Predicting rotation in supercell thunderstorms using supervised machine learning

Louise Aubet (Master Thesis from 09.2021 to 01.2022)
Supervisors : Monika Feldmann, Alexis Berne

Motivation

Background

- Supercell thunderstorms are severe weather phenomena that can cause serious damage.
- No existing machine learning algorithm is used to predict rotation in thunderstorms.
- Mesocyclone Detection Algorithm relies on Doppler velocity data.

Goals

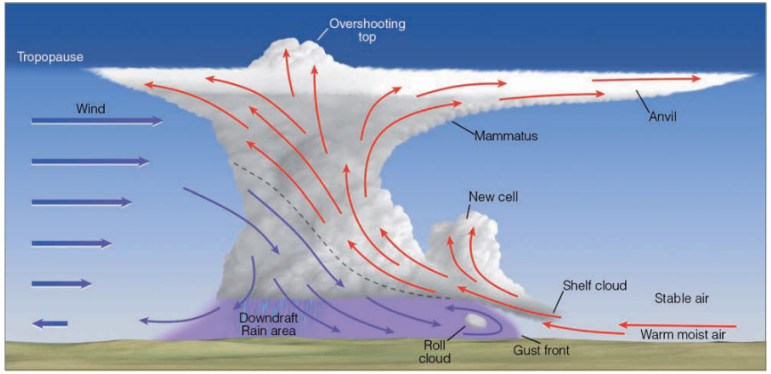
- Better predict rotation in thunderstorms in the case of limited radar data volume, without Doppler velocity data.
- Provide additional hazard assessment information on thunderstorms in the absence of a mesocyclone detection algorithm.

Introduction: Thunderstorms

Thunderstorms: convective storms

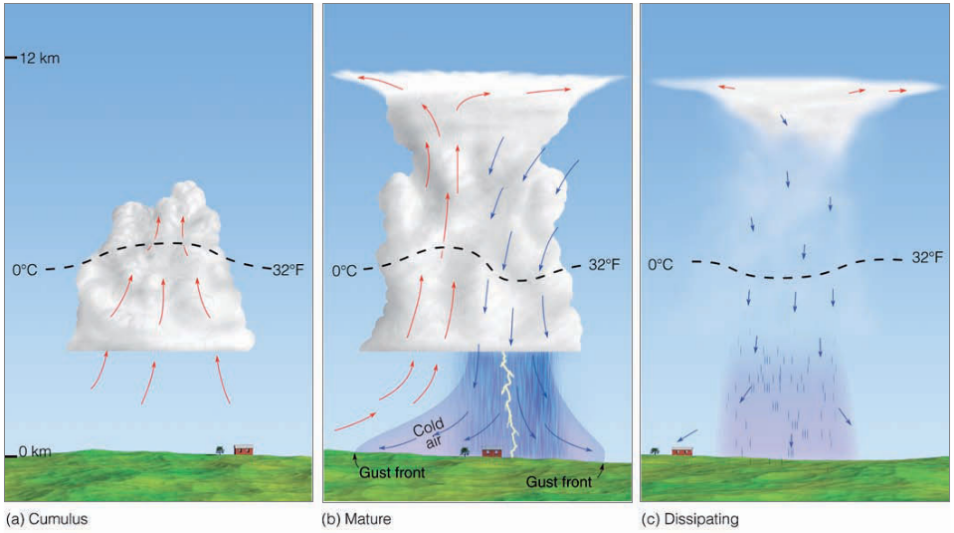
Different types:

