On hail in Switzerland – crowdsourcing, nowcasting and multi-day hail clusters

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von

Hélène Christine Louise Barras

von Broc, FR

Leiterin der Arbeit:

Prof. Dr. Olivia Romppainen-Martius

Geographisches Institut, Universität Bern

Ko-Leiter der Arbeit:

Dr. Urs Germann

MeteoSchweiz

On hail in Switzerland – crowdsourcing, nowcasting and multi-day hail clusters

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von

Hélène Christine Louise Barras

von Broc, FR

Leiterin der Arbeit:

Prof. Dr. Olivia Romppainen-Martius

Geographisches Institut, Universität Bern

Ko-Leiter der Arbeit:

Dr. Urs Germann

MeteoSchweiz

Von der Philosophisch-naturwissenschaftlichen Fakultät angenommen.

Bern, den 29. Juni 2021

Der Dekan:
Prof. Dr. Z. Balogh

Supervisors and advisors:
Prof. Dr. Olivia Romppainen-Martius
Institute of Geography
University of Bern

Dr. Urs Germann
Division for Radar, Satellite and Nowcasting
MeteoSwiss

Dr. Alessandro M. Hering
Division for Radar, Satellite and Nowcasting
MeteoSwiss

External co-referee:
Prof. Dr. Russ Schumacher
Department of Atmospheric Science
Colorado State University

This thesis was carried out as a collaborative project between the University of Bern and MeteoSwiss. The complete list of affiliations is the following:

- Mobiliar Lab for Natural Risks, University of Bern
- Institute of Geography, University of Bern
- Oeschter Centre for Climate Change Research, University of Bern
- Division for Radar, Satellite and Nowcasting, MeteoSwiss

Abstract

Hail is one of the costliest atmospheric hazards in Switzerland, causing substantial damage to crops, cars, buildings, and infrastructure every year. Currently, the Swiss population is warned operationally about thunderstorms, but no information is given on specific hazards such as hail, severe wind gusts and lightning. One reason is the gap in ground-based hail observations, without which predictions could not be verified. To address these gaps, this dissertation presents a multi-approach advancement to hail prediction. Three projects explore crowdsourced hail reports, create hail nowcasting models and characterize large- and local scale atmospheric conditions of multi-day hail clusters.

To close the gap in available ground-based hail observations, the first part of this dissertation uses crowdsourced hail size reports submitted via mobile application of the Swiss Federal Office for Meterorology and Climatology (MeteoSwiss). The reporting function was added in May 2015 and has collected more than 100'000 reports since. These reports are explored, filtered using an automatic plausibility filtering method based mainly on three criteria, and compared to two operational radar-based hail algorithms. The most important criterion guarantees a minimum proximity of reports to thunderstorms. Other criteria remove duplicate reports and artificial patterns and limit the time difference between the event time and the report submission time. If "no hail" reports are excluded, 53 % of reports collected until September 2020 remain after filtering. The comparison of crowdsourced hail reports with the algorithms probability of hail (POH) and maximum expected severe hail size (MESHS) indicates that some hail events were missed by the algorithms. While there is significant variability between size categories, the matched reports and radar-based algorithms correlate positively. MESHS values are typically 1.5 cm larger than the reported sizes. This study shows that the crowdsourced reports are invaluable for hail research and suggest that crowdsourcing could be applied to other atmospheric hazards.

In the second part of this dissertation, extreme gradient boosted tree (XGBoost) models are developed to nowcast the occurrence and size of hail for individual thunderstorms. Statistics of environmental variables from radar, satellite, lightning, topography and numerical weather models serve as features (also called predictors) to predict the maximum POH and MESHS, in 5-minute time steps, up to 45 minutes in advance. For each lead-time, binary XGBoost models predict the occurrence of hail (POH $\geq$ 10 %, MESHS $\geq$ 2 cm) and, subsequently, linear XGBoost models predict the non-zero maximum POH and MESHS values. Additional models with a reduced number of input features assess how many features are needed to reach the same nowcast quality as models using all features. The binary XGBoost models predict the occurrence of hail better than the Lagrangian persistence for all lead-times $\geq$ 10 minutes. For a lead-time of 5 minutes, both predictions skills are equal. About 500–1000 features are necessary to reach a similar skill to models that used all features. Although all data sources are present in the top 100 features, radar-based features are the most important. Features indicating an intense

