. the measure of disorder between the classes. This process is repeated on each new subset recursively. The recursion is stopped when the whole subset at one node has the same labels or when splitting no longer adds information. One issue is that a DT is a weak learner. It is prone to overfitting, i.e. having low bias and very high variance, especially if it is very deep. To overcome this issue, a RF averages the results of many DTs, in order to decrease the variance. The drawback is that the bias can increase and the model will be less interpretable.

A RF grows a combination of t DTs on t datasets using bootstrap sampling, i.e. sampling from the original dataset with replacement. At each node, a subset of features is randomly drawn. The size of this subset is a hyper-parameters that can be optimised during the tuning part. It is typically around $\sqrt{p}$. These two steps allow to guarantee heterogeneity between the DTs, so they are as independent as possible the one from the other. Then, bagging consists of aggregating the output of all the DTs. For classification, the majority vote on all DTs is chosen as the final label. This method reduces the variance through ensemble aggregation and reduces the bias through the depth of the DTs. As a RF is only separating the hyper-space defined by the features in different classes, it is not able to extrapolate. Therefore, the training set must be representative of all possible cases.

Another advantage of the RF is the simplicity of the hyper-parameter tuning compared to more complex models. The hyper-parameters of the RF model which need to be tuned using grid-search are summarised in Table IV. For every possible combination of parameters, a three-fold cross-validation is applied, to get an average score. The best combination of parameters is the one that minimises the accuracy, as defined in Table VIII. The dataset is divided into three subsets: one training set, one validation set and one test set. The training set is used to train the model, the validation set is used as a test set for the cross-validation and to select the decision threshold, and the test set is used to assess the performance of the model on unseen data.

Table IV – List of all hyper-parameters used in the RF algorithm, as well as the range of values that were tested.

Parameter	**Definition**	**Tested values**
t	Number of trees in the RF	50, 100, 150, 200
d	Maximum depth of the DTs	10, 20, 30, 40
m	Number of features randomly picked at a node split	10, 20, 30

3.1.3 Baseline model: Logistic Regression

The performance of the RF is compared to the performance of a baseline model, here a Logistic Regression (LR). It is the simplest machine learning model for classification. It can be seen as a modification of a linear regression in order to output binary results. This model is

based on the logistic function given below:

$$\mathbf{y} = \frac{\exp(\mathbf{z})}{1 + \exp(\mathbf{z})}, \text{ with } \mathbf{z} = \beta_\mathbf{0} + \sum_{j=1}^{p} \beta_\mathbf{j} \mathbf{X_j} \tag{5}$$

with β_j weights defined during the training of the model. The values of each weight is found by minimising a cost function using optimisation methods.

3.1.4 The issue of unbalanced data

The problem considered here is highly unbalanced as supercell thunderstorms are rare. The minority class is the class of most interest but is the smallest one. During training, the model is biased towards the majority class, as it aims to minimise the overall error rather than the error linked to the minority class in particular (He and Ma, 2013). Fortunately, there exist some methods to mitigate this issue. Data-level methods focus on resampling the dataset to modify the ratio between the two classes. The two main methods are random oversampling and undersampling. Random undersampling consists in randomly removing samples from the majority class. This results in a loss of information by discarding useful or important samples and it can increase the variance of the classifier. Random oversampling, on the opposite, acts on the minority class. Data points in the minority class are randomly replicated, up to a certain ratio between the two classes. Oversampling doesn't bring any new information but increases the weight of the minority class by increasing the number of samples inside it. The final ratio between the classes, # mesocyclones / # non-mesocyclones, doesn't necessarily need to reach 1. For example, a ratio of 0.2 is used by Ruiz-Gazen and Villa, 2007. The optimal ratio depends on the dataset and should be selected to maximise the performance. It can therefore be considered as a hyper-parameter. A mix of both oversampling and undersampling can be applied. Algorithm-level methods use the principle of cost-sensitive learning. One assigns a larger cost to misclassification of samples from the minority class. During training, the algorithm is therefore artificially biased towards the minority class.

3.2 Classification of storm tracks

In a second part, the aim is the classification of entire storm tracks in two classes, with or without at least one mesocyclonic timestep. Let $\mathbf{X}$ be the feature tensor and $\mathbf{y}$ the vector of labels:

$$\mathbf{X} \in \mathbb{R}^{n \times p \times t}, \ \mathbf{y} \in \{0, 1\}^n \tag{6}$$

with n the number of data points, p the number of features and t the number of timesteps in the track. The goal is to estimate the probability $P(Y = 1|X)$.

3.2.1 Data preparation

Some transformations are applied to the dataset as pre-processing steps :

1. Positive and negative rotation variables are combined into one single rotation variable. There are therefore two classes: storm with rotation and without rotation, as in Section 3.1.

2. Any non-numeric cells and missing values are replaced by the value −9999, as in Section 3.1.
3. Region labels and synoptic weather classifications are encoded to one-hot arrays as in Section 3.1. Also, cyclical variables like 'hour' and 'time of day' are encoded with sine and cosine functions in order to explicit the cyclical variations.
4. Storm tracks with a maximum RANK over the entire life-cycle that is < 25 are discarded. This allows to focus on severe thunderstorms only and mitigate the unbalance.
5. Storm tracks with a maximum median relative quality index over the entire life-cycle that is < 0.3 are discarded. Afterwards, the relative quality index statistics are removed from the dataset. They are not given as features to the model as they don't represent environmental data.
6. Min-max scaling between 0 and 1 is applied on feature columns. Normalisation parameters are computed on the training set and are then used to normalise training, validation and testing sets. This prevents the testing set from influencing the training procedure.
7. Tracks are padded at the end, to the size of the longest track, 129, with the value 0. Therefore, the data for each sample has the same shape: $p \times t$.

Table V summarises the number of mesocyclonic, non-mesocyclonic and total storms, after these pre-processing steps. The mesocyclonic storms represent 19.8% of the dataset.

Table V – Summary of the number of mesocyclonic, non-mesocyclonic and total storms, after the pre-processing steps.

	Number of timesteps
Mesocyclones	373
Non-mesocyclones	1′514
Total	1′887

3.2.2 Model building: Convolutional Neural Network

The model used for the classification of storm tracks is a Convolutional Neural Network (CNN, Fukushima and Miyake, 1982), a type of Artificial Neural Network (ANN). ANNs are collections of connected units, called neurons, inspired by the structure of neurons in a biological brain. An example of a simple ANN is illustrated in Figure 12. ANNs are composed of different layers. The input data goes from the input layer to the output layer by travelling through a certain number of hidden layers. The output of each neuron is computed by a non-linear function of the sum of the inputs and weights. The output of node j in layer l is given by:

$$x_j^{(l)} = \phi\left(\sum_i w_{i,j}^{(l)} x_i^{(l-1)} + b_j^{(l)}\right) \tag{7}$$

where $x_i^{(l-1)}$ are the inputs coming from nodes from the previous layer, $b_j^{(l)}$ is the bias term corresponding to the node j, $w_{i,j}^{(l)}$ are the weights corresponding to each of the nodes in the layer l and ϕ is the activation function. In classic ANNs, the hidden layers are generally dense layers, i.e. each neuron in one layer is connected to all neurons in the next layer. CNNs take a different

approach as a neuron in one layer only influences a limited number of neurons in the next layer. This method is useful for images or time-series because in these cases, the local dependencies are the most important. This type of layer is called convolutional layer as the output is given by the convolution of the input and a filter:

$$x^{(1)}[n] = \sum_{k=0}^{N-1} f[k]x^{(0)}[n-k] \tag{8}$$

with f a local time-invariant filter. The appropriate type of filter for a particular task, e.g. low pass or high pass filter, can be chosen during the hyper-parameter tuning. This convolutional layer reads sequences of input data and extracts features, followed by a pooling layer that reduces the dimensions by combining the outputs of neuron clusters. The size of the neuron clusters can be considered as a hyper-parameter. This structure is much more sparse and local, and using convolution instead of general matrix multiplication implies that there are fewer parameters to determine. In the end, CNNs are capable of extracting translation-invariant patterns and capturing different hierarchies of patterns (Chollet, 2018). Once the CNN is built, it can be trained on the training dataset. The tuning of the weights in the layers aims to minimise the value of the loss function with a method called backward propagation of errors (or back-propagation). Back-propagation finds the derivatives of the nodes' equations from the final layer to the initial one. The derivatives of each layer are multiplied to compute the derivatives of the initial layers, following the chain rule.

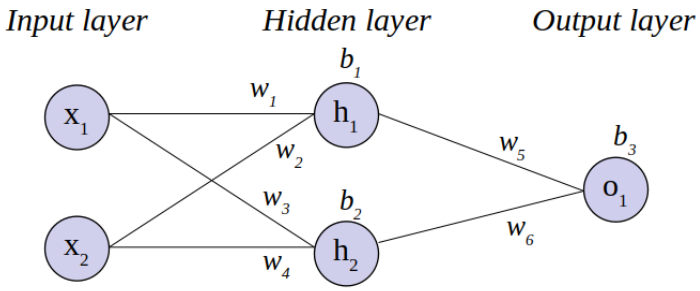

Figure 12 – Diagram showing a basic example of ANN.

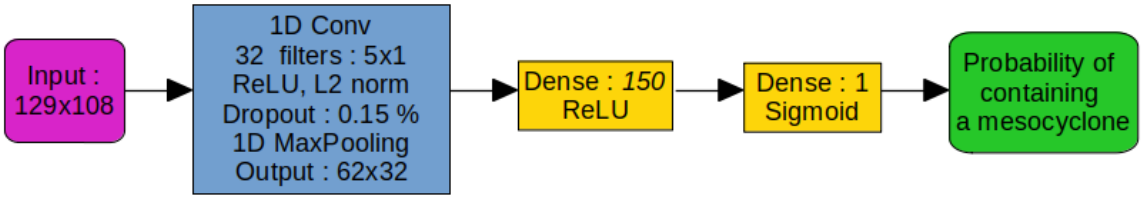

Figure 13 – Schematic diagram of a CNN like the one used in this work, with kernel size $k = 5$, number of filters $f = 32$, dropout $d = 0.15$ and number of neurons $u = 150$.

In this project, the input data consists of time-series that are one dimensional so a CNN with 1D filters is used. The model is implemented with Keras (Chollet et al., 2015). The structure of the CNN used is illustrated in Figure 13. A convolution step is applied. This step is composed of one convolutional layer, followed by a dropout and a max-pooling layer. The max-pooling layer uses the maximum value of each local cluster of neurons in the feature map.

In the convolutional layer, the rectified linear unit (ReLU, Nair and Hinton, 2010) is used as an activation function. It is the most commonly used activation function in CNNs and is defined by:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \tag{9}$$

Using the ReLU instead of the sigmoid or hyperbolic tangent in hidden layers allows avoiding saturation and the vanishing gradient problem. The vanishing gradient problem can happen during back-propagation if the gradient becomes extremely small and slows down or even stops the training of the model. Hyperbolic tangent has derivative values between 0 and 1, and when small derivatives are multiplied together for a large number of layers, the gradient can become very small. The He Uniform weight initialisation (He et al., 2015), associated with the ReLU function, is used. It draws initial values for the weights of the layers from a uniform distribution within $[-l, l]$, where $l = \sqrt{(6/n_{in})}$ and n_{in} is the number of input units in the weight tensor, as defined by He et al., 2015. To prevent overfitting, a dropout layer is added before the max pooling. Dropout is a regularisation method that randomly drops out some nodes in the layer. By making the network more sparse, it lowers its complexity. Another regularisation method used is L2, which consists in adding a regularisation term in the loss function. This term is defined as the L2-norm of the weights matrix. After the convolutional step, the features are flattened to one vector and pass through a fully connected layer before the output layer. The dense layer before the output layer provides a buffer between the learned features and the output, to facilitate the interpretation of the features before the prediction. The task is a binary classification so the model must output probabilities between 0 and 1. Therefore the activation function used in the output layer is a sigmoid function, defined by:

$$\phi(x) = \frac{1}{1 + \exp(-x)} \tag{10}$$

During training, the model minimises the loss on the training data by updating the weights in the convolutional layer using stochastic gradient descent. The loss function used is the binary cross-entropy, that measures the difference between two probability distributions :

$$\text{crossentropy} = -\frac{1}{N}\sum_{i=1}^{N} [y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \tag{11}$$

where p_i is the predicted label by the model and y_i is the true label. In each epoch, multiple batches of training examples are created and used as input data for the CNN. The hyper-parameters listed in Table VI are tuned during grid-search.

Table VI – List of all hyper-parameters used in the CNN, as well as the range of values that were tested.

Parameter	**Meaning**	**Tested values**
k	Kernel size	5, 9, 11
f	Number of filters	16, 32, 64
d	Dropout rate	0.1, 0.15, 0.2
u	Number of neurons in the dense layer	50, 100, 150

3.2.3 Baseline models

A basic meteorological baseline is created to have a reference point for the performance of the CNN. In this simple meteorological model, a storm is classified as mesocyclonic if the maximum RANK value over the entire life-cycle is superior to a certain threshold. The best threshold is selected by maximising the Critical Success Index value on the validation set. Then, this threshold is used to classify storms in the test set. Moreover, a basic ANN is created to serve as a machine learning baseline. It is composed of one dense layer with a number of neurons that can be optimised.

3.3 Model evaluation

After a cross-validation step to tune the hyper-parameters and the decision threshold τ, the performance of the model is evaluated using unseen data. In classification tasks, the output is categorical. There exist some specific metrics to evaluate the performance of classification models. The confusion matrix as shown in Table VII, gives a cross-classification of the predicted class by the true class. In this matrix, the number of True Positives (TP) is the number of observed mesocyclonic timesteps that are predicted as mesocyclonic, the number of True Negatives (TN) is the number of observed non-mesocyclonic timesteps that are predicted as non-mesocyclonic, the number of False Positives (FP) is the number of observed non-mesocyclonic timesteps that are predicted as mesocyclonic and the number of False Negatives (FN) is the number of observed mesocyclonic timesteps that are predicted as non-mesocyclonic. Many metrics computed from the confusion matrix emphasise one property or another. The definitions and formulas of the main metrics used in this work are reminded in Table VIII.

Table VII – Confusion matrix for a binary classification task.

		Predicted class	
		Class 1	Class 0
Actual class	Class 1	Hit TP	Miss FN
	Class 0	False alarm FP	Correct rejection TN

The accuracy is the most common metric. It takes into account all the well predicted samples. It is not representative of the real performance of the model when there is a strong unbalance between the two classes, as the high number of TN artificially increases it. The Receiver Operating Characteristic curve (ROC curve, Mason, 1982) shows the value of recall as a function of the False Alarm Rate (FAR) for each value of the decision threshold. Recall evaluates the number of TP among all the points that are positive in reality whereas the FAR evaluates the number of FP among all the points that are negative in reality. The Area Under the ROC Curve (AUC) synthesises the performance with a unique number. An AUC score higher than 0.5 has more skill than a random prediction (AUC = 0.5) and a score of 1 indicates the ability to perfectly discriminate between positive and negative events. The goal of a ROC curve is to show how well a model can distinguish between two classes. If there are wide disparities

in the cost of FN and FP, it may be critical to give more importance to one type of classification error. Similarly to accuracy, AUC isn't a useful metric for this type of optimisation.

Another possibility is to plot recall as a function of precision when varying the value of the decision threshold, which gives a performance diagram (Roebber, 2009). Precision counts the number of TP among all the points classified as positive. It evaluates how the model can correctly predict the rare events and how many truly relevant results are returned. It does not provide any information on how the model performs in the majority class. The F-scores combine precision and recall into a single metric. F1-score is defined as the harmonic mean of precision and recall. F2-score is a version of F1-score that gives less weight to precision and more weight to recall. Therefore, F2 gives more importance to minimising FN than FP. In this work, FN, i.e. predicting that there is no mesocyclone whereas in reality there is one, is considered as a more serious error than FP, so F2-score is preferred.

The reliability diagram (Hsu and Murphy, 1986) illustrates the difference between the forecast probability and the observed relative frequency. For a perfectly reliable forecast, the curve would closely follow the diagonal. It means that every time the predicted probability of an event is p then this event occurs a fraction p of the time. If the forecast probability is greater, respectively lower than the observed relative frequency then the model is over-forecasting, respectively under-forecasting. The Brier score (BS) measures the mean squared error in probability space:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(f_t - o_t)^2 \tag{12}$$

where f_t is the forecast probability of event t and o_t is its observed relative frequency. For a perfectly reliable forecast, $BS = 0$. The Brier Skill Score (BSS, Brier, 1950) evaluates the skill of the forecast compared to a reference forecast, generally the climatology :

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{\mathrm{BS}_{\mathrm{ref}}} \tag{13}$$

The best possible score is BSS = 1. If BSS < 0, the model is worse than the reference model. Finally, we can define the Matthews Correlation Coefficient (MCC, Matthews, 1975) that evaluates the quality of a binary classification. It can be interpreted as a correlation coefficient between the observed and predicted binary classifications and scales between −1 and +1, with +1 for perfect prediction, 0 for random and −1 for total disagreement. It is considered one of the most informative metrics because it takes into account the ratios of the four categories of the confusion matrix.