- Single-cell thunderstorms
- Multi-cell thunderstorms
- Supercell thunderstorms



Multi-cell thunderstorm

Life cycle of a single cell thunderstorm



Source: Donald Ahrens and Henson, 2015

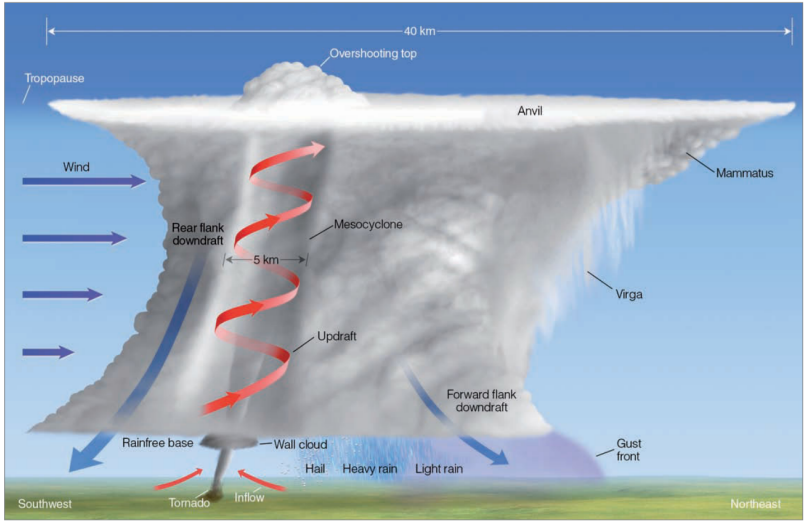
Introduction: Supercells

Mesocyclone: deep persistently rotating updraft

Image from Meteoswiss Blog

Characteristics:

- Require strong wind shear
- Annual and diurnal cycle: maximum around June-July and in late afternoon (Feldmann et al., 2021, Wapler et al., 20216)
- Tend to be larger, faster and last longer than non-mesocyclonic storms (Wapler, 2021)



Source: Donald Ahrens and Henson, 2015

Data

Output from detection algorithms in the Rad4Alp network:

Thunderstorm Radar Tracking algorithm (Hering et al., 2008):

- General information: storm ID, time
- Position: longitude/latitude
- Thunderstorm intensity variables: area, propagation speed, ET15, ET45, VIL, reflectivity, precipitation, POH, MESHS, RANK

RANK: heuristic attribute evaluating severity of a storm cell

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

Mesocyclone Detection Algorithm (Feldmann et al., 2021):

- Class label for each timestep (positive or negative rotation)
- Rotational properties: rotational velocity, vorticity, diameter, rank, altitude

Data

Additional data:

- Synoptic weather class

(from Meteoswiss)

- Topographical data: aspect, slope, TPI, altitude

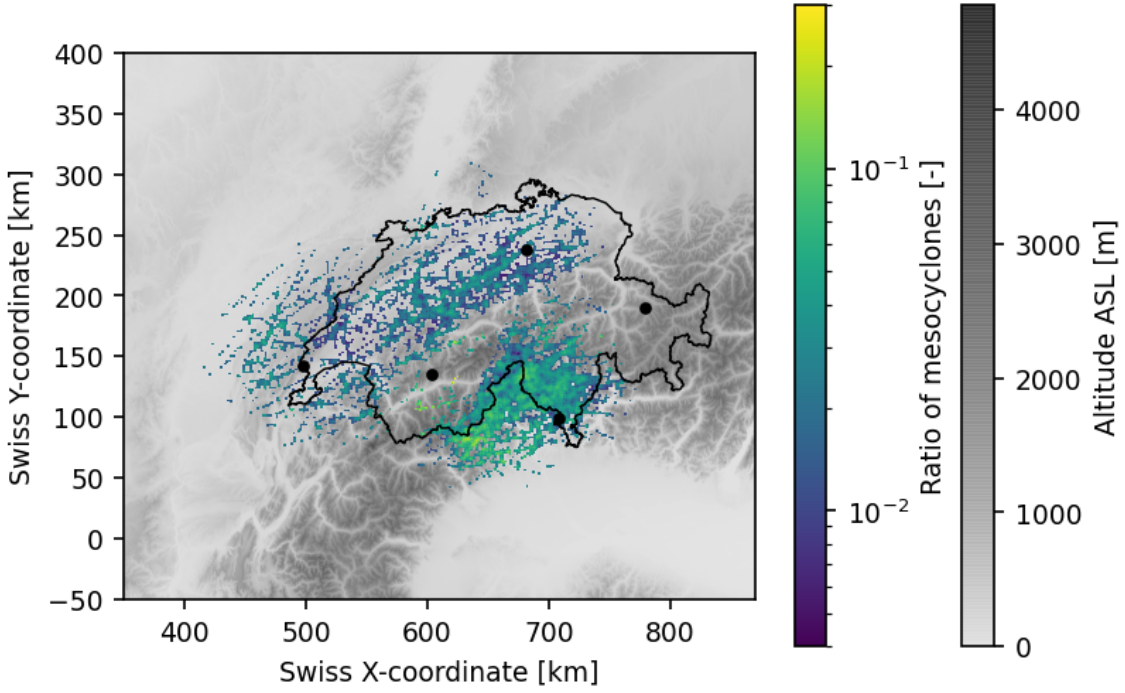
(from Swisstopo, Jarvis et al., 2008)