thunderstorm activity at the most recent time step increase the probability of POH $\geq$ 10 % and MESHS $\geq$ 2cm. The Lagrangian persistence predict the POH values with a smaller standardized centered root-mean squared error than linear XGBoost models, up to a lead-time of 25 minutes. A likely reason is the smaller sample size used to train and test linear XGBoost models. This chapter demonstrated the effectiveness of machine learning in nowcasting and will serve as a base for future projects.

Multi-day hail clusters cause significant damage in a short time. To increase their predictability, the third part of this dissertation explores the large- and local-scale atmospheric conditions during and up to three days before multi-day hail clusters and isolated hail days. Hail days between 2002–2019 are defined for two regions, north and south of the Alps, within 140 km of the Swiss radar network. The conditions are described using a weather type classification, reanalysis data, objectively identified fronts and atmospheric blocks. For both regions, composite atmospheric variables indicated a more stationary and meridionally amplified atmospheric flow during multi-day hail clusters. North of the Alps, blocks are more frequent over the North Sea and surface fronts are located farther from Switzerland on clustered hail days than on isolated hail days. Furthermore, clustered hail days are characterized by significantly higher convective available potential energy (CAPE) values, warmer daily maximum surface temperatures, and a higher atmospheric moisture content than isolated hail days. South of the Alps, these differences in CAPE, temperature and moisture are not as significant. However, the mean sea level pressure is significantly deeper on isolated hail days. For both regions, the Rossby waves are already more amplified three days before multi-day hail clusters, than before isolated hail days. Furthermore, prior to more than 10 % of clustered hail days, atmospheric blocks occur over Scandinavia, which is not the case for isolated hail days. This chapter shows that the temporal clustering of hail days is coupled to specific large- and local-scale flow conditions, providing an added value for short- to medium-range forecasts of hail in Switzerland. Furthermore, the conditions during multi-day hail clusters north of the Alps raise the question, whether multi-day hail clusters may occur more frequently with global warming.

Altogether, this dissertation explores a way of closing the hail observation gap, creates nowcasting models for hail, which could lead to an operational hail warning system, and characterizes the atmospheric conditions during and before multi-day hail clusters and isolated hail days. The latter provides an added value for hail forecasts. This dissertation makes a further step towards warning the Swiss population of hail and preventing its damage.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

This motivation focuses mainly on the activities related to severe weather and hail in Switzerland. A plethora of key literature on hail from other regions in the world are presented in more detail in the individual chapter's introductions.

Hail regularly occurs in the summer season and is among the costliest natural hazards in Switzerland (VKF, 2013; FOEN, 2016). It destroys crops, threatens aviation and damages cars, buildings and infrastructure. The cost of hail damage has increased over the past decades, because the number of buildings has increased and because new infrastructure is more valuable and vulnerable (FOEN, 2016). The increased vulnerability stems from the rising pressure on building envelopes to be more performant, combined with a significant decrease in uncertainty around structural safety margins in the past decades. This development has essentially enabled the use of building materials with safety factors closer to the minimum allowable levels (Madsen et al., 2006; Stucki and Egli, 2007; Maydl and Schulter, 2013). Roller shutters, solar panels and more delicate facade insulations are typical examples of newer building components that are vulnerable to hail (Donner, 2020).