3.4 Feature importance and interpretability of the model

Investigating feature importance and interpreting results of the model are crucial as they bring insight into the learning process of the model and help to understand the physical relations. It can eventually lead to a better understanding of how to improve the model and therefore more trust. To do so, we use SHapley Additive exPlanations (SHAP, Lundberg and Lee, 2017), which is a mathematical method to explain individual predictions of a model. It combines two interpretability methods : Shapley values (Shapley, 2016) and Local Interpretable Model-agnostic Explanations (LIME, Ribeiro et al., 2016). Shapley values estimate the contribution of each feature to the final prediction using coalition game theory (Branzei et al., 2008). This

Table VIII – List of the definition and formula of performance metrics used in this work. We define n_{meso} the number of mesocyclones, n_{nomeso} the number of non-mesocyclones and $R_{\text{meso}} = \frac{n_{\text{meso}}}{n}$.

Long name	**Formula**	**Value range**	**No skill value**	**Optimal skill value**	**Definition**
Accuracy	$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$	$[0, 1]$	0.5	1	How many good predictions among all predictions.
Recall	$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$	$[0, 1]$	0.5	1	How many good predictions among all events in class 1.
False Alarm Rate	$\text{FAR} = \frac{\text{FP}}{\text{FP+TN}}$	$[0, 1]$	0.5	0	How many wrong predictions among all events in class 0.
AUC	Area under ROC curve	$[0, 1]$	0.5	1	How well a model discriminates between two classes.
Precision	$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$	$[0, 1]$	R_{meso}	1	Probability of correct detection of positive values.
F1-score	$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}}$	$[0, 1]$	$\frac{R_{\text{meso}}}{R_{\text{meso}}+0.5}$	1	Harmonic mean of precision and recall.
F2-score	$\text{F2} = \frac{5 \cdot \text{Precision} \cdot \text{Recall}}{4 \cdot \text{Precision+Recall}}$	$[0, 1]$	$\frac{5R_{\text{meso}}}{8R_{\text{meso}}+1}$	1	Harmonic mean of precision and recall, with more weight for recall.
Critical Sucess Index	$\text{CSI} = \frac{\text{TP}}{\text{TP+FP+FN}}$	$[0, 1]$	$\frac{0.5n_{\text{meso}}}{n_{\text{meso}}+0.5n_{\text{nomeso}}}$	1	How many observed and forecast events were correctly predicted.
Matthews correlation coefficient	$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{((TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN))}}$	$[-1, 1]$	0	1	Correlation coefficient between observation and prediction.

method describes a way to distribute the value of the prediction among the features. If we look at the prediction as a game, each feature represents one player and the difference between this particular prediction and the average prediction for all instances is the gain. The Shapley value of a feature is defined as the average contribution of this feature to the prediction, in all possible coalitions, i.e. combinations of the other features. The Shapley value is not the difference in prediction we would obtain if we remove the feature from the model. LIME aims to create local linear models to approximate the individual predictions of the model of interest. To do so, new data points that are perturbed versions of the original data point are created. The predictions for these new points are computed using the ML model. These data points are then weighted proportionally to their distance to the original point, creating a local linear model, that can explain the original data point. SHAP combines both principles by defining an explanation model g of the original model f, that uses Shapley values:

$$f(x) = g(x') = \phi_0 + \sum_{i=0}^{M} \phi_i x'_i \tag{14}$$

where ϕ_i are the Shapley values, x' is a simplified input with some omitted features. The mathematical definition of ϕ_i is:

$$\phi_i = \sum_{S \in N \setminus i} \frac{|S|!(M - |S| - 1)!}{M!} [f(S \cup \{i\}) - f(S)] \tag{15}$$

where N is the set of all input features, S is the set of non-zero indexes in x' and M is the number of features. The complexity of computing exact Shapley values grows with 2^M. Lundberg et al., 2018 provide an efficient estimation approach for tree-based models, called TreeSHAP. This method is used for the RF.

For one sample, in particular, the SHAP value represents how much knowing the value of this feature changes the prediction of the model. SHAP is a method for local interpretability, however, by averaging the SHAP values for all the predictions, we can obtain a global explanation of the model. A feature importance ranking displays the features which have large absolute SHAP values, in descending order. A summary plot brings more information than the feature importance graph as it takes into account both feature importance and feature effect. Therefore, it is possible to assess the impact of one feature value on the predicted probability, in particular, if it is a positive or negative impact. Finally, a dependence plot shows the SHAP value as a function of the feature value and allows to study in more detail the correlation between feature and SHAP value.

4 Results and discussion

4.1 Classification of timesteps

4.1.1 Mitigation of unbalance

The first step is to apply various methods to mitigate the unbalance of the dataset, as described in Section 3.1.4. Figure 14 shows the performance diagram for RFs with various methods to mitigate the unbalance of the dataset, using the test data. Each point in the curve corresponds to one probability threshold. The colourmap shows the corresponding value of the CSI. It is used to select a decision threshold. The best point is situated in the top-right corner and the worst point in the bottom-left corner. The decision threshold that maximises the value of the CSI on the validation set is marked by a triangle. The LR is used as a machine learning baseline. A RF without any technique for mitigating the unbalance and other models with class weights and random resampling with different ratios are compared. We can see that all the RF methods perform better than the LR. Furthermore, mitigation methods that undersample or add class-weight perform slightly worse than the RF without mitigation methods. On the opposite, oversampling is very slightly increasing the performance.

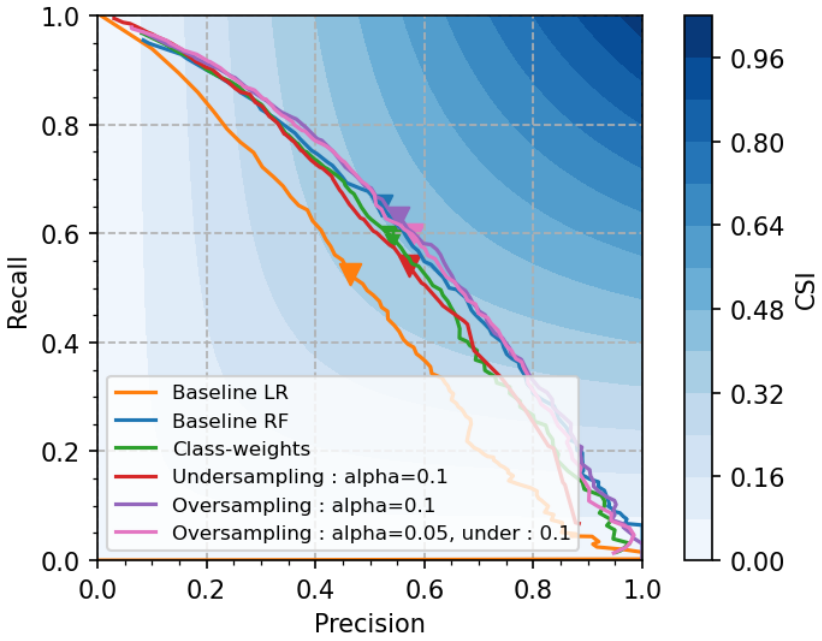

Figure 14 – Performance diagram for the LR and RFs with various resampling methods, evaluated on test data. The colormap shows the corresponding value of the CSI. The triangle corresponds to the decision threshold that maximises the CSI on the validation data.

Table IX – Confusion matrices showing the ratios of TP, TN, FP and FN [%], for the LR and RFs with various resampling methods. They are evaluated on the test set.

Model	Predicted	Observation	
		1	**0**
LR Baseline	**1**	0.33	0.38
	0	0.29	99
RF Baseline	**1**	0.4	0.37
	0	0.22	99
RF with weights	**1**	0.37	0.31
	0	0.25	99
RF with undersampling	**1**	0.34	0.25
	0	0.28	99
RF with oversampling	**1**	0.39	0.32
	0	0.23	99
RF with over and undersampling	**1**	0.37	0.27
	0	0.25	99

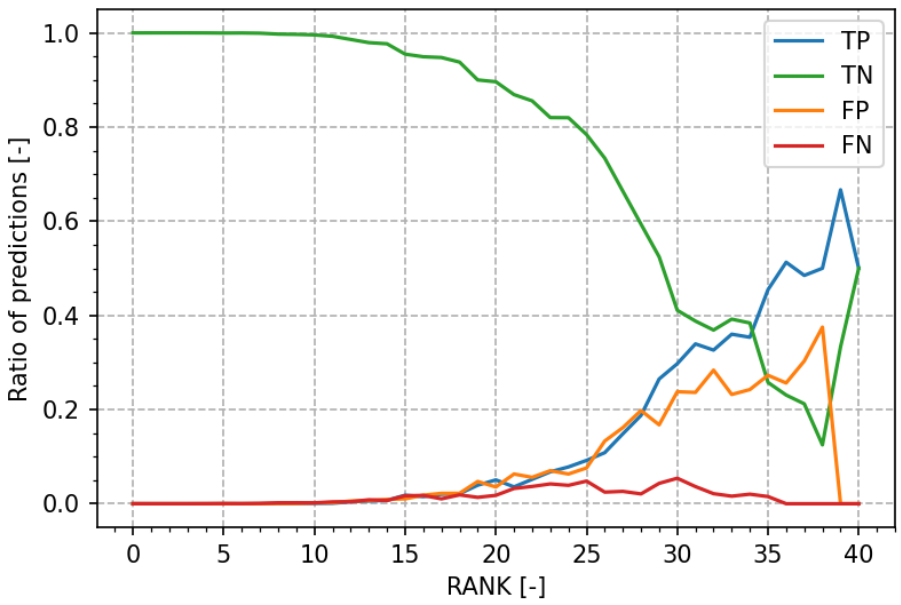

Figure 15 – Distribution of the ratios of TP, TN, FP and FN predictions depending on the RANK of the data point, for the baseline RF.

Table XII illustrates the confusion matrices for each of the models, allowing to inspect the results in more detail. They are created using the decision thresholds shown in Figure 14. For the LR baseline, there are more FP than TP and almost as many FN as TP. For the RF baseline, there are almost as many FP as TP and almost twice less FN than TP. Mitigation methods make the number of FP decrease but the number of FN tends to increase. In the end, both FP and FN are still high for hazardous events like supercells. Therefore, it seems that resampling methods are not helping to improve the performance of the RF. Figure 15 displays the distribution of the predictions (TP, TN, FP and FN) of the baseline RF, as a function of the value of the RANK. On one hand, the good predictions, TP and TN, mostly occur for severe supercells and weak thunderstorms respectively. On the other hand, the wrong predictions, FP and FN, mostly occur for severe thunderstorms and moderate supercells respectively. The persistently high ratios of FP and FN across the different models might suggest the presence of too many mislabelled points in the dataset that are disturbing the learning process of the model: events labelled as mesocyclones by the MDA algorithm but that are not mesocyclones, or the opposite.

4.1.2 Filtering on RANK

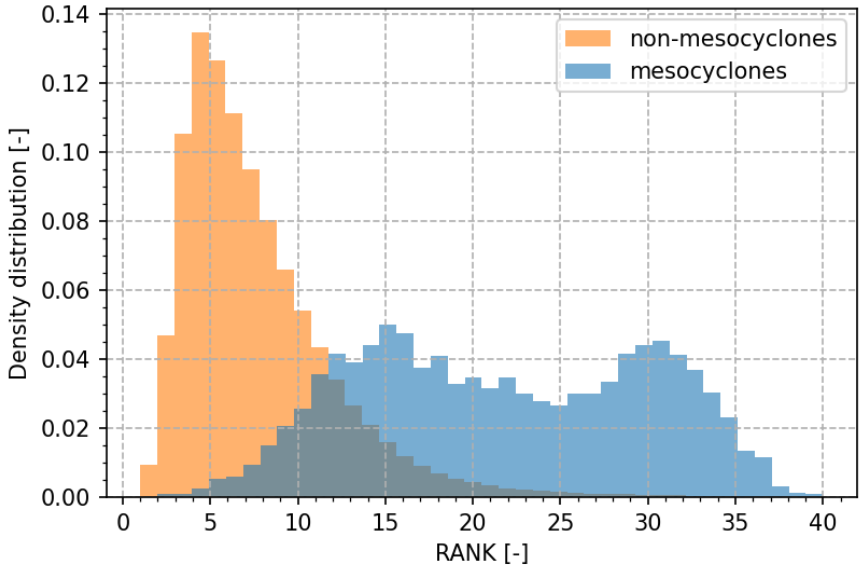

Figure 16 – Density distribution of RANK [-] values for mesocyclones and thunderstorms.

An approach to reduce the impact of mislabelled data points is to focus on severe thunderstorms and mesocyclones. A reasonable hypothesis is that the severe mesocyclonic thunderstorms are less likely to be mislabelled as non-mesocyclonic as the radar signature would be clearer. Furthermore, these severe supercells are the ones that forecasters are the most interested in, as they are the most hazardous. Figure 16 shows the density distribution of the RANK of thunderstorms and mesocyclones. Thunderstorms' distribution is quite typical, with the majority of thunderstorms being weak thunderstorms and the density decreasing for more severe thunderstorms. Mesocyclones' distribution is more unusual, with two peaks around 15 and 32. Our hypothesis leads us to think that the left part of the mesocyclones' distribution would contain moderate mesocyclones that would be more easily missed by the model. Using the categorisation of thunderstorms given in Table I, we can focus on severe and very severe thunderstorms only, by filtering and keeping only data points that have a RANK > 25. The

updated number of mesocyclonic, non-mesocyclonic and total timesteps is given in Table X. This filtering discards a lot of weak and very weak non-mesocyclonic thunderstorms so the ratio between the two classes increases to 27.4%, which helps mitigate the unbalance of the dataset.

Table X – Summary of the number of mesocyclonic, non-mesocyclonic and total timesteps, after the filtering using the RANK.

Category	**Number of timesteps**
Mesocyclones	2′703
Non-mesocyclones	7′173
Total	9′876

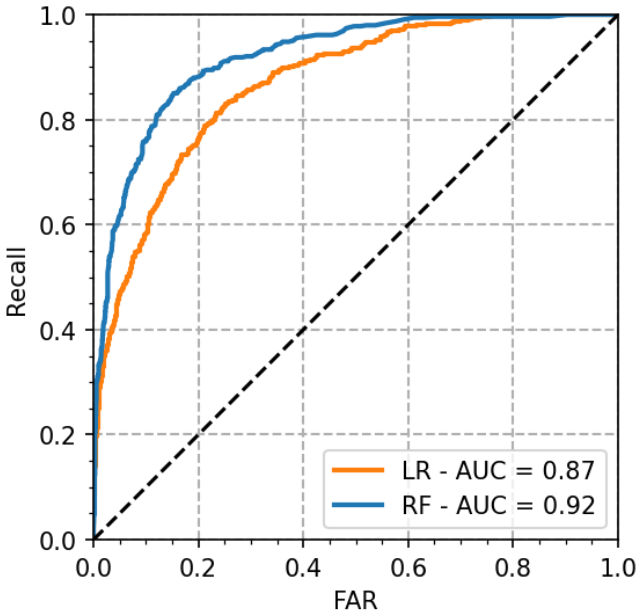

Figure 17 – ROC curve and AUC value for the RF and the LR. The points from low decision thresholds are in the top-left corner and the points from high decision thresholds are in the bottom-right corner. The black dashed line indicates the performance of a no-skill model outputting random predictions.

With this new dataset containing severe thunderstorms only, the performance of the RF is compared to the performance of a LR. The RF is tuned using grid-search on training data and the best hyper-parameters are : $t = 200$, $d = 20$ and $m = 30$. Figures 17 and 18 show the performance of the RF and the LR on test data. AUC is significantly higher for the RF and it consistently performs better across all probability thresholds. AUC for the RF is above 0.9, which is considered high because the computation of FAR takes into account TN which are very numerous. The performance diagram ignores the number of TN and therefore moderately widens the gap between the two models, as the CSI value is 1.2 times higher for the RF than for the LR. CSI is a metric that is sensitive to event frequency, unlike AUC. To get a high CSI value, we need a high recall and a high precision. It means that the model must correctly identify a large portion of mesocyclones without creating too many FP, which is very difficult when the event frequency is low. Furthermore, we can see on the graph that the decision thresholds selected on the validation set, $\tau_{RF} = 0.374$ and $\tau_{LR} = 0.364$, also almost maximise the CSI on the test set. This means that the validation set is representative of the test set.