- Relative quality index

(from Feldmann et al., 2021)

Dataset:

- 6 years: 2016-2021
- Convective season: April-October
- 5 min timesteps
- Extremely unbalanced: 0.26%



	Number of detections	Number of storm tracks	Number of active days
Mesocyclones	6′717	584	92
Thunderstorms	2′518′600	594′060	216
Total	2′525′317	594′066	216

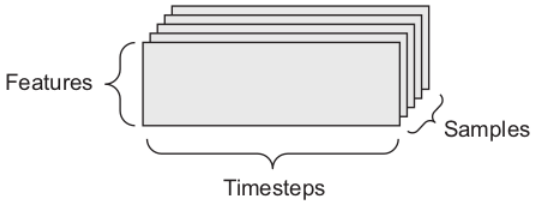
Methods

First part:

- A **Random Forest** is used to classify timesteps as mesocyclonic or non-mesocyclonic.
- It is compared to a Logistic Regression (Machine Learning baseline).

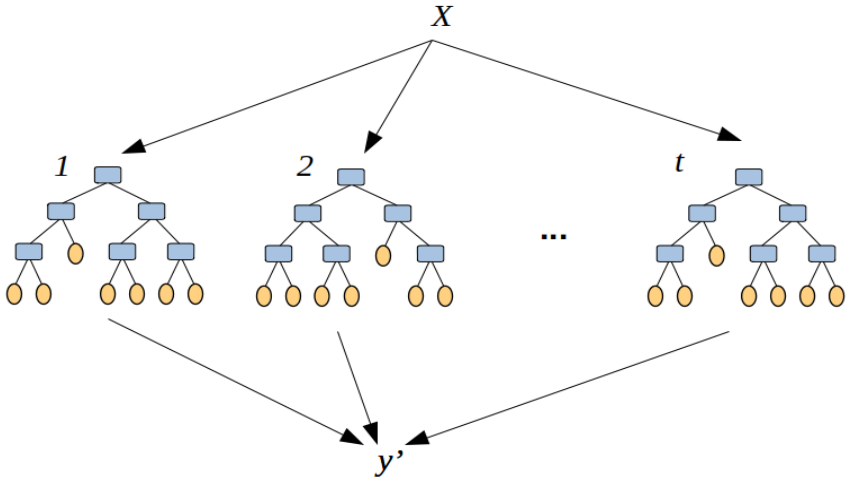
Second part:

- A **Convolutional Neural Network** is used to classify whether or not storms contain at least one mesocyclonic timestep.
- It is compared to a basic Artificial Neural Network (ANN), the Machine Learning baseline, and to a meteorological baseline using a threshold on the RANK.



Methods

Random Forest (RF): ensemble learning method that combines decision trees and bagging.



Steps:

- Bootstrap sampling to obtain K subsets.
- Grow K trees, randomly select some features for each node.
- Take majority vote among trees.

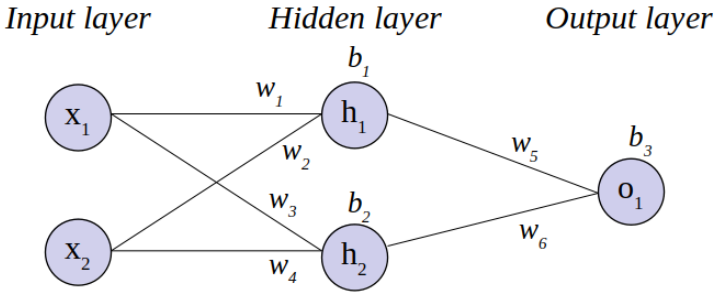
Advantages:

- Reduces variance through ensemble aggregation.
- Reduces bias through depth of trees.
- Hyperparameter tuning is simple.

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

Methods

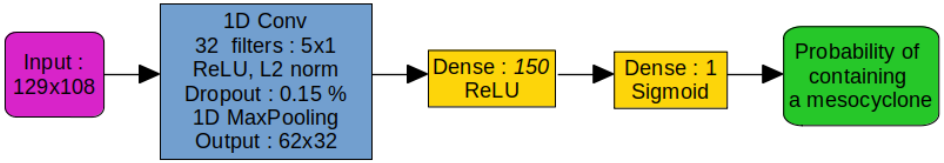
Convolutional Neural Network (CNN): one type of Artificial Neural Network.