A state of the art operational hail warning system can contribute to reducing damage. Such warning systems prevent damage, for example by rolling up roller shutters automatically and alerting the population to the hailstorms. Developing hail warning systems is, however, not straightforward. Modern high-resolution numerical weather prediction systems struggle with simulating the exact location and time of thunderstorms correctly (e.g., James et al., 2018). Reasons are that the onset location and time of convection is stochastic and can depend on small variations in boundary layer- and surface temperature and moisture (Crook, 1996; Trefalt et al., 2018). Like the butterfly effect (Lorenz, 2000), small differences in initial conditions may determine whether a thunderstorm develops and creates large hail, or no hail at all. Complex orography further enhances, reduces or anchors storms in a way that is very specific to the storm's location, time, severity and movement (e.g., Lean et al., 2009; Barrett et al., 2015; Trefalt et al., 2018; Bachmann

et al., 2020; Heim et al., 2020).

Thunderstorms are detected in real-time in data from five weather radars, located in different regions in Switzerland (Fig. 1.1). The thunderstorm radar tracking algorithm (TRT; Hering et al., 2008) automatically tracks thunderstorms and estimates their intensity in real-time. The thunderstorms are extrapolated forward in time in the direction of their recent movement (e.g., Fig. 1.1). Since 2005, warnings are issued for the future thunderstorm positions (Panziera et al., 2016, e.g., 1.1). Since 2020, these warnings are provided fully automatically and operationally to the end users via mobile application (app) of the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss; see Hering et al., 2015). So far, these warnings have not provided any information on specific hazards such as wind gusts, lightning or hail.

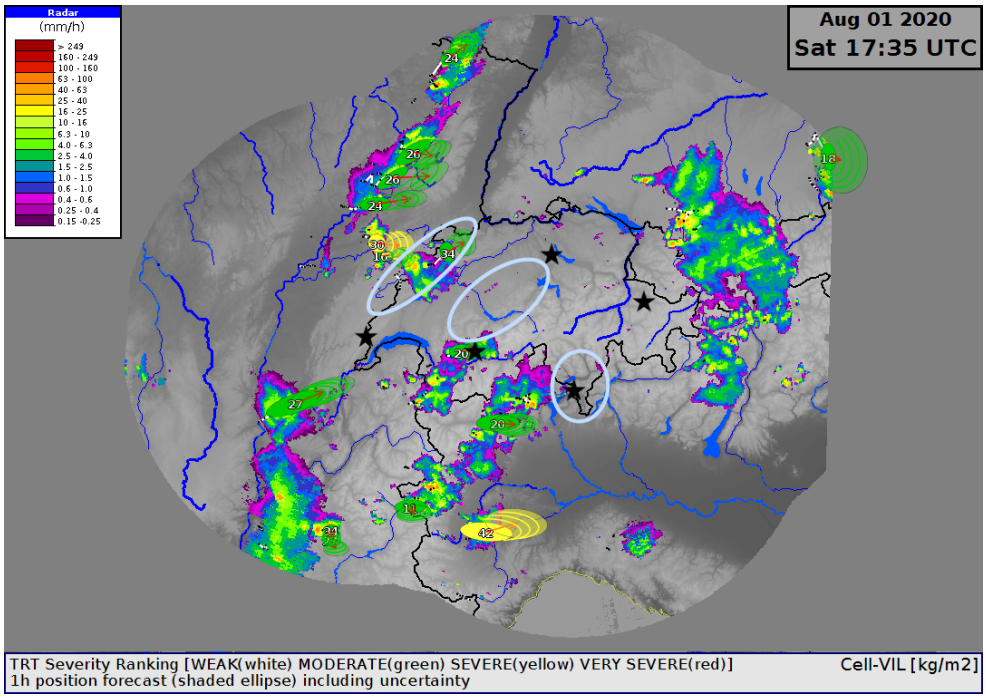

Figure 1.1: Radar image composite from the Swiss radar network for August 1 2020 at 17:35 UTC. Shown are the precipitation intensity (colored contours), the thunderstorm location as tracked by TRT and their future locations (colored filled ellipses) forward extrapolated along the vectors of motion (red arrows). Numbers inside TRT tracked cells indicate the maximum vertically integrated liquid content in kg m^{-2}, black stars show the locations of the five Swiss radar stations and the light blue ellipses the three Swiss hail hotspot regions as detected by Nisi et al. (2016) and NCCS (2021).