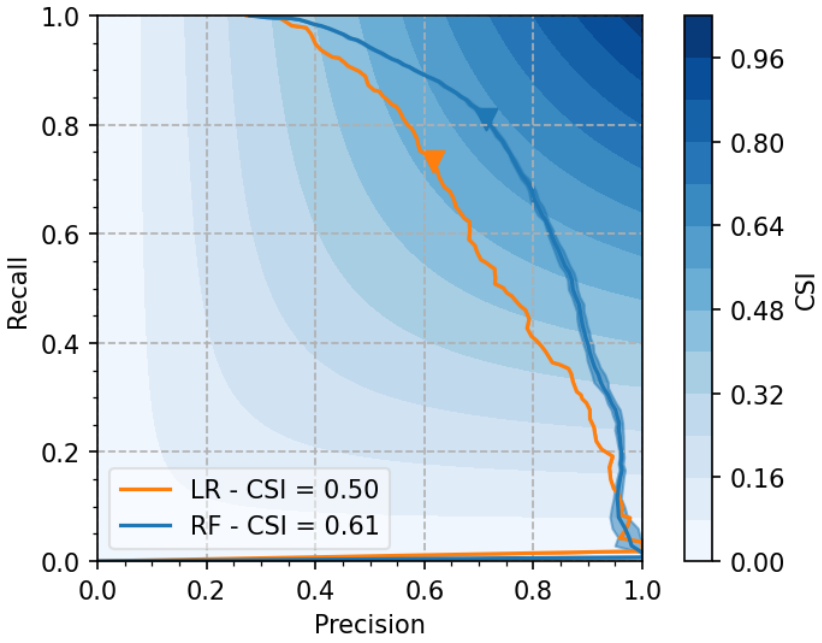

Figure 18 – Performance diagram for the RF and the LR on test data. The dark lines show the mean and light shadings show the standard deviation. The colormap shows the corresponding value of the CSI. The triangles correspond to the decision thresholds that maximise CSI on the validation data. They are $\tau_{RF} = 0.374$ and $\tau_{LR} = 0.364$.

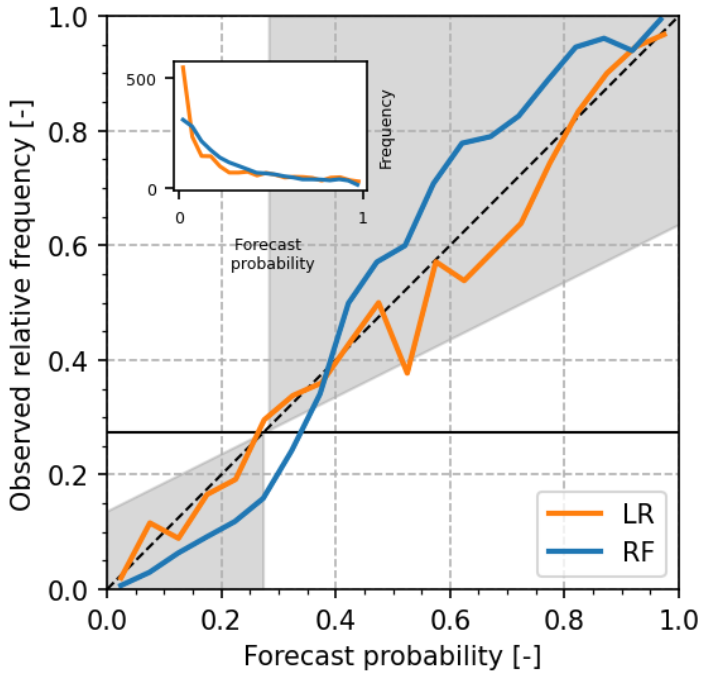

Figure 19 – Reliability diagram for the RF and the LR. The dashed diagonal indicates a perfectly reliable forecast. The horizontal line indicates the climatological probability, representing no resolution. The light grey area indicates the forecasts with positive BSS.

The reliability diagram represented in Figure 19 shows the observed relative frequency as a function of the forecast probability, for the RF and the LR. The dashed diagonal indicates a perfectly reliable forecast. The horizontal line indicates the climatological probability of the event, which corresponds here to the ratio between the two classes. The light grey area indicates the area containing forecasts with a positive BSS. Both curves are almost always situated in this area. The RF is under-forecasting events between 40 and 100% (when the curve is above the diagonal) and over-forecasting events between 0 and 40%. The model is, therefore, over-forecasting rare events which is what is generally asked from a model predicting rare events. The LR curve is less smooth but it follows more closely the diagonal.

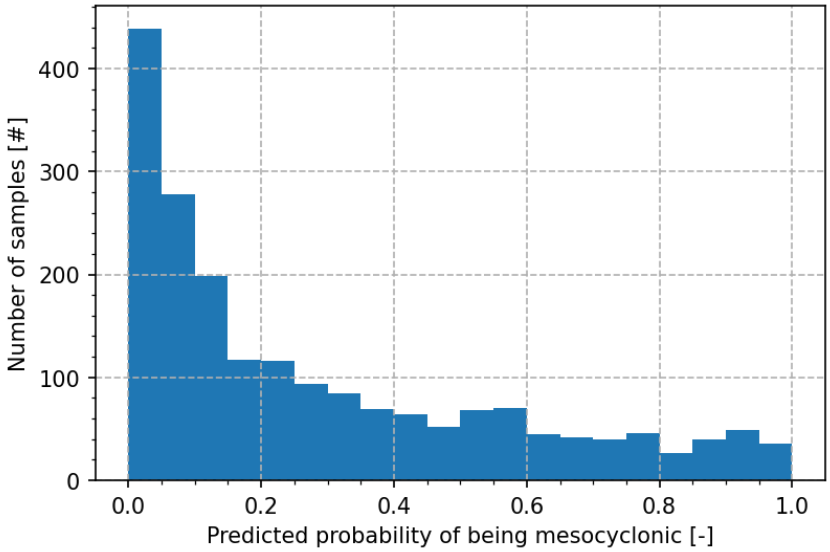

Figure 20 – Distribution of the predicted probability that the timestep is mesocyclonic, output by the RF on the test set.

Table XI – Confusion matrices showing the ratios of TP, TN, FP and FN [%], for the LR and the RF on the test set.

Model	**Predicted**	**Observation**	
		1	**0**
LR	**1**	20	12
	0	8	60
RF	**1**	22	9
	0	5	64

Figure 20 displays the distribution of the predicted probability that the timestep is mesocyclonic, output by the RF on the test set. The ideal distribution would be two peaks around 0 and 1, corresponding to non-mesocyclones and mesocyclones, without any predictions in between. In the test set, there are 1435 non-mesocyclonic samples and 541 mesocyclonic samples. In this figure, a peak of predictions can be observed around 0. There are 720 predictions of probability < 0.1. Then, the number of predictions is decreasing with increasing probability and there

is no clear peak around 1. The high peak around 0 is emphasised by the higher amount of non-mesocyclonic timesteps compared to mesocyclonic ones, even though this unbalance is now reasonable. This distribution means that the model can classify many timesteps as non-mesocyclonic with very high confidence but cannot classify many mesocyclonic timesteps with the same confidence. As a lot of predicted probabilities are situated around 0.5, a small change in the value of the decision threshold might considerably impact the ratios of TP, TN, FP and FN. Table XI illustrates the confusion matrices for the LR and the RF, using the decision thresholds selected in Figure 18. When going from the LR to the RF, the ratio of TN increases from 60% to 64% and the ratio of TP increases from 20% to 22%. The ratio of FP is 1.3 times lower and the ratio of FN is 1.6 times lower for the RF than for the LR. Table XII summarises the values of various metrics for the RF and the LR. The RF is performing better than the LR for all the metrics and the difference is always statistically significant.

Table XII – Evaluation metrics for the RF and the LR. The scores are averaged over 10 runs.

Model	**AUC**	**Precision**	**Recall**	**F2**	**MCC**	**CSI**
LR	0.875	0.615	0.731	0.705	0.532	0.502
RF	0.924	0.711	0.811	0.789	0.661	0.610

The spatial distribution of the predictions of the RF is illustrated in Figure 21. There are no obvious spatial biases. A higher concentration of points can be found around Monte Lema radar and is explained by the higher concentration of thunderstorms in this region as shown by Feldmann et al., 2021. A study of the minimum elevation scan distribution depending on the type of predictions (TP, TN, FP, FN) doesn't highlight any biases in altitude.

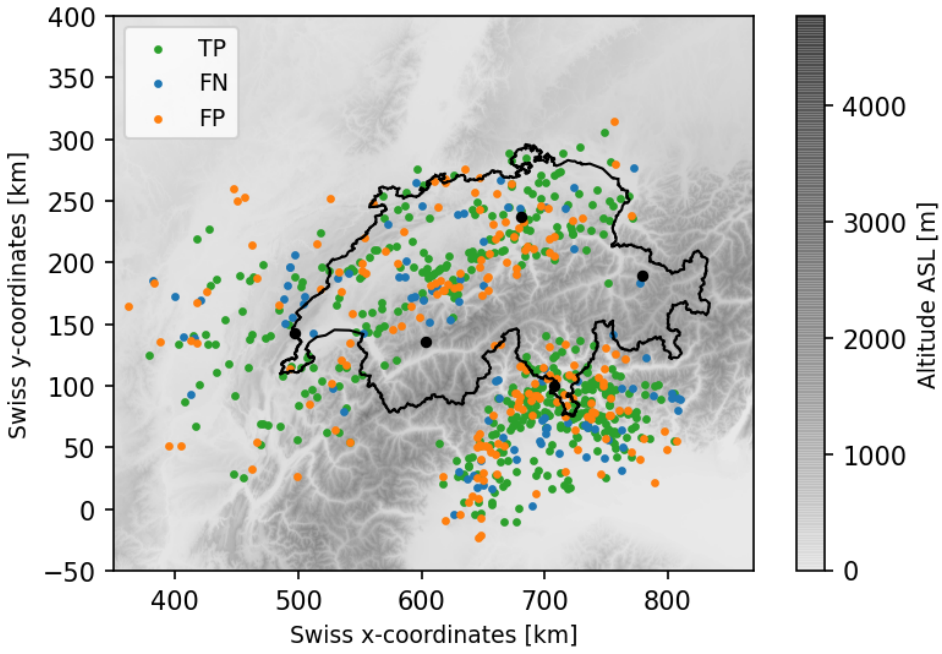

Figure 21 – Map showing the spatial distribution of the TP, FP and FN predictions of the RF. TN predictions are not shown because they are too numerous. The altitude ASL [m] is also shown.

4.1.3 Addition of a new feature: convexity

As explained in Section 3.1.1, convexity is a new feature that is computed and added to the dataset. A RF is trained on this updated dataset and its performance is compared to the one of a RF trained on the old dataset and a LR trained on the updated dataset too. Figure 22 displays the performance diagram. The CSI value for the RF using the convexity feature is 0.01 higher than the CSI value for the RF that doesn't use it. This difference is not statistically significant. The RF with convexity still performs better than LR with convexity. Therefore, the addition of convexity as a new feature doesn't improve the performance of the model. The main hypothesis is that in Switzerland, because of the terrain, the signatures on the radar images are less clear and the hook-echo cannot be seen as often as in areas with flat terrain such as in the USA.

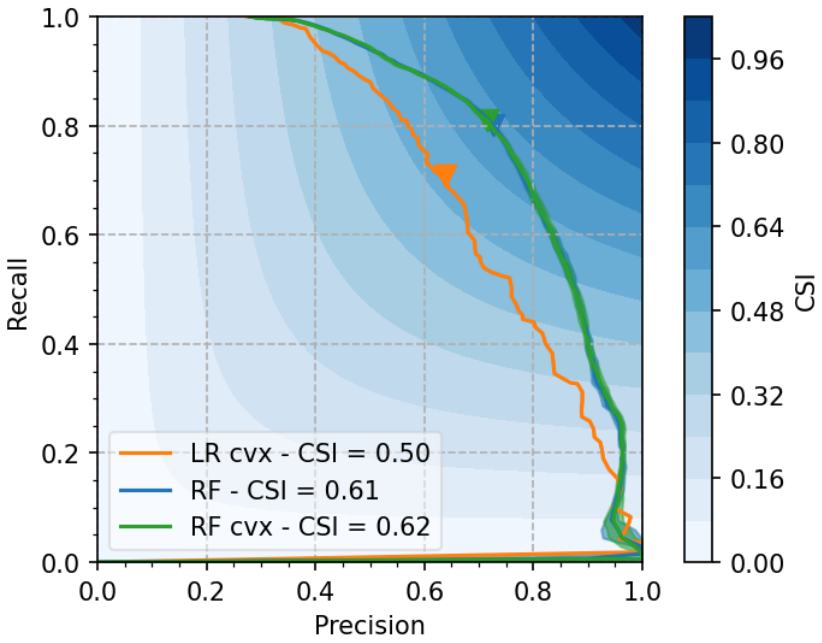

Figure 22 – Performance diagram for the LR and RF with the addition of the convexity in the features. The dark lines show the mean and light shadings show the standard deviation. The colourmap shows the corresponding value of the CSI. The triangles correspond to the decision thresholds that maximise CSI on the validation data.

4.1.4 Feature importance and interpretability

Feature importance is assessed using the SHAP method to construct a feature importance graph as displayed in Figure 23. It displays the 20 most important features in decreasing order. For each feature, the composition ratio is the relative importance compared to the other features. The cumulative ratio doesn't reach 100% as some less important features are not shown in the graph. We can see that the area is the most important feature, far above the others. In the top features, we can also find statistics for the reflectivity, shape descriptors such as the axis of the cell ellipse, x and y velocities and statistics for the slope and TPI. Some features are related, such as the area and ellipse descriptors, the x and y velocities or the reflectivity statistics. These features describe similar physical quantities and are therefore correlated with each other.

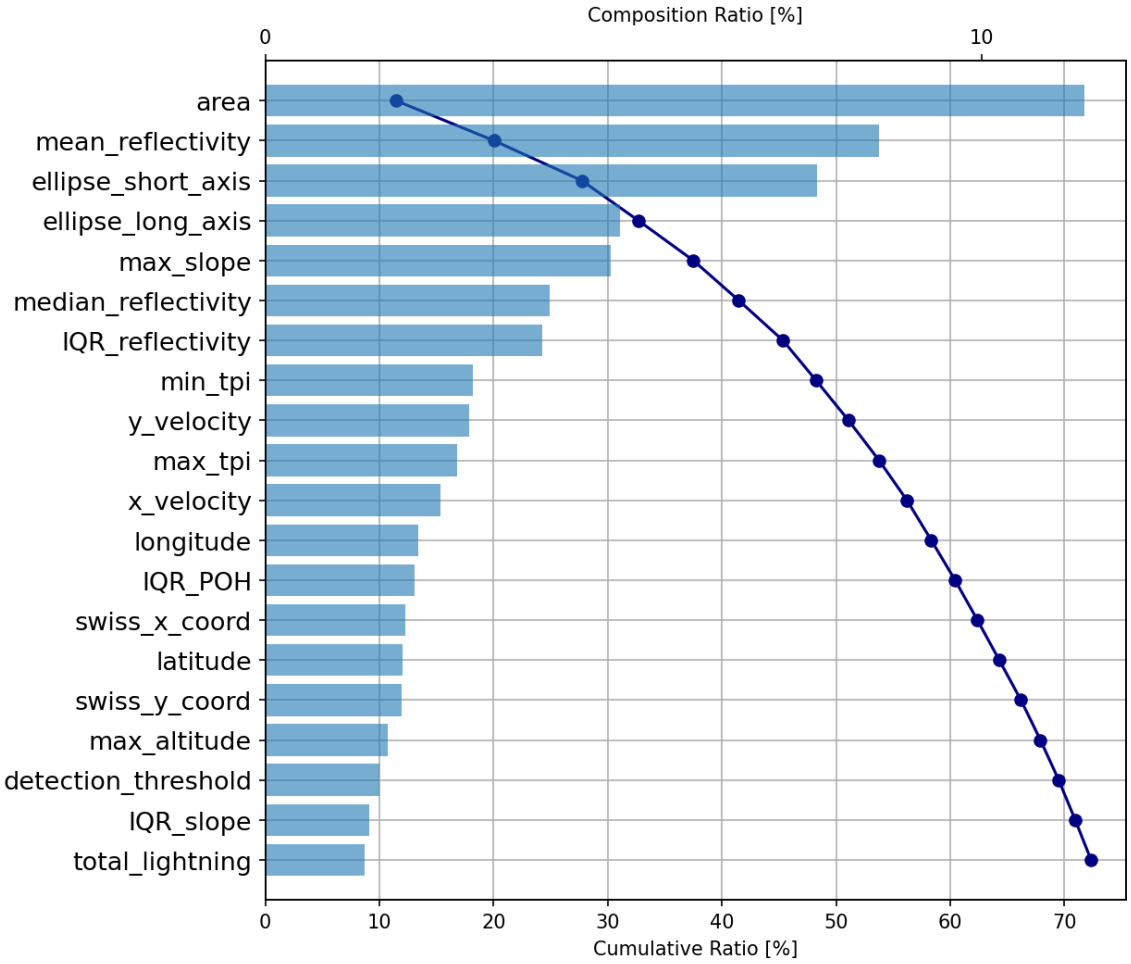

Figure 23 – Feature importance ranking for the RF, computed on test data using the SHAP method. Only the 20 most important features are shown.