Properties:

- Convolutional layer: one neuron only influences a limited amount of neurons in the next layer, the ones that are close to it.

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



Advantages:

- Fewer parameters so faster to train.
- Less complex and less connected so less prone to overfitting.

Methods

Data preparation:

- Positive and negative rotation together

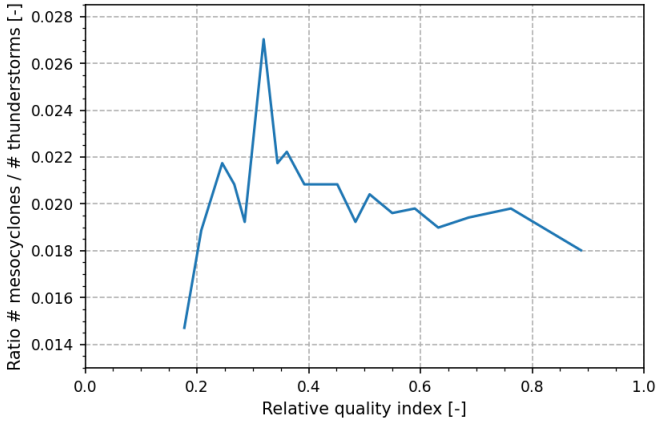
Two classes: 1 for rotation and 0 for no rotation.

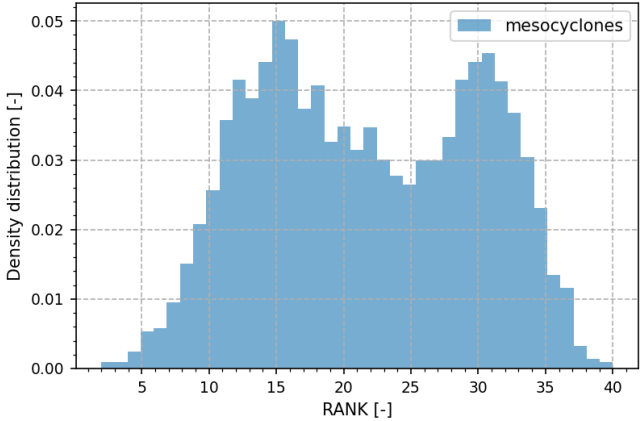
- Threshold on the relative quality index

Only data points with quality index > 0.3.

- Threshold on the RANK

Only severe thunderstorms: RANK > 25.
A consequence is the mitigation of unbalance: ratio = 27%.

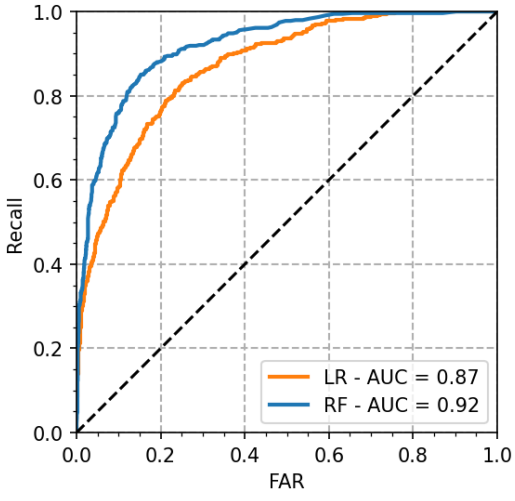


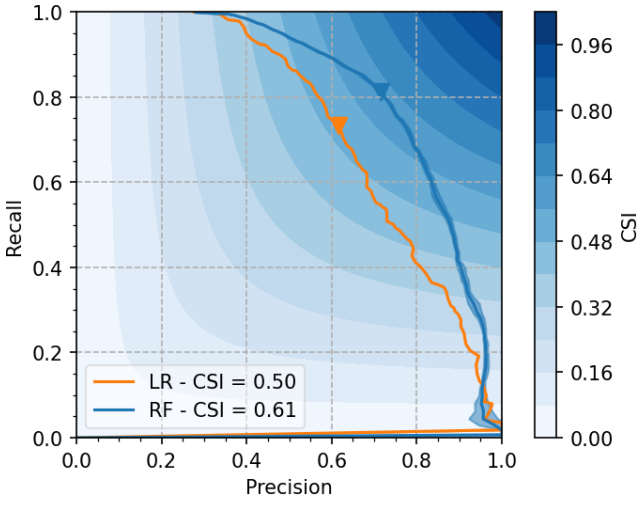


Results: timestep classification

Random Forest is compared to Logistic Regression.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \qquad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$





ROC curve:

- FAR artificially low due to the number of TN.