Thanks to measurement campaigns in the 1970s (Federer et al., 1986) and 90s (Treloar, 1998), radar-based algorithms were developed as proxies for hail and its size at the ground. Two such algorithms have been operational at MeteoSwiss since 2008 and 2009 and were reprocessed for the years 2002 and after (Nisi et al., 2016). The algorithm probability of hail (POH; Waldvogel et al.,

1979) was developed in the 70s from hail data gathered during the field campaign Grossversuch IV in central Switzerland with hailpads. POH provides an estimate of the probability that hail will occur at the ground. Since its implementation at MeteoSwiss, POH is estimated in quasi real-time from data provided by the Swiss radar network on a 1 by 1 km grid over Switzerland. The second algorithm, Treloar's Maximum Expected Severe Hail Size (MESHS; Treloar, 1998) provides an estimate of the largest hail diameter expected within each 1 by 1 km area.

Nisi et al. (2016), (2019) and (2020) used these algorithms to create climatologies of hail size and hail streaks, i.e. the hail accumulations at the ground, and analyzed their diurnal cycle over the period 2002–2017. The national project "Hail climate Switzerland" (see www.hailclimatology.ch) updated these maps and added new hail climatology products adapted to user needs (NCCS, 2021). The climatologies indicate three hail hotspot regions in Switzerland: in the Jura, in Entlebuch, and in southern Switzerland (Fig. 1.1).

Until recently, developing any hail warning system in Switzerland was a challenge, because no systematic, direct hail observations had been available since the Grossversuch IV. Yet observations of hail at the ground are essential to validate radar-based hail estimates and model-based hail predictions. Importantly, these observations need to be contemporary to the analyzed data period. Insurance damage claims (Bider, 1954; Willemse, 1995) have given valuable evidence that hail of a moderately certain size occurred at the ground. Morel (2014) used car damage claims to validate POH. Furthermore, hail damage to crops can be detected in satellite images (Gallo et al., 2012; Bell and Molthan, 2016; Bell et al., 2020). However, all these damage-based methods only give an imprecise estimate of hail size.

In 2015, the MeteoSwiss and the Mobiliar Lab for Natural Risks of the University of Bern initiated two projects to fill the hail size observation gap. First, a hail crowdsourcing function was added to the MeteoSwiss app, with which users could report the occurrence and approximate size of hail. Since May 2015, $>$ 100'000 reports have been submitted, mostly from populated locations in and close to Switzerland. Second, a pilot network of 11 automatic hail sensors (Löffler-Mang et al., 2011) measured hail stone impacts in 2015–2017. Noti (2016) conducted a first comparison of radar-based algorithms with MeteoSwiss crowdsourced reports and hail sensor data collected in 2015 and 2016. Noti found a positive correlation between the radar-based hail algorithms and the reported size. The sample size of crowdsourced hail reports at that point was, however, too small for stable statistical results on large hail diameters. A more comprehensive comparison of crowdsourced hail reports with radar-based hail algorithms is needed to better understand the reports and the algorithms. The pilot hail sensor network has been extended with 80 new sensors installed in the hail hot spot regions in 2018 (Mobiliar Lab for Natural Risks, 2021). These sensors have successfully captured several hail events and are likely to provide a promising dataset for future hail research.

Past studies have conveyed the importance of atmospheric conditions and processes across dif-

ferent scales to characterize and predict hail in Switzerland. Nisi et al. (2020) found that the topography in Switzerland strongly influences the frequency of hailstorms and the diurnal cycle of convection initiation. Trefalt (2017) documented significant differences in diurnal cycle of the local environmental characteristics between non-hail days, hail days, small hail days and large hail days, north and south of the Alps. On the synoptic-scale, hail in Switzerland preferentially occurs when the flow over central Europe is westerly or southwesterly (Nisi et al., 2018) and up to 45 % of all detected hail cells in northeastern and southern Switzerland form in pre-frontal environments (Schemm et al., 2016). The strong year-to-year variability in hail occurrence suggests that it is strongly controlled by large-scale weather patterns (Nisi et al., 2016). Furthermore, a case study on a severe alpine hailstorm in June 2015 highlighted the interplay of large-scale atmospheric patterns and local processes (Trefalt et al., 2018). Moreover, Madonna et al. (2018) used a Poisson regression approach to model monthly hail occurrences in Northern Switzerland using large-scale environmental variables. The differences in hail environments across the Europe and the Atlantic connected with hail day rich and hail day poor months suggest that large-scale dynamics influence hail day clustering. These studies indicate that hail prediction needs to incorporate processes that are relevant at different scales.