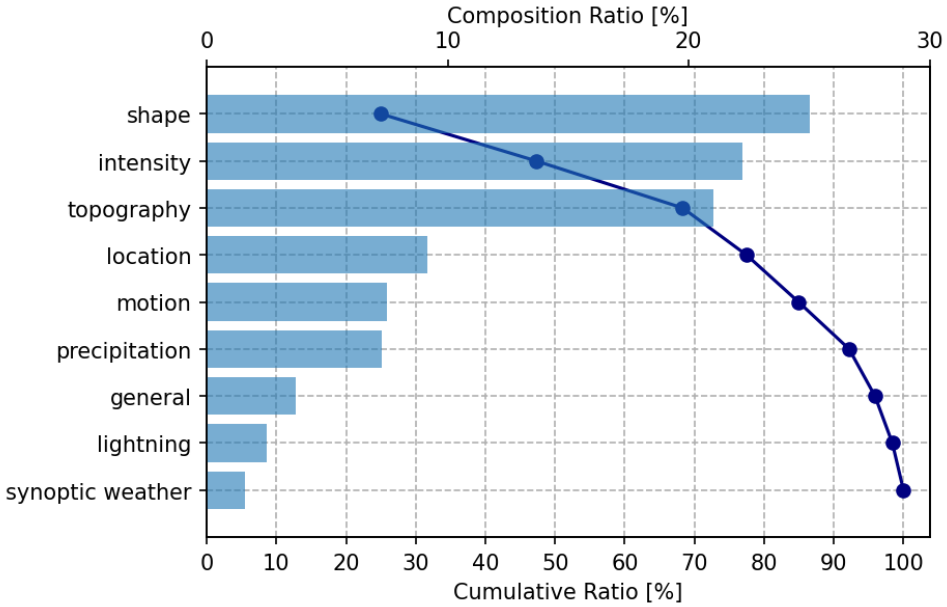

Figure 24 – Feature importance ranking for the RF, computed on test data using the SHAP method, by category of features.

To get a better sense of which type of features plays a bigger role, features are gathered in categories that describe each physical property of the thunderstorm: general, location, shape, motion, topography, synoptic weather, intensity and lightning. The details of the categories are given in Table XVI in Appendix. Figure 24 shows the feature importance using these categories. Shape features prove to be very important as this category has the highest composition ratio: 25%. It is followed by intensity, topography and precipitation. This graph allows us to have a better idea of which features are crucial for the learning of the model. It can also be useful in the future, to efficiently reduce the number of features without decreasing the performance of the model too much. Studying these two types of graphs together might also be relevant in the case of very dependent features. The longitude and the Swiss x coordinate are good examples as they represent a similar physical parameter. Having two features that are so strongly correlated makes the feature importance smaller and shared among them. Comparing the importance of one of these features to the other is not interesting but comparing the total importance of these two features with other features is more interesting. This is why computing the importance of groups of features might be meaningful.

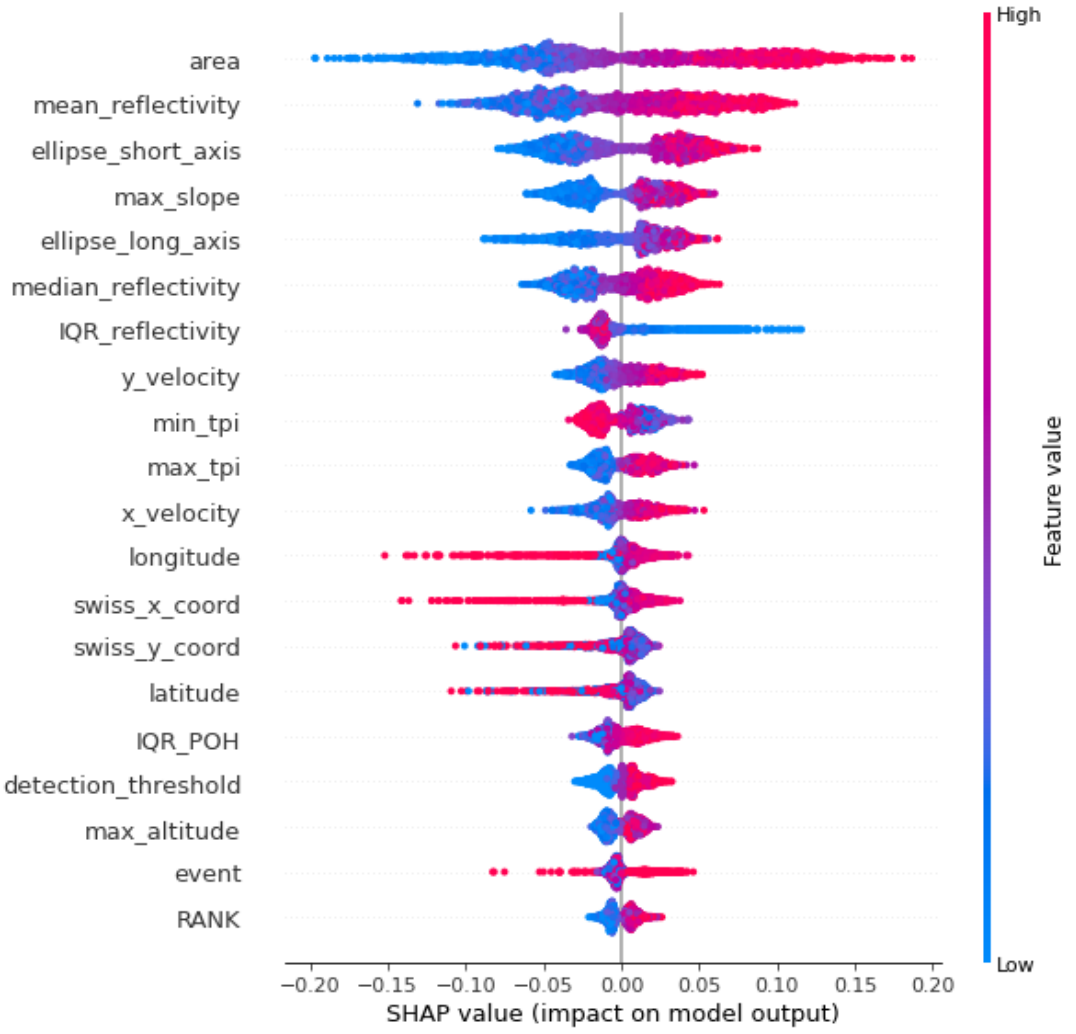

Figure 25 – Summary plot of the SHAP method, detailing feature importance ranking for the RF as well as the feature effect. It is computed on the test set. Only the 20 most important features are shown.

To provide more insight into the reasons why each variable is important, Figure 25 displays the summary plot of the SHAP method, which adds information about the feature effect on the prediction. Each point corresponds to one sample in the dataset. The colour scale indicates the value of the feature. The SHAP value illustrates how much the value of the feature is changing the output of the model. On the horizontal axis, the SHAP value indicates if the feature has a positive (for positive SHAP value) or a negative (for negative SHAP value) impact on the prediction, along with the intensity of the impact. Most of the features have a positive impact on the final prediction. For example, we can see that for higher area values, the SHAP value is positive so the feature area makes the predicted probability of being a mesocyclone increase. Similar reasoning can be done for the various reflectivity statistics and the x and y velocities. A study by Wapler, 2021 on physical characteristics of thunderstorms highlights that supercell thunderstorms are on average larger, faster and exhibit higher reflectivity values than other types of thunderstorms. Therefore, the model seems to have extracted physically relevant information from the features. The distribution of the points for the maximum slope is more mixed, as we can see a certain number of blue points among the red ones, that have positive SHAP values. Also, longitude and the Swiss x coordinates have very similar distributions, which is coherent as they represent the same physical parameter. A similarity can also be seen between the distributions of latitude and the Swiss y coordinate. Finally, maximum and minimum TPI have reversed distributions in terms of colours. High maximum TPI and low minimum TPI values tend to increase the predicted probability of being a mesocyclone. High maximum TPI and low minimum TPI values indicate that the storm cells are situated in areas with ridges and valleys, which also correspond to areas with strong slopes. These data points with high maximum TPI, low minimum TPI and high maximum slope mostly correspond to data points situated in the Prealps, as only a few data points are situated over the Alps, as it can be seen in Figure 7.

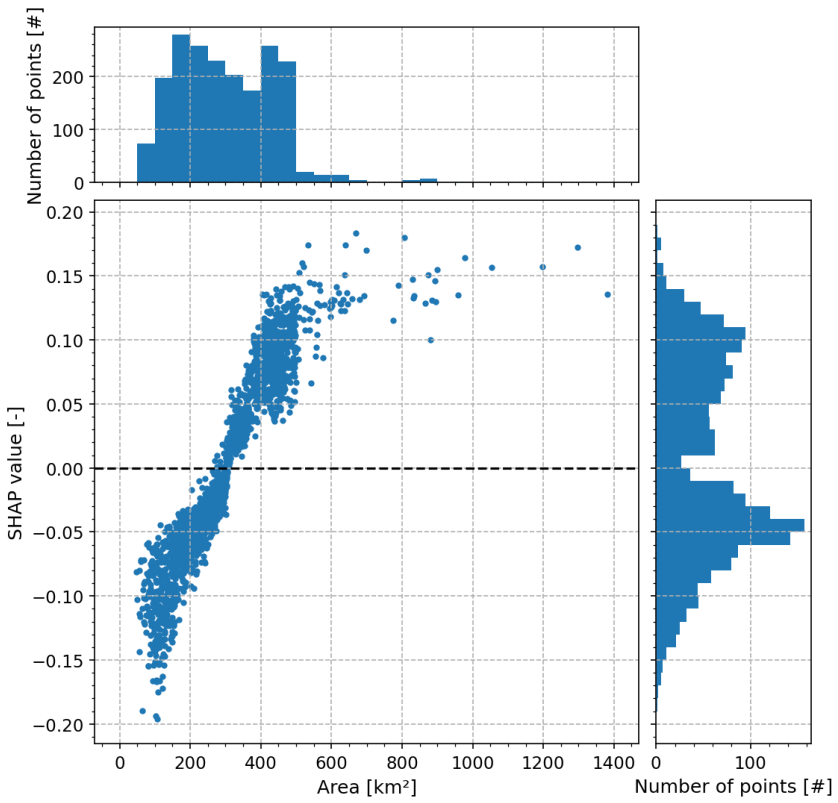

Figure 26 – Dependence plot for the area [km^2] in the RF, computed on the test set. The scatter plot is surrounded by two histograms showing the distribution in number of points [#].

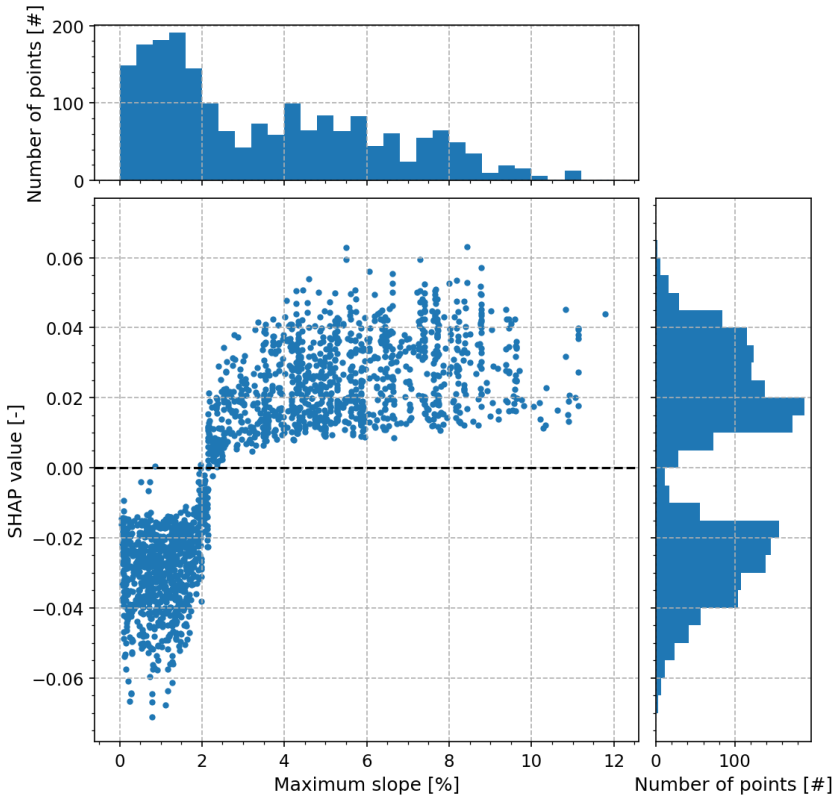

Figure 27 – Dependence plot for the maximum slope [%] in the RF, computed on the test set. The scatter plot is surrounded by two histograms showing the distribution in number of points [#].

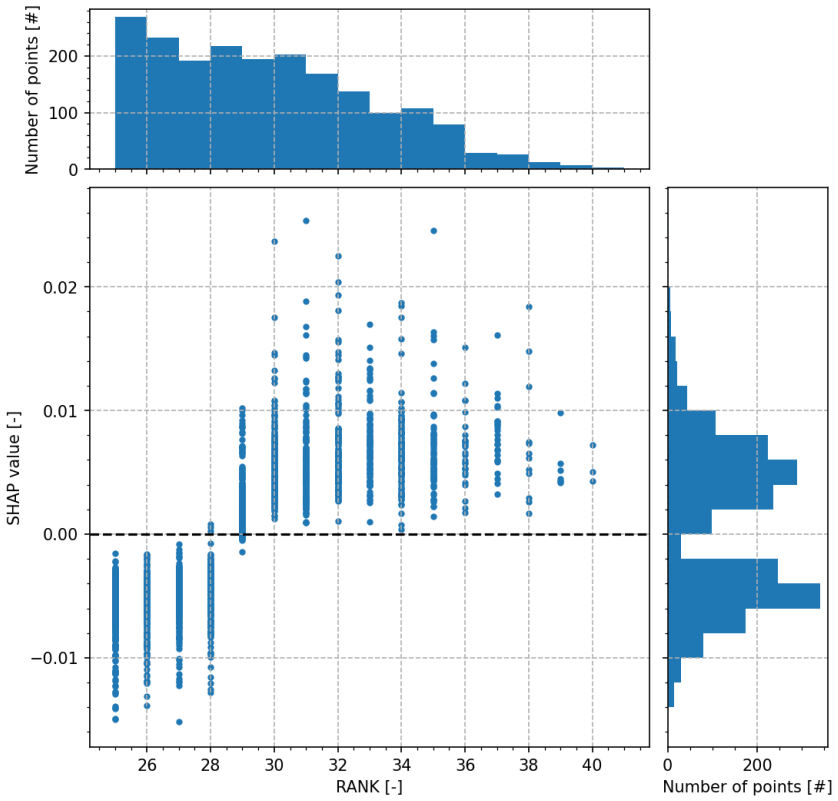

Figure 28 – Dependence plot for the RANK [-] in the RF, computed on the test set. The scatter plot is surrounded by two histograms showing the distribution in number of points [#].

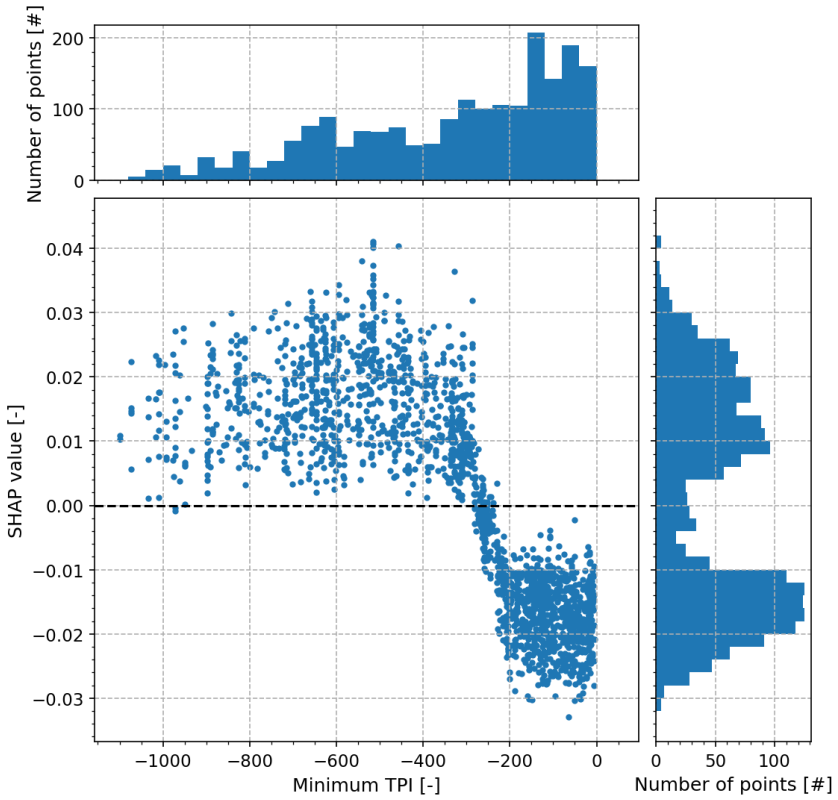

Figure 29 – Dependence plot for the minimum TPI [-] in the RF, computed on the test set. The scatter plot is surrounded by two histograms showing the distribution in number of points [#].