Performance diagram:

- Wider gap between the models.
- Validation test is representative of the test set.

Results: timestep classification

Logistic Regression

[%]	Real class 1	Real class 0
Predict class 1	20	12
Predict class 0	8	60

Random Forest

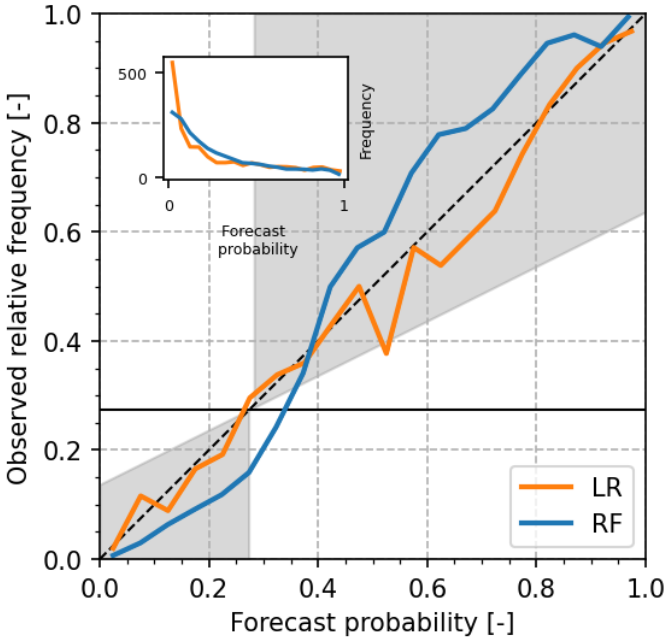
[%]	Real class 1	Real class 0
Predict class 1	22	9
Predict class 0	5	64

Confusion matrices:

- Ratio of FN is 1.6 times lower and ratio of FP is 1.3 times lower for RF rather than for LR.

Reliability diagram:

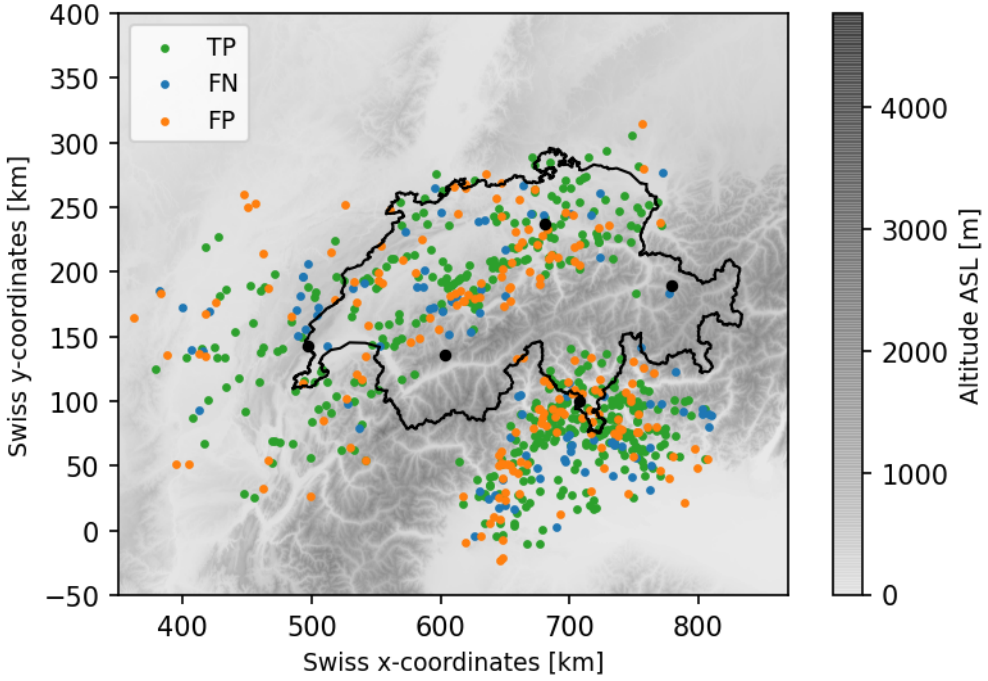
- RF is over-forecasting rare events and under-forecasting more frequent events.