A method that incorporates variables from different scales and sources in prediction models is machine learning (ML). In contrast to numerical weather prediction models, ML uses statistical tools to uncover patterns and knowledge that has not been explicitly programmed (Samuel, 1959; Koza et al., 1996). It is typically used to discover patterns and linear- and non-linear interactions in large data. ML has become very popular in severe weather research for several reasons. The rise in computational power availability and amount of data has made ML more accessible and necessary (Chen and Lin, 2014). Furthermore, convective hazards are typically associated with sub-grid spatial scales and, therefore, implicitly parametrized in conventional operational weather prediction models (e.g., Goyette, 2008; Pennelly et al., 2014; Cassola et al., 2015; Adams-Selin and Ziegler, 2016; Stucki et al., 2016). Parameterization of sub-grid scale phenomena remains one of the greatest challenges in numerical weather modeling (Pielke, 2013). ML provides the option of "learning" to predict convective phenomena using both model simulations and observations (e.g., Miyoshi et al., 2016; McGovern et al., 2017; Bouttier and Marchal, 2020). The use of machine learning in weather modeling is still relatively new but the uptake is accelerating. Marzban and Witt (2001) and Manzato (2013) were among the first to apply ML in hail prediction. Later, several studies have used ML to detect and predict hail (e.g., Gagne et al., 2015, 2017, 2018; McGovern et al., 2017, 2019b; Czernecki et al., 2019; Pullman et al., 2019; Flora et al., 2020; Hill et al., 2020; Yao et al., 2020) and other natural hazards (e.g., Lagerquist et al., 2017, 2020; Herman and Schumacher, 2018a,b; Zhou et al., 2019). In Switzerland, ML methods have been applied to nowcast foehn wind events (Sprenger et al., 2017) and to predict the growth and decay of precipitation (Foresti et al., 2019). However, ML methods have not yet been applied to nowcast hail.

1.2 Aims and outline of this thesis

These projects, past analyses and available methods motivated me to conduct three hail research projects, presented in self-contained articles:

1. The first part of this thesis (chapter 2, Barras et al., 2019) explores the MeteoSwiss crowdsourced hail reports until 2018. Plausibility filters are developed, and the reports are compared systematically to POH and MESHS. Chapter 3 contains an update of the analysis to 2020. The research questions are:

   - What are the characteristics of MeteoSwiss crowdsourced hail reports? How large is the fraction of reports remaining after applying plausibility filters?
   - What is the utility and what are the limitations of the MeteoSwiss crowdsourced reports?
   - How do crowdsourced hail reports compare to the radar-based hail algorithms POH and MESHS?

2. The second part (chapter 4) makes a direct step towards developing automatic hail warnings. The machine learning algorithm XGBoost (extreme gradient boosted trees) is applied to nowcast POH and MESHS for individual thunderstorms. The machine learning models detect interactions between the target variables (POH and MESHS) and more than 10'000 predictor variables extracted from multiple data sources along thunderstorm paths in 2018. This chapter answers the following questions:

   - For which lead-times between 5 and 45 minutes do machine learning models predict the probability and maximum size of hail in Switzerland better than the Lagrangian persistence?
   - How many features are necessary for these XGBoost models to perform well?
   - Which data sources do the XGBoost models use and which features are most important?
   - Which information on thunderstorm environments can we gain using the Shapley Additive Explanations (SHAP) interpretation method?