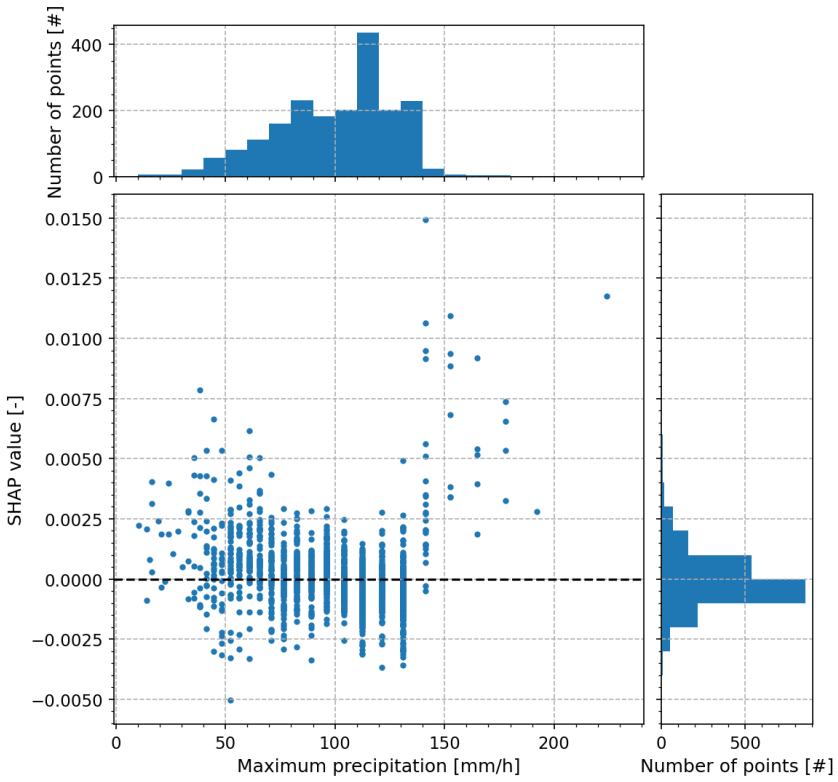

Figure 30 – Dependence plot for the maximum precipitation [mm/h] in the RF, computed on the test set. The scatter plot is surrounded by two histograms showing the distribution in number of points [#].

Dependence plots for the RF are shown in Figures 26 to 30. For each feature, the scatter plot displays the SHAP value as a function of the value of the feature. Each point on the graph represents one sample in the dataset. It is surrounded by two histograms showing the distributions, in number of points []. The first 4 features, i.e. area, maximum slope, RANK and minimum TPI, are among the most important features and the fifth feature, the maximum precipitation, is one of the least important features. These figures confirm what has been observed in Figure 25 and bring more insight into the relationship between the feature and the SHAP value. On one hand, for the area, the maximum slope, the RANK and the minimum TPI, in Figures 26 to 29, the values of the feature are highly correlated to the SHAP values. On the other hand, for the maximum precipitation, we can see in Figure 30 that the point cloud is centred around the horizontal axis and that the correlation between the feature and the SHAP values is low. There are similarities between Figures 26 to 28. The cloud of points seems to have a sigmoid shape. A saturation effect can be observed as the SHAP values don't increase indefinitely. At some point in the range of feature values, when the value of the feature is too high or too low, the impact on the prediction is more or less stable. Furthermore, around the horizontal line $y = 0$, represented on the figures by a black dashed line, the clouds of points are approximately linear and tend to be less dispersed. For example, for the area, there is a very clear threshold: areas > 300 km^2 all influence positively the prediction whereas areas < 300 km^2 almost all influence negatively the prediction. Above 500 km^2, the saturation effect is visible and larger areas have approximately the same SHAP value as an area of 500 km^2: the SHAP values are not higher than 0.17. Figure 29 exhibits similar properties but the correlation with the SHAP values is negative. The histograms emphasise the impact on the SHAP values as for the first 4 features, the SHAP values distributions exhibit 2 peaks on both sides of the horizontal axis. For the maximum precipitation on the opposite, the distribution exhibits only one peak centred on the horizontal axis. It is important to remember that this doesn't explain any causality effects. The correlations between these features and the SHAP values are coherent with results from studies showing that there are statistically significant differences in the distributions of these features for mesocyclones and thunderstorms. Therefore, the model can identify, for the most important features, meaningful correlations between features and labels.

4.2 Classification of storm tracks

4.2.1 Performance of the models

The second part of this project aims to take into account the entire storm track, to better classify the storm as mesocyclonic or non-mesocyclonic. The CNN is compared to an ANN and a meteorological baseline, defined in Section 3.2. The CNN is tuned using grid-search on training data and the best hyper-parameters are : $u = 150$, $k = 5$, $d = 0.15$ and $f = 32$. The same procedure is applied to the ANN and the optimised number of neurons in the dense layer is set to 100. The ROC curves for the CNN and the basic ANN are shown in Figure 31. We can see that the two curves are very similar and the AUC values are both equal to 0.85. Figure 32 shows the performance diagram for the same models, with the addition of the meteorological baseline. The average performance of the ANN is higher than the average performance of the CNN, and the CSI value is 0.42 for the ANN against 0.39 for the CNN. However, parts of the error bars are overlapping, which means that for one particular run, the CNN might perform equally or better than the ANN. The difference in performance between the two models is not statistically significant. The two models are performing better than the meteorological baseline, which has a CSI value of 0.21. This curve doesn't exhibit any error bar because there is no randomness in the model. The selected decision thresholds are $\tau_{\text{meteo}} = 32.2$, $\tau_{\text{ANN}} = 0.283$

and $\tau_{CNN} = 0.384$. For the meteorological baseline, it means that all storms with a maximum RANK over the life-cycle >32.2 are categorised as mesocyclones. Even if these graphs cannot be directly compared to the graphs from Section 4.1.2 because the dataset is different and the task is different, one could have expected the models to perform better for this task as more information is available. Part of the explanation can be due to the dataset. Indeed, in this part, the minority class represents 19.8% of the dataset, which is a smaller fraction than 27.4% in the first part. More importantly, the dataset is considerably smaller as one sample represents one storm. This means that the model has fewer samples to train on so it is less able to generalise the extracted patterns. Furthermore, small datasets tend to lead to more overfitting (Chollet, 2018), so this issue needs to be addressed carefully. Resampling methods have been applied to the ANN and the CNN but similarly to what has been observed in Section 4.1.1, they were not helpful to improve the performance of the models.

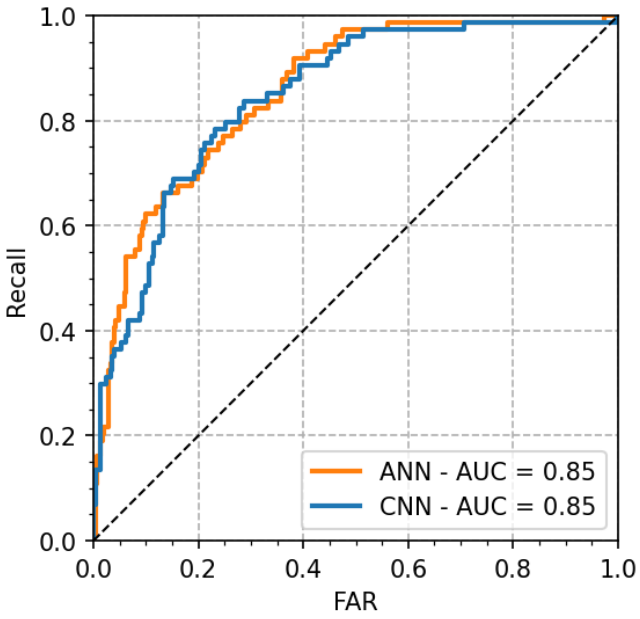

Figure 31 – ROC curve and AUC value for the CNN and the basic ANN on test data. The points from low decision thresholds are in the top-left corner and the points from high decision thresholds are in the bottom-right corner. The black dashed line indicates the performance of a no-skill model outputting random predictions.

Figure 33 displays the distribution of the predicted probability that the timestep is mesocyclonic, output by the CNN on the test set. There is a peak around 0 with 220 predictions < 0.1 and a smaller peak around 1 with 30 predictions. Between these two peaks, the number of predictions is low. Compared to Figure 20, the peak around 1 is more visible and this would mean that the CNN can classify a storm as mesocyclonic with a greater degree of certainty as the RF can classify a timestep as mesocyclonic. Table XIII shows the confusion matrices for the CNN, the basic ANN and the meteorological baseline. To obtain these confusion matrices, the decision thresholds selected in Figure 32 are applied. Compared to the meteorological baseline, both neural networks have higher ratios of TP and TN as well as lower ratios of FP and FN. When comparing the ANN and the CNN, we can observe that the CNN has a ratio of TP that is 1.25 times higher but a ratio of TN that is 1.1 times lower. Furthermore, it has a FP ratio that is 2 times bigger and a FN ratio that is 1.5 times lower. Table XIV lists the values of some performance metrics for the three models. It can be seen that the CNN performs better than the two other models in terms of recall and F2-score. Indeed, in the confusion matrices, the ratio

of FN is smaller for the CNN. The ANN performs better than the two other models in terms of AUC, precision, MCC and CSI. The difference in scores between the CNN and the ANN is not always statistically significant. For the MCC and the CSI scores, the error intervals for the CNN and the ANN are overlapping. This is represented by the symbol * next to the score. These results show that the RANK alone doesn't contain enough information to reliably classify mesocyclones. The neural networks are using more environmental variables and perform better than the meteorological baseline. The comparison between ANN and CNN shows that the utility of convolutional layers in the neural network is limited.

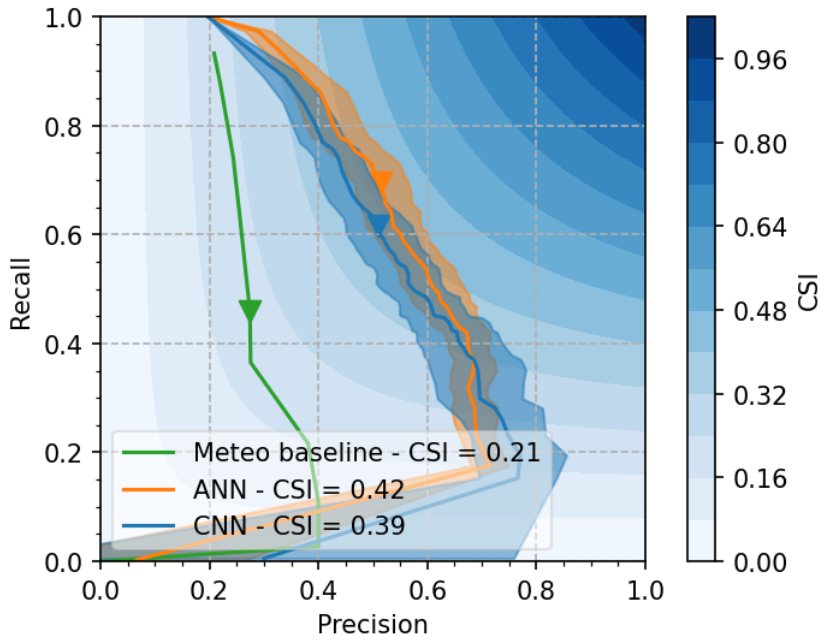

Figure 32 – Performance diagram for the CNN, the basic ANN and the meteorological baseline on test data. Dark lines show the mean and light shadings show the standard deviation. The triangles correspond to the decision thresholds that maximise CSI on the validation data : $\tau_{\text{meteo}} = 32.2$, $\tau_{\text{ANN}} = 0.242$ and $\tau_{\text{CNN}} = 0.283$.

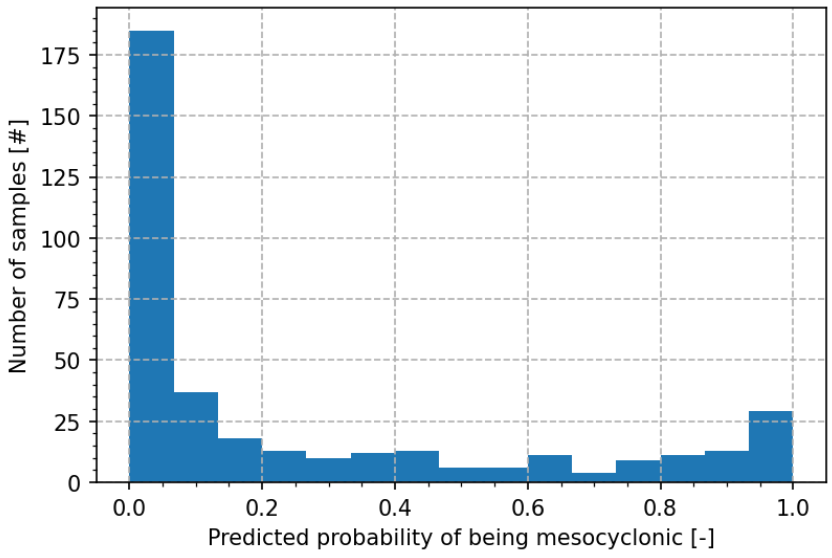

Figure 33 – Distribution of the predicted probability that the storm contains at least one mesocyclonic timestep, output by the CNN on the test set.

Table XIII – Confusion matrices showing the ratios of TP, TN, FP and FN [%], for the meteorological baseline, the ANN and the CNN.

Model	**Predicted**	**Observation**	
		1	**0**
Meteo Baseline	**1**	9	24
	0	11	56
ANN	**1**	12	6.6
	0	7.4	74
CNN	**1**	15	13
	0	5	67

Table XIV – Evaluation metrics for the CNN, the basic ANN and the meteorological baseline. The scores are averaged over 10 runs. The symbol * next to some of the ANN scores means that the error intervals between the ANN score and the CNN score are overlapping and therefore the difference with the CNN model is not statistically significant.

Model	**AUC**	**Precision**	**Recall**	**F2**	**MCC**	**CSI**
Meteorological baseline	0.581	0.274	0.459	0.404	0.137	0.207
Basic ANN	**0.856***	**0.688**	0.568	0.588	**0.544***	**0.451***
CNN	0.851	0.500	**0.716**	**0.659**	0.478	0.417

4.2.2 Feature importance and interpretability

As in Section 4.1.4, an analysis of the feature importance is carried out for the CNN. The feature importance graph is illustrated in Figure 34 for individual features and Figure 35 by categories of features. We can see that the detection threshold is the most important feature with a ratio of 5%. Among the ranking, we can also find some statistics related to precipitations: MESHS, POH and precipitation. In addition, topographical features like aspect at the centroid of the storm cell, minimum and maximum TPI, as well as maximum slope are in the list of the 20 most important features. Finally, we can find features containing information about the location such as categorical features for the Southern Prealps, the Northern Prealps, the Po Valley, Baden-Württemberg and none of the sub-domains. The category with precipitation features is the second most important one in Figure 35, after the category with topographical features which has a ratio of approximately 23%. Location is the third most important category of features.