Results: timestep classification

Spatial distribution of predictions:

- No obvious spatial biases.
- Higher concentration of FP and FN around Monte Lema, because higher concentration of mesocyclones in this region.



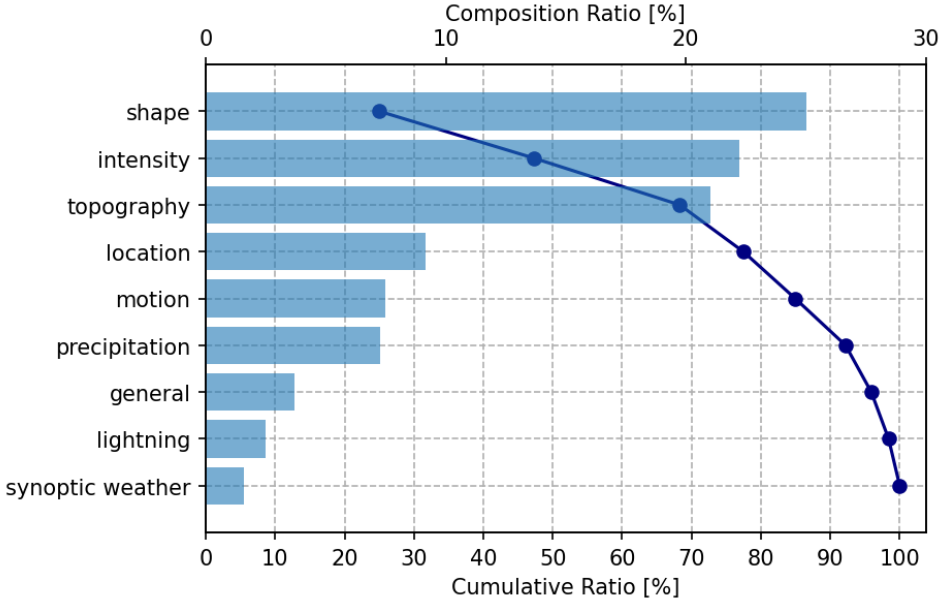
Results: timestep classification

Feature importance:

- Using SHAP value: method to explain output of ML models using coalition game theory (Lundberg et al., 2017)

Aggregation in categories:

- Shape: area, ellipse axis
- Intensity: RANK, detection threshold, composite reflectivity, echotop altitude
- Topography: altitude, slope, aspect, TPI
- Precipitation: POH, MESHS, VIL, combiprecip
- Motion: x and y velocities



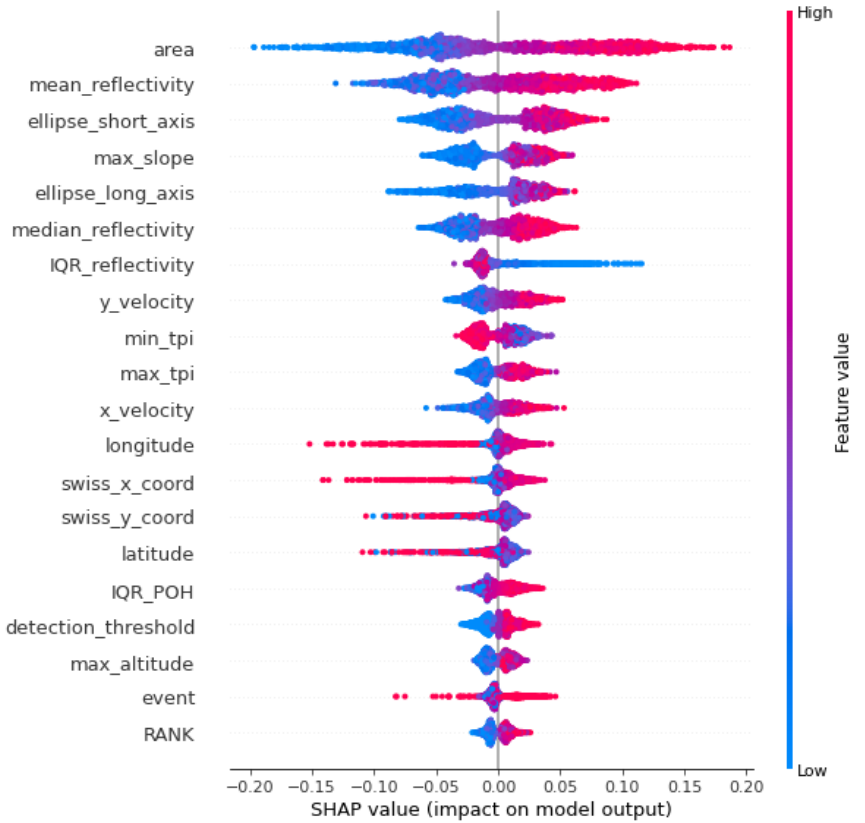
Results: timestep classification

Feature importance:

Summary plot: shows feature importance and feature effect

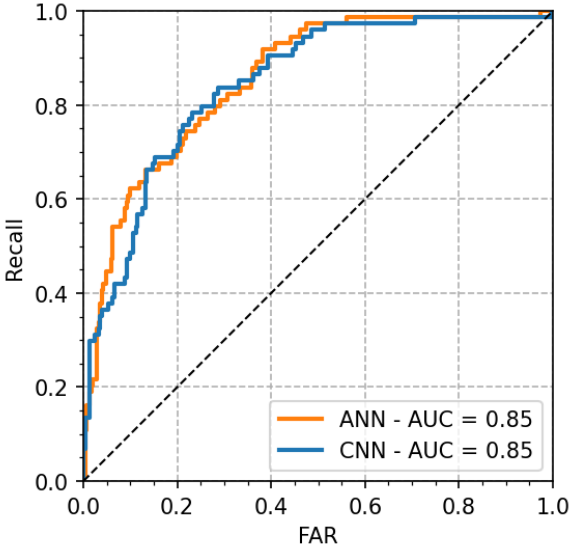
Area, velocity and reflectivity have a positive impact : higher values increase the probability of a storm cell to be mesocyclonic.

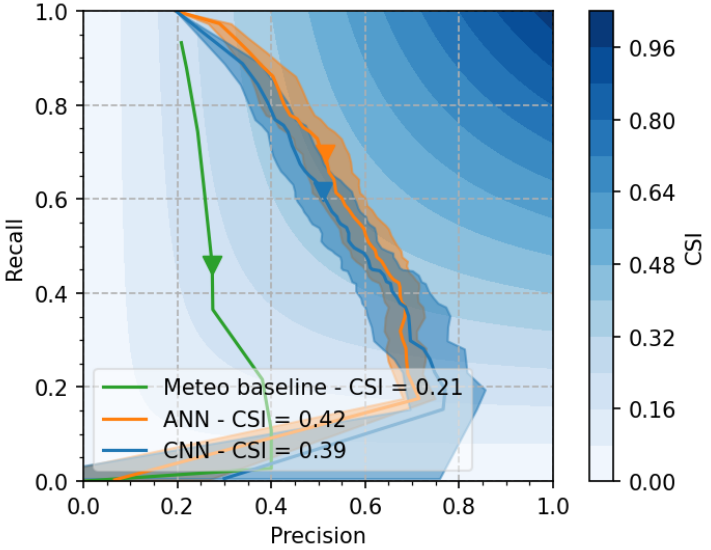
Coherent with physical characteristics of storm over the entire life cycle: larger, faster, higher refectivity (Wapler, 2021).



Results: storm classification

- CNN performs better than the meteorological baseline but as well as a basic ANN.
- Difference in performance with the RF can be explained by the size of the dataset which is smaller.