3. Finally, the third part (chapter 5; Barras et al., 2021 (in review)) focuses on improving the hail prediction through process understanding. The large- and synoptic scale atmospheric conditions during and before multi-day hail clusters are contrasted to the situations during and before isolated hail days. This chapter, addresses the following questions:

   - Which atmospheric conditions are associated with and differentiate multi-day clusters and isolated hail days in Switzerland, north and south of the Alps during 2002–2019?
   - Which atmospheric conditions occur on days before multi-day and isolated hail events?

Besides chapters 2, 3, 4 and 5, the remainder of this thesis is structured as follows: In chapter 6, I summarize the main findings, give some concluding remarks and suggest some ideas for future research avenues. Supporting information to chapters 2, 4 and 5 are presented in Appendices A, B and C.

Chapter 2

Experiences with >50,000 crowdsourced hail reports in Switzerland

This chapter contains an article that was written together with Alessandro Hering, Andrey Martynov, Pascal-Andreas Noti, Urs Germann and Olivia Martius. It was published in 2019 with the title "Experiences with >50,000 crowdsourced hail reports in Switzerland" in the Bulletin of the American Meteorological Society (Barras et al., 2019). The subsequent chapter 3 gives a short update on the crowdsourced reports that were collected until September 2020. Chapter A in the Appendix compares the MeteoSwiss crowdsourced hail reports with other available hail data sets. This comparison was done in response to reviews during the publication of the article.

2.1 Capsule Summary

Fifty-nine thousand crowdsourced hail size reports, gathered in Switzerland since May 2015, are presented, assessed, and compared to two operational radar-based hail detection algorithms.

2.2 Abstract

Crowdsourcing is an observational method that has gained increasing popularity in recent years. In hail research, crowdsourced reports bridge the gap between heuristically defined radar hail algorithms, which are automatic and spatially and temporally widespread, and hail sensors, which provide precise hail measurements at fewer locations. We report on experiences with and first results from a hail size reporting function in the app of the Swiss National Weather Service. App users can report the presence and size of hail by choosing a predefined size category. Since May 2015, the app has gathered >50,000 hail reports from the Swiss population. This is an unprecedented wealth of data on the presence and approximate size of hail on the ground. The reports are filtered automatically for plausibility. The filters require a minimum radar reflectivity value in a neighborhood of a report, remove duplicate reports and obviously artificial patterns, and

limit the time difference between the event and the report submission time. Except for the largest size category, the filters seem to be successful. After filtering, 48% of all reports remain, which we compare against two operationally used radar hail detection and size estimation algorithms, probability of hail (POH) and maximum expected severe hail size (MESHS). The comparison suggests that POH and MESHS are defined too restrictively and that some hail events are missed by the algorithms. Although there is significant variability between size categories, we found a positive correlation between the reported hail size and the radar-based size estimates.

2.3 The hail observation gap

Hail fall in Switzerland at a specific location is infrequent, typically very localized, and characterized by a high spatial variability in hailstone sizes. That said, in the hail hot spots, hail occurs about 2–3 times per square kilometer per year (Nisi et al., 2016; Punge and Kunz, 2016). As a consequence, ground observations require a very dense observational network and are therefore very expensive. Similar challenges exist for hail observations worldwide. Since the 1990s, researchers have attempted to fill the gap by involving the general public in gathering weather observations. Examples include the Community Collaborative Rain, Hail and Snow Network (CoCoRaHS) in North America (Cifelli et al., 2005; Reges et al., 2016), the European Severe Weather Data Base (ESWD, Dotzek et al., 2009), the European Weather Observer application (app) (EWOB, Groenemeijer et al., 2017), and the Mobile Precipitation Identification Near the Ground Project (mPING) mostly in North America (Elmore et al., 2014). Ground observations are essential for developing, verifying, and improving indirect hail detection and hail size estimation algorithms based on remotely sensed data such as weather radar observations.