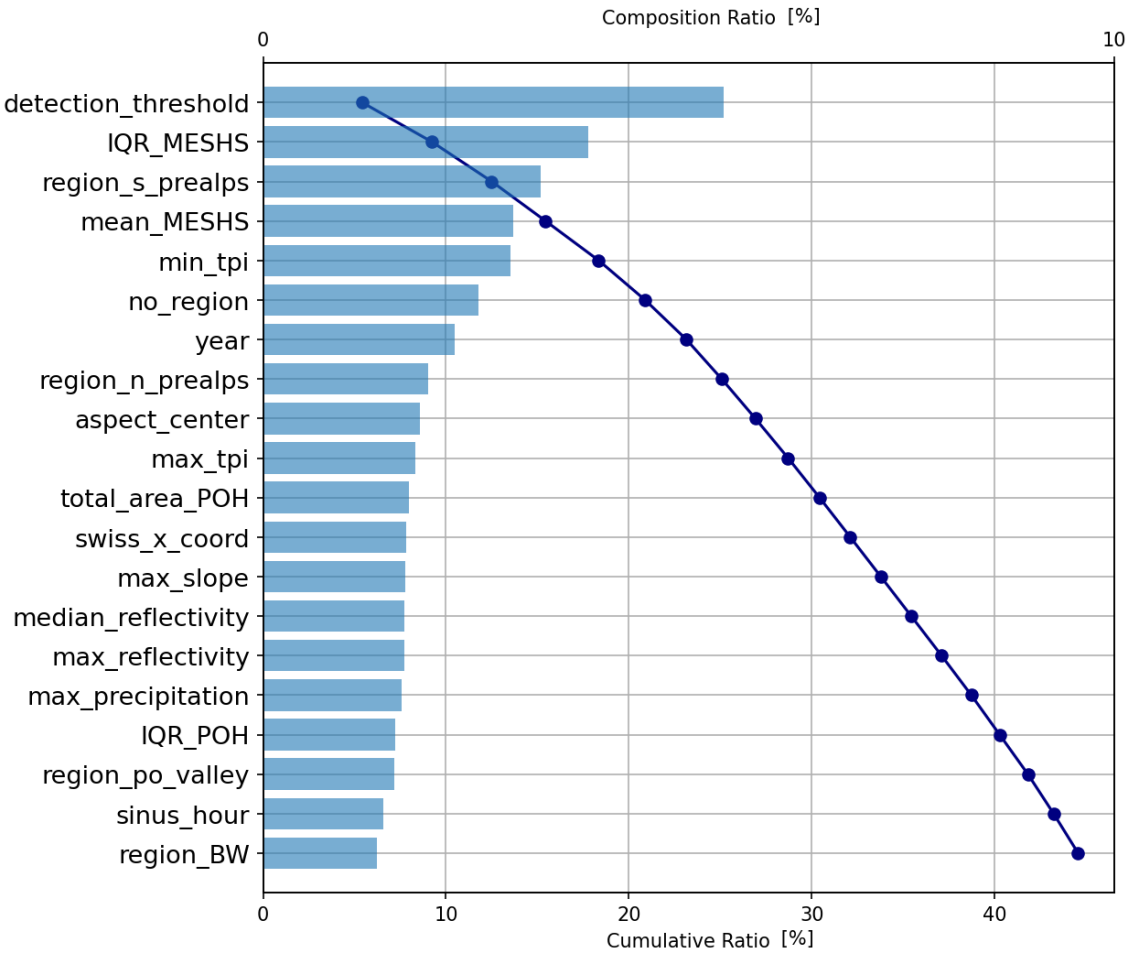

Figure 34 – Feature importance ranking in the CNN, computed on test data using SHAP method. Only the 20 most important features are shown.

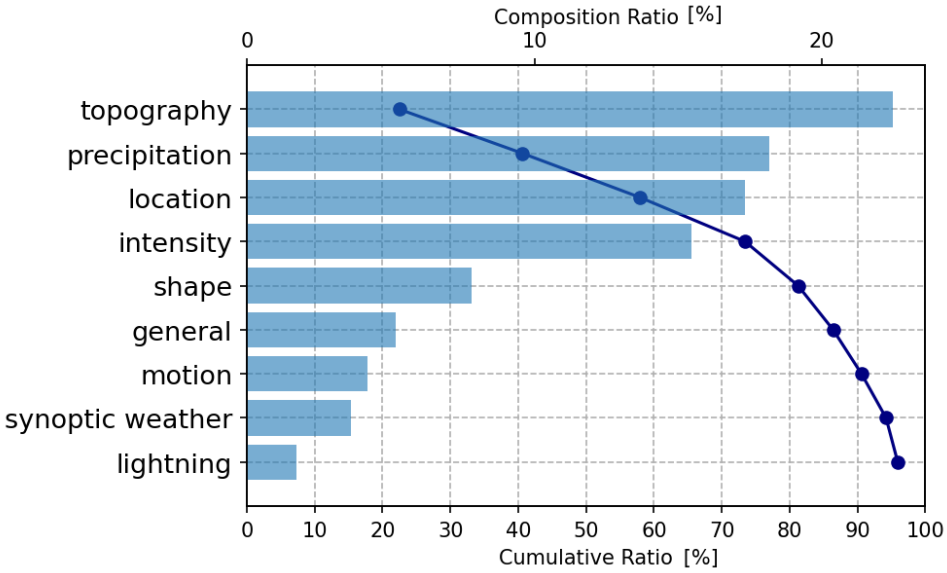

Figure 35 – Feature importance ranking in the CNN, computed on test data using SHAP method, by category of features.

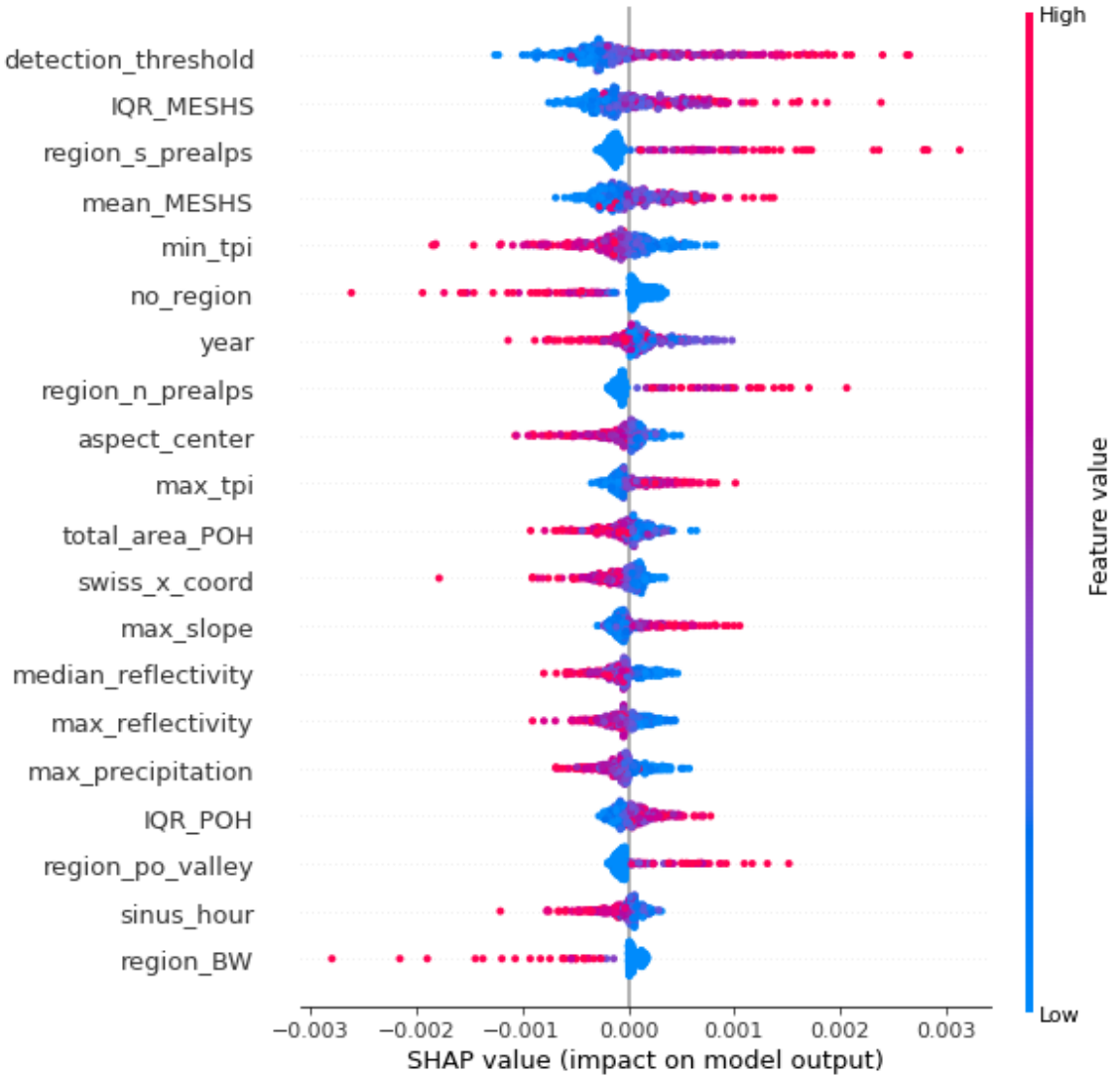

Figure 36 – Summary plot of the SHAP method, detailing feature importance ranking in the RF as well as the feature effect. It is computed on the test set.

Figure 36 shows the summary plot for the CNN, computed on test data. To plot one SHAP value per feature, the SHAP values are averaged over the entire storm track for each feature. As the storm tracks are on average 29 timesteps long, compared to the maximum length which is 129, most of them contain a high number of zero timesteps at the end, which make the output SHAP value lower. This explains the difference of order of magnitude with SHAP values computed in Section 4.1.4. Indeed, there is a difference of two orders of magnitude between SHAP values shown in Figures 25 and 36. This averaging introduces a bias towards long tracks: as they have fewer zero timesteps at the end, their average SHAP value would tend to be higher. Location features are categorical features, i.e. 1 or 0 if the data point belongs to the sub-domain or not. For Southern and Northern Prealps and the Po Valley, data points with a high feature value, i.e. 1, tend to have a positive SHAP value which increases the probability

of being mesocyclonic. Therefore, being a data point situated in one of these three sub-domains makes on average the predicted probability increase. For the Southern Prealps, the impact on the predicted probability is higher than for the Northern Prealps, which in turn is higher than for the Po Valley. On the opposite, for the Baden-Württemberg sub-domain and the rest of the domain, higher feature values are associated with negative SHAP values that make the predicted probability decrease. Therefore, being a data point situated in Baden-Württemberg or none of the sub-domains makes the predicted probability decrease. The distributions of the location features are not symmetric with respect to the y-axis: low values of the features, i.e. 0, have a low SHAP value in absolute value. It means that for a data point that is not situated in the Southern Prealps for example, the information that the point isn't located in the Southern Prealps has a very small impact on the prediction. Indeed, more crucial information is the sub-domain in which the data point is located. A study by Feldmann et al., 2021 highlighted that the Southern and Northern Prealps are the most active regions in Switzerland in terms of thunderstorms and mesocyclones. Moreover, Figure 7 illustrates the ratio between the number of detected mesocyclones by the number of detected thunderstorms per 1 km^2. We can observe that the highest ratios are located in the Southern Prealps and the Po Valley, and more sparsely in the Northern Prealps. Consequently, the model synthesises the positive correlation between the Southern Prealps, the Northern Prealps and the Po Valley, and the probability that the storm contains a mesocyclone. Results for Baden-Württemberg and outside of the sub-domains have to be taken with care. The relative quality index is lower in the Baden-Württemberg sub-domain. If the data point doesn't belong to any of the sub-domains, it is at the boundaries of the domain, at a great distance from the radars so the relative quality index is in general also low. A lower quality index increases the risk of underestimating thunderstorm occurrence and minimising the severity of the thunderstorm Feldmann et al., 2021. For features related to precipitation, the relation between feature value and SHAP value is more nuanced, as high and low feature values can be seen for the same SHAP value. Finally, concerning the topography, minimum TPI and maximum TPI have opposite effects on the predicted probability, as observed in Figure 25 for the RF.

Figures 37 to 39 display dependence plots for the CNN, for three features: detection threshold and minimum TPI, which are among the most important features and change in x velocity which is one of the least important features. As in Section 4.1.4, the positive and negative correlations between feature values and SHAP values can be seen for the most important features whereas for the change in x velocity, no clear slope can be observed. In Figures 37 and 38, the distinction between positive and negative impact is less strict than for the RF. Indeed, the distributions of the SHAP values only exhibit one peak, on one side of the horizontal axis. Furthermore, for high values of the features, the range of SHAP values is wide. For example, for two data points with a normalised minimum TPI at 0.54, the impact on the prediction can vary from $-1.0 \cdot 10^{-4}$ to $-1.9 \cdot 10^{-3}$, one order of magnitude. Part of the explanation might be the averaging as mentioned before: for two tracks with different lengths, the SHAP values resulting from the average can be very different.

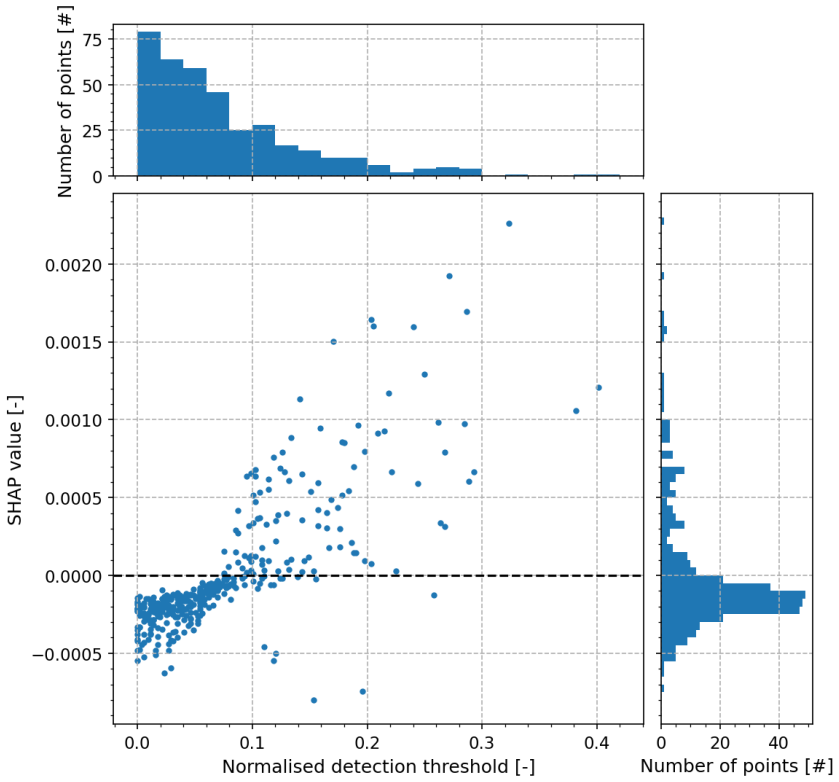

Figure 37 – Dependence plot for the normalised detection threshold [-] in the CNN, computed on the test set. The scatter plot is surrounded by two histograms showing the distribution in number of points [#].

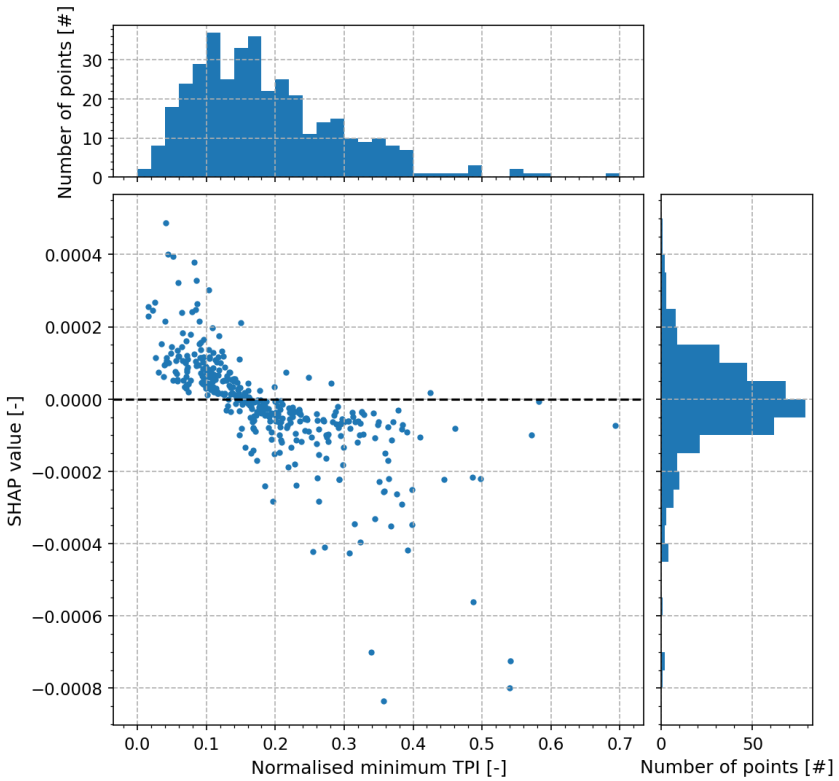

Figure 38 – Dependence plot for the minimum TPI [-] in the CNN, computed on the test set. The scatter plot is surrounded by two histograms showing the distribution in number of points [#].