ANN

[%]	Real class 1	Real class 0
Predict class 1	12	6.6
Predict class 0	7.4	74

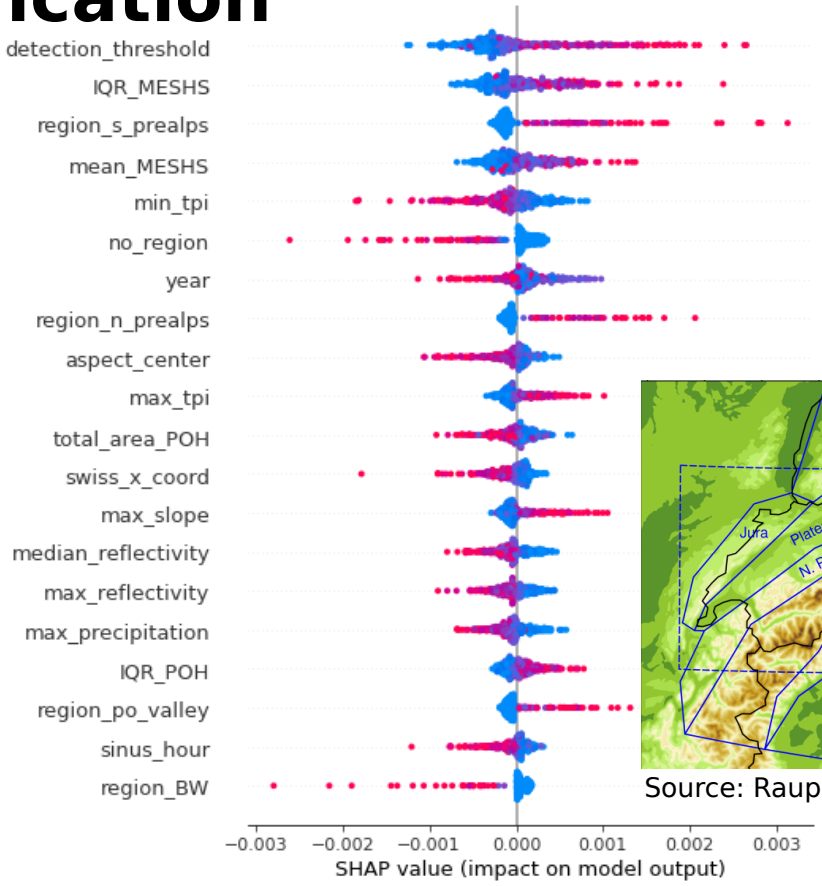
CNN

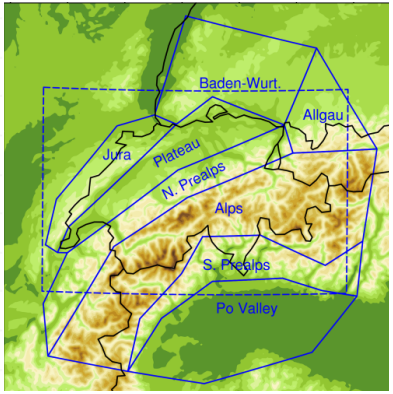
[%]	Real class 1	Real class 0
Predict class 1	15	13
Predict class 0	5	67

Results: storm classification

Feature importance:

- For storms in the Southern, Northern Prealps and in the Po Valley, the probability of being mesocyclonic is higer than the average.
- Coherent with a study showing that the Southern and Northern Prealps are the most active regions in Switzerland in terms of thunderstorms and mesocyclones (Feldmann et al., 2021).
- Being located in Baden-Würt. and outside of these regions is negatively correlated to the SHAP value.





Source: Raupach et al., 2021

Conclusion

- The Random Forest performs better than the Logistic Regression.
- Shape features are the most important for the Random Forest and the effect of the features is in accordance with previous studies.

- The CNN performs as well as the basic ANN and both models perform better than the meteorological baseline.
- Location features are more important for the CNN and the effect of the features is in accordance with previous studies.

Outlook:

- Add more features to the model (i.e. near-storm environment data).
- Build machine learning model to predict rotation intensity or rotation tendency in the next timesteps.

References

- Feldmann, M., Germann, U., Gabella, M., and Berne, A.: A Characterisation of Alpine Mesocyclone Occurrence, Weather Clim. Dynam. Discuss., in review, 2021

- Wapler, K. : Mesocyclonic and non-mesocyclonic convective storms in Germany: Storm characteristics and life-cycle, Atmospheric Research, 2021

- Wapler K., Hengstebeck T., Groenemeijer P., Mesocyclones in Central Europe as seen by radar, Atmospheric Research, 2016

- Hering A., Morel C., Galli G., Senesi S., Ambrosetti P. & Boscacci M. : Nowcasting thunderstorms in the Alpine region using a radar based adaptive thresholding scheme, Proc. ERAD Conference, 2004

- D. Wolfensberger, M. Gabella, M. Boscacci, U. Germann, and A. Berne. Rainforest: a random forest algorithm for quantitative precipitation estimation over Switzerland. AtmosphericMeasurement Techniques, 2021

- Nisi, L., Hering, A., Germann, U. and Martius, O.. A 15-year hail streakclimatology for the alpine region. Quarterly Journal of the Royal Meteorological Society, 2018.

- C. Donald Ahrens and R. Henson., Meteorology today: an introduction to weather, climate and the environment. Cengage Learning: Boston, 2015.

Thank you for your attention.
Any questions ?