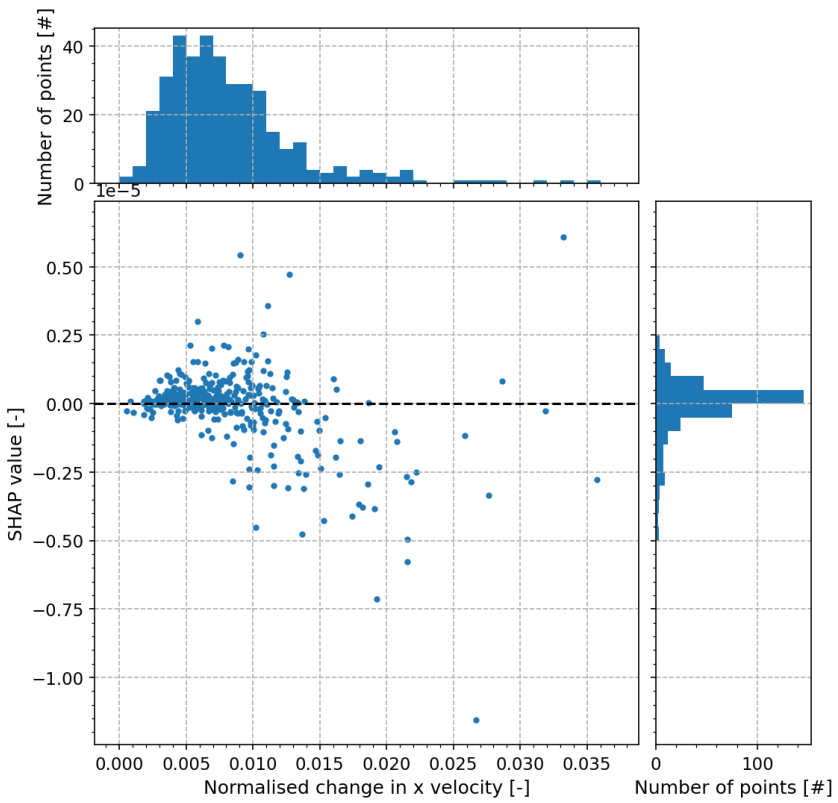

Figure 39 – Dependence plot for the change in x velocity [-] in the CNN, computed on the test set. The scatter plot is surrounded by two histograms showing the distribution in number of points [#].

5 Conclusion

In this report, we present the results of a mesocyclone classification task over the Alpine area using RFs and CNNs. The data consists of intensity variables extracted from 6-year radar images from the Swiss radar network. The TRT algorithm detects thunderstorm tracks in the radar images and the MDA algorithm identifies supercells among these detected storm cells using Doppler velocity data. Additionally, features containing information about synoptic weather, location and topography are used. A RF is used to predict the probability that a particular timestep is mesocyclonic while a CNN is used to predict the probability that the entire storm contains at least one mesocyclonic timestep. The RF, respectively the CNN, is compared to a LR, respectively an ANN, which plays the role of the machine learning baseline. There is no existing meteorological baseline to compare the performance of the RF with. A meteorological baseline for the CNN is defined here as a condition on the RANK, a heuristic variable that aggregates various features and evaluates the severity of the storm: storms with life-cycle maximum RANK higher than a threshold are considered as mesocyclones. The models are trained and tuned on training data using grid-search and cross-validation and then evaluated on test data.

In the first part, the RF performs significantly better than the LR. Using resampling methods to mitigate the unbalance and adding a convexity feature don't improve the performance. On the opposite, focusing on severe thunderstorms with RANK > 25 improves the performance of the models, with AUC = 0.92 and CSI = 0.61 for the RF. More particularly, the RF has a ratio of FN that is 2 times lower than the LR, which is encouraging because missing an event is the most critical error for hazardous events like supercell thunderstorms. Reasons for this performance improvement can be the unbalance mitigation created by the filtering on the RANK, as the percentage of mesocyclones in the dataset increases from 0.6% to 27%. Also, discarding moderate thunderstorms that are more easily misclassified by the model might help. The study of potential biases shows that the distribution of the predictions has no obvious biases, neither in altitude nor in location. Furthermore, a reliability diagram reveals that the RF is over-predicting rare events and under-predicting more frequent events, which is what would be expected as supercells are dangerous and rare events. Finally, the analysis of feature importance and feature effect on the predictions shows that shape, intensity and topography features are the most important ones. Furthermore, the model seems to be able to extract relationships such as the fact that a bigger area and a higher lightning rate increases the probability of a storm cell to be mesocyclonic, that are in accordance with what has been observed in another study on mesocyclones life-cycle (Wapler, 2021).

In the second part, the CNN performs better than the meteorological baseline but approximately as well as the classic ANN. The CNN outperforms the ANN in terms of recall and F2-score but is outperformed by the ANN in terms of AUC, precision, MCC and CSI. Nonetheless, some of these discrepancies between the two models are not statistically significant. The performance of the CNN might be improved by choosing a different structure for the layers and a more in-depth tuning of the hyper-parameters, but the small difference in performance between the CNN and the ANN suggests that the convolutional step might not be useful. The performance reached by the CNN, AUC = 0.85 and CSI = 0.39, are much smaller than the performance reached by the RF, even though they cannot be compared directly. The main reason for that must be the smaller size of the dataset that reduces the capability of the model to generalise. Interpretation of the model using feature importance shows that the categories containing topography, precipitation and location features play the biggest role. A strong positive correlation is observed between being located in the Southern and Northern Prealps, the Po Valley sub-domains and the predicted probability of being mesocyclonic. A previous study on

the occurrence of mesocyclones shows that they are the most active regions for mesocyclones (Feldmann et al., 2021). Therefore, as for the RF, the CNN seems to be able to extract information that is coherent with our knowledge of mesocyclone occurrences. The RF and CNN prove that it is possible to classify mesocyclonic and non-mesocyclonic timesteps or storms without using any Doppler velocity data. They could be helpful to classify mesocyclones further back in time and therefore extend the database of mesocyclones for climatology studies. In this work, we focus on supercell thunderstorms in the Alpine area. The location is very particular as the topography is complex. Furthermore, the Swiss radar network has its inherent specificity so extrapolation to another geographic area should not be undertaken without re-training the model with data from the radar network of this specific area.

The machine learning models used in this work have room for improvement. First, a deeper analysis could investigate some case studies for FN and FP predictions in comparison to TP and TN. This would allow us to better understand what are the patterns leading to false predictions. The SHAP method would be a useful tool to interpret these individual predictions. To synthesise more general patterns, another possibility would be to identify extreme cases among the predictions, as the best TP, the mesocyclonic events with the highest probabilities and the worst FP, the mesocyclonic events with the lowest probabilities and similarly for the non-mesocyclonic events, as done by Lagerquist et al., 2020. Instead of focusing on single cases, this method would give a more global view of the potential biases of the models. It would also be interesting to add more features to the dataset. From a meteorological point of view, ideal information for this classification task would be near-storm environment data such as wind, temperature profile or humidity. These features would allow to better capture patterns during the life-cycle of the thunderstorm. This type of information is not easily available from the radar network. Another possibility can be the use of radar images directly instead of statistics computed from these images. The pre-processing steps would be simplified, and more information would be contained in the data. One pre-processing step that would still be necessary is the transformation from polar to Cartesian grid. The CNN would be a perfectly appropriate model to treat this type of data. Also, sounding data from Payerne and Milan can be added to give more information about the synoptic weather. For the RF, however, the synoptic weather classifications are not among the most important features and the performance improvement might be marginal. When adding more data to the models, we must be careful with the risk of overfitting. Strategies to avoid this issue can be the selection of some features based on feature importance analysis or the computation of feature statistics. Another direction for future studies would be to take into account the time-dependency between the timesteps, to predict the intensity (regression task) or the tendency (classification task) of the mesocyclone's rotation for future timesteps. One type of model that is particularly suitable for this task is Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber, 1997), one type of Recurrent Neural Networks (RNN). We would consider entire tracks and use rotational properties from the MDA algorithm. Contrary to ANNs or CNNs, RNNs have feedback connections that allow them to keep information from one step to another. They are particularly useful to process sequences of data like time series. We have briefly explored this approach with the implementation of an LSTM to predict the rotation tendency for the next timestep among four cases: stable, intensification, decay or no rotation. The performance of the LSTM was compared to a baseline: a persistence model which predicts the forecast timestep as equal to the current timestep. This approach was not pursued due to time constraints. A machine learning model that would be able to efficiently predict mesocyclones' rotation evolution would be very useful for real-time forecasting. Furthermore, using interpretation methods on the model could bring insight into the dynamic relationships.

A Appendix

Table XV – Table listing the features available in the data set, as well as the definition and the units. The column 'Short name' refers to the name given in the dataset used in the algorithm and the 'Name' refers to a more explicit name. Explicit names are not given to the rotational variables as they are not used in this work.

Name	Short name	Definition	Units
ID_storm	ID	Thunderstorm identification number	-
time	time	Time – YYYYMMDDHHmm	-
longitude	lon	Thunderstorm centroid longitude	°E
latitude	lat	Thunderstorm centroid latitude	°N
swiss_x_coord	chx	Swiss x coordinate	m
swiss_y_coord	chy	Swiss y coordinate	m
ellipse_long_axis	ell_L	Major axis of the ellipse	km
ellipse_short_axis	ell_S	Minor axis of the ellipse	km
ellipse_orientation	ell_or	Ellipse orientation	°
area	area	Thunderstorm area	km^2
x_velocity	vel_x	Zonal velocity	km/h
y_velocity	vel_y	Meridional velocity	km/h
x_change_velocity	Dvel_x	Change of zonal velocity	km/h/15min
y_change_velocity	Dvel_y	Change of meridional velocity	km/h/15min
RANK	RANKr	Thunderstorm RANK	-
	RANK	Thunderstorm RANK rounded between 0 and 4	-
neg_lightning	CG-	Cloud to ground flashs in the last 5 min with negative polarisation	#
pos_lightning	CG+	Cloud to ground flashs in the last 5 min with positive polarisation	#
total_lightning	CG	Total cloud to ground flashs in the last 5 min	#
ratio_pos_lightning	%CG+	Percentage of positive cloud to ground flashs in the last 5 min	‰
max_height_echotop45	ET45	Maximum height of ECHO TOP 45dBZ	km
median_height_echotop45	ET45m	Median height of ECHO TOP 45dBZ	km
max_height_echotop15	ET15	Maximum height of ECHO TOP 15dBZ	km
median_height_echotop15	ET15m	Median height of ECHO TOP 15dBZ	km
max_height	maxH	Maximum height of max. radar signal	km
median_height	maxHm	Median height of max. radar signal	km
vertical_liquid_water	VIL	Vertical integrated liquid water	kg/m^2
max_precipitation	max_CPC	Maximum combiprecip (precipitation)	mm/h
mean_precipitation	mean_CPC	Mean combiprecip (precipitation)	mm/h
median_precipitation	median_CPC	Median combiprecip (precipitation)	mm/h
IQR_precipitation	IQR_CPC	IQR combiprecip (precipitation)	mm/h
max_POH	max_POH	Maximum Probability Of Hail	‰
mean_POH	mean_POH	Mean Probability Of Hail	‰
median_POH	median_POH	Median Probability Of Hail	‰
IQR_POH	IQR_POH	IQR Probability Of Hail	‰
max_MESHS	max_MESHS	Maximum Expected Severe Hail Size	cm
mean_MESHS	mean_MESHS	Mean Expected Severe Hail Size	cm
median_MESHS	median_MESHS	Median Expected Severe Hail Size	cm
IQR_MESHS	IQR_MESHS	IQR Expected Severe Hail Size	cm
max_reflectivity	max_CZC	Maximum Composite Reflectivity	dBZ
mean_reflectivity	mean_CZC	Mean Composite Reflectivity	dBZ
median_reflectivity	median_CZC	Median Composite Reflectivity	dBZ
IQR_reflectivity	IQR_CZC	IQR Composite Reflectivity	dBZ
altitude_center	altitude	Terrain altitude at centroid	m
slope_center	slope	Terrain slope at centroid	‰
aspect_center	aspect	Terrain aspect at centroid	°
max_altitude	max_alt	Maximum Altitude	m
mean_altitude	mean_alt	Mean Altitude	m
median_alt	median_alt	Median Altitude	m
IQR_alt	IQR_alt	IQR Altitude	m
min_alt	min_alt	Minimum Altitude	m
max_slope	max_slp	Maximum Slope	‰

Table XV – Table listing the features available in the data set (... continued)

Name	Short name	Definition	Units
mean_slope	mean_slp	Mean Slope	‰
median_slope	median_slp	Median Slope	‰
IQR_slope	IQR_slp	IQR Slope	‰
min_slope	min_slp	Minimum Slope	‰
max_aspect	max_asp	Maximum Aspect	°
mean_aspect	mean_asp	Mean Aspect	°
median_aspect	median_asp	Median Aspect	°
IQR_aspect	IQR_asp	IQR Aspect	°
max_aspect	min_asp	Minimum Aspect	°
max_tpi	max_tpi	Maximum Topographic Position Index	-
mean_tpi	mean_tpi	Mean Topographic Position Index	-
median_tpi	median_tpi	Median Topographic Position Index	-
IQR_tpi	IQR_tpi	IQR Topographic Position Index	-
min_tpi	min_tpi	Minimum Topographic Position Index	-
max_qual	max_qual	Maximum Relative Quality Index	-
mean_qual	mean_qual	Mean Relative Quality Index	-
median_qual	median_qual	Median Relative Quality Index	-
IQR_qual	IQR_qual	IQR Relative Quality Index	-
min_qual	min_qual	Minimum Relative Quality Index	-
date	event	Date in YYjjj (day-of-year format)	-
year	year	Year	-
day_of_year	doy	Day of year	-
time	tod	Time of day	-
hour	hour	Time of day rounded to hour	h
GWT8	GWT8	Weather classification with 8 classes	-
GWT10	GWT10	Weather classification with 10 classes	-
GWT18	GWT18	Weather classification with 18 classes	-
GWT26	GWT26	Weather classification with 26 classes	-
direction	dir	Direction of storm	°
smoothed_direction	dir_s	Smoothed direction of storm	°
distance	dis	Distance to last detection	m
difference_altitude	dalt	Altitude difference to last detection	m
difference_slope	dslope	Slope difference to last detection	‰
pos_rot	pos	Label for positive rotation	0/1
neg_rot	neg	Label for negative rotation	0/1
total_area_POH80	area_POH	Total area with POH > 80	km^2
contiguous_area_POH80	c_area_POH	Contiguous area with POH > 80	km^2
total_area_MESHS2	area_MESHS2	Total area with MESHS > 2	km^2
contiguous_area_MESHS2	c_area_MESHS2	Contiguous area with MESHS > 2	km^2
total_area_MESHS4	area_MESHS4	Total area with MESHS > 4	km^2
contiguous_area_MESHS4	c_area_MESHS4	Contiguous area with MESHS > 4	km^2
total_area_reflectivity41	area_ref	Total area with reflectivity > 41	km^2
contiguous_area_reflectivity41	c_area_ref	Contiguous area with reflectivity > 41	km^2
label_severe_hail	s_hail	Timestep with severe hail (c_area_MESHS4 > 5)	0/1
region	region	Region of timestep	-
region_weighted	region_weighted	Region weighted by storm peak, entire track same region	-
label_meso	meso	Timestep with mesocyclone	0/1
label_hail	hail	Timestep with hail (c_area_MESHS2 or c_area_POH > 5)	0/1
label_mesostorm	mesostorm	Storm with mesocyclone (entire track labelled)	0/1
label_hailstorm	hailstorm	Storm with hail	0/1
label_severe_hail	s_hailstorm	Storm with severe hail	0/1
label_mesohail	mesohail	Timestep with mesocyclone and hail	0/1
label_mesohailstorm	mesohailstorm	Storm with mesocyclone and hail	0/1
label_severe_hailstorm	meso_s_hail	Timestep with mesocyclone and severe hail	0/1
label_severe_mesohailstorm	meso_s_hailstorm	Storm with mesocyclone and severe hail	0/1
label_normal	normal	Normal thunderstorm (c_area_MESHS2 < 4)	0/1
	p/n_radar	Radars detecting positive/negative rotation : [A, D, L, P, W]	-
	p/n_x	x coordinate at centroid of rotation	m

Table XV – Table listing the features available in the data set (... continued)

Name	Short name	Definition	Units
	p/n_y	y coordinate at centroid of rotation	m
	p/n_dz	Vertical extent	m
	p/n_A	Label for detection by Albis radar	0/1
	p/n_D	Label for detection by La Dole radar	0/1
	p/n_L	Label for detection by Monte Lema radar	0/1
	p/n_P	Label for detection by Plaine Morte radar	0/1
	p/n_W	Label for detection by Weissfluhgipfel radar	0/1
	p/n_A_range	Range of detection by Albis radar	km
	p/n_D_range	Range of detection by La Dole radar	km
	p/n_L_range	Range of detection by Monte Lema radar	km
	p/n_P_range	Range of detection by Plaine Morte radar	km
	p/n_W_range	Range of detection by Weissfluhgipfel radar	km
	p/n_A_n	Number of detections by Albis radar	#
	p/n_D_n	Number of detections by La Dole radar	#
	p/n_L_n	Number of detections by Monte Lema radar	#
	p/n_P_n	Number of detections by Plaine Morte radar	#
	p/n_W_n	Number of detections by Weissfluhgipfel radar	#
	p/n_A_el	Number of elevations with detection by Albis radar	#
	p/n_D_el	Number of elevations with detection by La Dole radar	#
	p/n_L_el	Number of elevations with detection by Monte Lema radar	#
	p/n_P_el	Number of elevations with detection by Plaine Morte radar	#
	p/n_W_el	Number of elevations with detection by Weissfluhgipfel radar	#
	p/n_size_sum	Sum of all detected areas	km^2
	p/n_size_mean	Mean of all detected areas	km^2
	p/n_vol_sum	Sum of detected volume	km^3
	p/n_vol_mean	Mean of detected volume	km^3
	p/n_z_0	Minimum of altitude of all detections	m
	p/n_z_10	10th percentile of altitude of all detections	m
	p/n_z_25	25th percentile of altitude of all detections	m
	p/n_z_50	50th percentile of altitude of all detections	m
	p/n_z_75	75th percentile of altitude of all detections	m
	p/n_z_90	90th percentile of altitude of all detections	m
	p/n_z_100	Maximum of altitude of all detections	m
	p/n_z_IQR	IQR of altitude of all detections	m
	p/n_z_mean	Spatial mean of altitude of all detections	m
	p/n_r_0	Minimum rotational velocity of all detections	m/s
	p/n_r_10	10th percentile of rotational velocity of all detections	m/s
	p/n_r_25	25th percentile of rotational velocity of all detections	m/s
	p/n_r_50	50th percentile of rotational velocity of all detections	m/s
	p/n_r_75	75th percentile of rotational velocity of all detections	m/s
	p/n_r_90	90th percentile of rotational velocity of all detections	m/s
	p/n_r_100	Maximum rotational velocity of all detections	m/s
	p/n_r_IQR	IQR of rotational velocity of all detections	m/s
	p/n_r_mean	Spatial mean of rotational velocity of all detections	m/s
	p/n_v_0	Minimum vorticity of all detections	s^{-1}
	p/n_v_10	10th percentile of vorticity of all detections	s^{-1}
	p/n_v_25	25th percentile of vorticity of all detections	s^{-1}
	p/n_v_50	50th percentile of vorticity of all detections	s^{-1}
	p/n_v_75	75th percentile of vorticity of all detections	s^{-1}
	p/n_v_90	90th percentile of vorticity of all detections	s^{-1}
	p/n_v_100	Maximum of vorticity of all detections	s^{-1}
	p/n_v_IQR	IQR of vorticity of all detections	s^{-1}
	p/n_v_mean	Spatial mean of vorticity of all detections	s^{-1}
	p/n_d_0	Minimum diameter of all detections	m
	p/n_d_10	10th percentile of diameter of all detections	m
	p/n_d_25	25th percentile of diameter of all detections	m
	p/n_d_50	50th percentile of diameter of all detections	m
	p/n_d_75	75th percentile of diameter of all detections	m

Table XV – Table listing the features available in the data set (... continued)

Name	Short name	Definition	Units
	p/n_d_90	90th percentile of diameter of all detections	m
	p/n_d_100	Maximum diameter of all detections	m
	p/n_d_IQR	IQR of diameter of all detections	m
	p/n_d_mean	Spatial mean of diameter of all detections	m
	p/n_rank_0	Minimum rank of rotation of all detections	-
	p/n_rank_10	10th percentile of rank of rotation of all detections	-
	p/n_rank_25	25th percentile of rank of rotation of all detections	-
	p/n_rank_50	50th percentile of rank of rotation of all detections	-
	p/n_rank_75	75th percentile of rank of rotation of all detections	-
	p/n_rank_90	90th percentile of rank of rotation of all detections	-
	p/n_rank_100	Maximum rank of rotation of all detections	-
	p/n_rank_IQR	IQR of all rank of rotation of all detections	-
	p/n_rank_mean	Spatial mean of all rank of rotation of all detections	-

Table XVI – List of the categories for feature importance analysis and the features inside each category.

General	Location	Shape	Motion	Topography	Synoptic weather	Intensity	Precipitation	Lightning
time	lon	area	vel_x	altitude	GWT8	det	VIL	CG
ID	lat	ell_L	vel_y	slope	GWT10	maxH	hail	CG+
event	chx	ell_S	Dvel_x	aspect	GWT18	maxHm	s_hail	CG-
year	chy	ell_or	Dvel_y	dalt	GWT26	RANKr	max_POH	%CG+
doy	region_weighted	area_POH	dir	dslope		ET45	mean_POH	
tod		c_area_POH	dir_s	max_alt		ET45m	median_POH	
hour		area_MESHS2	dis	mean_alt		ET15	IQR_POH	
		c_area_MESHS2		median_alt		ET15m	max_MESHS	
		area_MESHS4		IQR_alt		max_CZC	mean_MESHS	
		c_area_MESHS4		min_alt		mean_CZC	median_MESHS	
		area_ref		max_slp		median_CZC	IQR_MESHS	
		c_area_ref		mean_slp		IQR_CZC	max_CPC	
				median_slp			mean_CPC	
				IQR_slp			median_CPC	
				min_slp			IQR_CPC	
				max_asp				
				mean_asp				
				median_asp				
				IQR_asp				
				min_asp				
				max_tpi				
				mean_tpi				
				median_tpi				
				IQR_tpi				
				min_tpi				

Bibliography

R. Branzei, D. Dimitrov, and S. Tijs. *Models in cooperative game theory*, volume 556. Springer Science & Business Media, 2008.

L. Breiman. Random forests. *Machine Learning*, 2001:5–32, 2001. doi:10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

G.W. Brier. Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, 78(1):1 – 3, 1950. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

A. Burke, N. Snook, D.J. Gagne II, S. McCorkle, and A. McGovern. Calibration of machine learning–based probabilistic hail predictions for operational forecasting. *Weather and Forecasting*, 35(1):149 – 168, 2020. doi:10.1175/WAF-D-19-0105.1. URL https://journals.ametsoc.org/view/journals/wefo/35/1/waf-d-19-0105.1.xml.

M. Chantry, H. Christensen, P. Dueben, and T. Palmer. Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft ai. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200083, 2021. doi:10.1098/rsta.2020.0083. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0083.

F. Chollet. Deep learning with python. Technical report, 2018. URL https://github.com/fchollet/keras.

Francois Chollet et al. Keras, 2015. URL https://github.com/fchollet/keras.

B. E. Coffer, M. Kubacki, Y. Wen, T. Zhang, C. A. Barajas, and M. K. Gobbert. Using machine learning techniques for supercell tornado prediction with environmental sounding data. 2020.

C. Donald Ahrens and R. Henson. *Meteorology today: an introduction to weather, climate and the environment*. Cengage Learning: Boston, 2015. ISBN 978-1-305-11358-9.

C. A. Doswell. Societal impacts of severe thunderstorms and tornadoes: lessons learned and implications for europe. *Atmospheric Research*, 67-68:135–152, 2003. ISSN 0169-8095. doi:https://doi.org/10.1016/S0169-8095(03)00048-6. URL https://www.sciencedirect.com/science/article/pii/S0169809503000486. European Conference on Severe Storms 2002.

F. Fabry. *Meteorology and radar*, page 1–7. Cambridge University Press, 2015. doi:10.1017/CBO9781107707405.002.

M. Feldmann, U. Germann, M. Gabella, and A. Berne. A characterisation of alpine mesocyclone occurrence. *Weather and Climate Dynamics*, 2(4):1225–1244, 2021. doi:10.5194/wcd-2-1225-2021. URL https://wcd.copernicus.org/articles/2/1225/2021/.

K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In S. Amari and Michael A. Arbib, editors, *Competition and Cooperation in Neural Nets*, pages 267–285, Berlin, Heidelberg, 1982. Springer Berlin Heidelberg. ISBN 978-3-642-46466-9.

D. J. Gagne, A. McGovern, S.E. Haupt, R.A. Sobash, J.K. Williams, and M. Xue. Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and Forecasting*, 32(5):1819 – 1840, 2017. doi:10.1175/WAF-D-17-0010.1. URL https://journals.ametsoc.org/view/journals/wefo/32/5/waf-d-17-0010_1.xml.

D. J. Gagne II, S. E. Haupt, D. W. Nychka, and G. Thompson. Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, 147(8):2827 – 2845, 2019. doi:10.1175/MWR-D-18-0316.1. URL https://journals.ametsoc.org/view/journals/mwre/147/8/mwr-d-18-0316.1.xml.

U. Germann, M. Boscacci, M. Gabella, and M. Schneebelie, 2016.

H. He and Y. Ma. Imbalanced learning: Foundations, algorithms, and applications. *Imbalanced Learning: Foundations, Algorithms, and Applications*, 06 2013. doi:10.1002/9781118646106.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

A. Hering, C. Morel, G. Galli, S. Senesi, P. Ambrosetti, and M. Boscacci. Nowcasting thunderstorms in the alpine region using a radar based adaptive thresholding scheme. page 206–211, 01 2004.

A. Hering, M. Boscacci, U. Germann, and S. Senesi. Operational nowcasting of thunderstorms in the alps during map d-phase. 2008.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi:10.1162/neco.1997.9.8.1735.

P. Hoeppe. Trends in weather related disasters – consequences for insurers and society. *Weather and Climate Extremes*, 11:70–79, 2016. ISSN 2212-0947. doi:https://doi.org/10.1016/j.wace.2015.10.002. URL https://www.sciencedirect.com/science/article/pii/S2212094715300347. Observed and Projected (Longer-term) Changes in Weather and Climate Extremes.

R. A. Houze. Chapter 8 - cumulonimbus and severe storms. In R. A. Houze, editor, *Cloud Dynamics*, volume 104 of *International Geophysics*, pages 187–236. Academic Press, 2014. doi:https://doi.org/10.1016/B978-0-12-374266-7.00008-1. URL https://www.sciencedirect.com/science/article/pii/B9780123742667000081.

W. Hsu and A.H. Murphy. The attributes diagram a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, 2(3):285–293, 1986. ISSN 0169-2070. doi:https://doi.org/10.1016/0169-2070(86)90048-8. URL https://www.sciencedirect.com/science/article/pii/0169207086900488.

H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33, 2019. doi:10.1007/s10618-019-00619-1. URL https://doi.org/10.1007/s10618-019-00619-1.

A. Jarvis, E. Guevara, H.I. Reuter, and A.D. Nelson. Hole-filled srtm for the globe : version 4 : data grid, 2008. URL https://srtm.csi.cgiar.org/srtmdata/. Published by CGIAR-CSI on 19 August 2008.

G. E. Jergensen, A. McGovern, R. Lagerquist, and T. Smith. Classifying convective storms using machine learning. *Weather and Forecasting*, 35(2):537 – 559, 2020. doi:10.1175/WAF-D-19-0170.1. URL https://journals.ametsoc.org/view/journals/wefo/35/2/waf-d-19-0170.1.xml.

E. Kabir, S. Guikema, and S. Quiring. Predicting thunderstorm-induced power outages to support utility restoration. *IEEE Transactions on Power Systems*, PP:1–1, 05 2019. doi:10.1109/TPWRS.2019.2914214.

S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Moncef Gabbouj, and D.J. Inman. 1d convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, 151:107398, 2021. ISSN 0888-3270. doi:https://doi.org/10.1016/j.ymssp.2020.107398. URL https://www.sciencedirect.com/science/article/pii/S0888327020307846.

F. P. Kuhl and C. R. Giardina. Elliptic fourier features of a closed contour. *Computer Graphics and Image Processing*, 18, 1981. doi:0146-664X/82/030236-2302.00/0.

R. Lagerquist, A. McGovern, C. R. Homeyer, D. J. Gagne II, and T. Smith. Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Monthly Weather Review*, 148(7):2837 – 2861, 2020. doi:10.1175/MWR-D-19-0372.1. URL https://journals.ametsoc.org/view/journals/mwre/148/7/mwrD190372.xml.

U. Lohmann, F. Lüönd, and F. Mahrt. *Storms and cloud dynamics*, page 285–322. Cambridge University Press, 2016. doi:10.1017/CBO9781139087513.011.

S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles. *CoRR*, abs/1802.03888, 2018. URL http://arxiv.org/abs/1802.03888.

I. Mason. A model for assessment of weather forecasts. *Australian Meteorological Magazine*, 30:291–303, 1982.

B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, 1975. ISSN 0005-2795. doi:https://doi.org/10.1016/0005-2795(75)90109-9. URL https://www.sciencedirect.com/science/article/pii/0005279575901099.

M. J. Molina, D. J. Gagne, and A. F. Prein. Deep learning classification of potentially severe convective storms in a changing climate. *Earth and Space Science Open Archive*, page 28, 2020. doi:10.1002/essoar.10504498.1. URL https://doi.org/10.1002/essoar.10504498.1.

A. Mostajabi, D. L. Finney, M. Rubinstein, and F. Rachidi. Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques.

Climate and Atmospheric Science, 2, 2019. doi:10.1038/s41612-019-0098-0. URL https://doi.org/10.1038/s41612-019-0098-0.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

L. Nisi, A. Hering, U. Germann, and O. Martius. A 15-year hail streak climatology for the alpine region. *Quarterly Journal of the Royal Meteorological Society*, 144(714):1429–1449, 2018. doi:https://doi.org/10.1002/qj.3286. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3286.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau. Scikit-learn: Machine learning in python. Technical report, 2011. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0083.

T. Pucik, P. Groenemeijer, A. T. Radler, L. Tijssen, G. Nikulin, A. F. Prein, E. van Meijgaard, R. Fealy, D. Jacob, and C. Teichmann. Future changes in european severe convection environments in a regional climate model ensemble. *Journal of Climate*, 30(17):6771 – 6794, 2017. doi:10.1175/JCLI-D-16-0777.1. URL https://journals.ametsoc.org/view/journals/clim/30/17/jcli-d-16-0777.1.xml.

T. H. Raupach, A. Martynov, L. Nisi, A. Hering, Y. Barton, and O. Martius. Object-based analysis of simulated thunderstorms in switzerland: application and validation of automated thunderstorm tracking with simulation data. *Geoscientific Model Development*, 14(10):6495–6514, 2021. doi:10.5194/gmd-14-6495-2021. URL https://gmd.copernicus.org/articles/14/6495/2021/.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi:10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

P. J. Roebber. Visualizing multiple measures of forecast quality. *Weather and Forecasting*, 24(2):601 – 608, 2009. doi:10.1175/2008WAF2222159.1. URL https://journals.ametsoc.org/view/journals/wefo/24/2/2008waf2222159_1.xml.

A. Ruiz-Gazen and N. Villa. Storms prediction : Logistic regression vs random forest for unbalanced data. *Case Studies in Business, Industry and Government Statistics*, 1(2):91–101, November 2007. URL https://hal.archives-ouvertes.fr/hal-00270176.

A. T. Rädler, P. H. Groenemeijer, E. Faust, R. Sausen, and T. Púčik. Frequency of severe thunderstorms across europe expected to increase in the 21st century due to rising instability. *npj Climate and Atmospheric Science*, 2019, 2019. doi:10.1038/s41612-019-0083-7. URL https://doi.org/10.1038/s41612-019-0083-7.

L. S. Shapley. *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, 2016. doi:doi:10.1515/9781400881970-018. URL https://doi.org/10.1515/9781400881970-018.

Swisstopo. URL https://www.swisstopo.admin.ch/de/geodata/height/dhm25200.html.

M. Taszarek, J. Allen, T. Pucik, P. Groenemeijer, B. Czernecki, L. Kolendowicz, K. Lagouvardos, V. Kotroni, and W. Schulz. A climatology of thunderstorms across europe from a synthesis of multiple data sources. *Journal of Climate*, 32(6):1813 – 1837, 2019. doi:10.1175/JCLI-D-18-0372.1. URL https://journals.ametsoc.org/view/journals/clim/32/6/jcli-d-18-0372.1.xml.

K. Wapler. Mesocyclonic and non-mesocyclonic convective storms in germany: Storm characteristics and life-cycle. *Atmospheric Research*, 248:105186, 2021. ISSN 0169-8095. doi:https://doi.org/10.1016/j.atmosres.2020.105186. URL https://www.sciencedirect.com/science/article/pii/S0169809520311224.

K. Wapler, T. Hengstebeck, and P. Groenemeijer. Mesocyclones in central europe as seen by radar. *Atmospheric Research*, 168:112–120, 2016. ISSN 0169-8095. doi:https://doi.org/10.1016/j.atmosres.2015.08.023. URL https://www.sciencedirect.com/science/article/pii/S0169809515002719.

A. Weiss. Topographic position and landforms analysis. 200, 2001.

T. Weusthoff. Weather type classification at meteoswiss : Introduction of new automatic classification schemes. Technical report, 2011.

D. Wolfensberger, M. Gabella, M. Boscacci, U. Germann, and A. Berne. Rainforest: a random forest algorithm for quantitative precipitation estimation over switzerland. *Atmospheric Measurement Techniques*, 14(4):3169–3193, 2021. doi:10.5194/amt-14-3169-2021. URL https://amt.copernicus.org/articles/14/3169/2021/.