In Switzerland, two radar-based hail algorithms have been in operation since 2008: the probability of hail (POH) and the maximum expected severe hail size (MESHS). They are used for nowcasting applications and for insurance loss estimates, and they were used to create the first Swiss radar-based hail climatology (Nisi et al., 2016). Recently, the algorithms were used to analyze the initiation and lifetime of hail cells and their swaths in complex topography (Nisi et al., 2018). A first verification of POH in Switzerland by Nisi et al. (2016) is based on insurance car loss data. Insurance loss data primarily provide information on the presence or absence of hail in areas with insured assets; the hail size is estimated from the damage type. However, the claims are often georeferenced to a ZIP code rather than to the actual hail event location. Spatially widespread information regarding the size of hail on the ground has so far been missing in Switzerland.

In May 2015, the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss) started to fill this observational gap by simultaneously launching a pilot network of 11 automatic hail sensors and a hail size crowdsourcing function in the MeteoSwiss app. These hail sensors record the impact of individual hailstones on a Makrolon disk using piezo-electric microphones. The signal correlates positively with the kinetic energy and momentum of the hailstone, and thus, the hailstone diameter can be estimated from these measurements. For more information on the hail sensors, see Löffler-Mang et al. (2011). As of 2018, this pilot network is being extended to include a total of 80 automatic sensors that will measure the kinetic energy and momentum of

hailstones for at least 8 years in the three hail hot spot regions of Switzerland (see Nisi et al., 2016).

The crowdsourced reports, the radar-based hail algorithms, and the automatic hail sensor network combine three sources of hail data that are of great complementary value. The radar hail algorithms provide automatic, spatially and temporally continuous estimates of the likelihood and size of hailstones at the ground. Automatic hail sensors have the advantage of measuring hail at the ground in a precise manner, but only at their exact location. The crowdsourced reports are numerous and account for much larger areas than automatic hail sensors, but provide subjective and less precise information of the true size of hail.

Trefalt et al. (2018) combined these hail data sources, as well as a newly developed dual-polarization radar-based hydrometeor classification (Besic et al., 2016, 2018), in a case study of an intense hailstorm in the northern Prealps. This case study showed good agreement between POH, MESHS, and the hailstone sizes sourced from the MeteoSwiss app. Kunz et al. (2018) and Wapler et al. (2015) emphasized the benefit of combining multiple data sources in similar case studies on hail storms in Germany.

This article will introduce the MeteoSwiss crowdsourced hail reports, demonstrate a strategy to automatically filter them for plausibility, comment on their utility and limitations, and present a comparison to the two radar-based hail algorithms, POH and MESHS.

2.4 Radar and crowdsourced data

2.4.1 Radar-based hail products

We compare the reports with two operational radar-based hail algorithms, i) POH [Foote et al. (2005b) based on Waldvogel et al. (1979)] and ii) MESHS [Joe et al. (2004) based on Treloar (1998)]. POH is a measure for the likelihood of hail occurrence, ranging from 0% to 100%. MESHS estimates the largest expected hail diameter in units of centimeters, starting at 2 cm. In Switzerland, POH and MESHS are used operationally and derived by combining freezing-level height information from the analysis (in real-time applications from the forecast) of the Consortium for Small-Scale Modeling numerical weather prediction model COSMO with the maximum height (echo top or ET) at which a radar reflectivity of at least 45 dBZ for POH (50 dBZ for MESHS) is detected (Donaldson, 1961). Both algorithms are described in detail in sections 3.1 and 3.2 in Nisi et al. (2016). MESHS differs from the maximum estimated size of hail (MESHS; Witt et al., 1998), a radar-based hail product that is commonly used in North America and that integrates the reflectivity greater than 40 dBZ above the melting layer. The ET information stems from the Swiss radar network which consists of five dual-polarization Doppler C-band radars. The radars scan the atmosphere at 20 elevations from -0.2°to 40°every 5 min (Germann et al., 2015, 2016). POH and MESHS 2D mosaic fields are available in real time every 5 min on a 1 x 1 km^2 Cartesian grid covering Switzerland and surrounding areas.

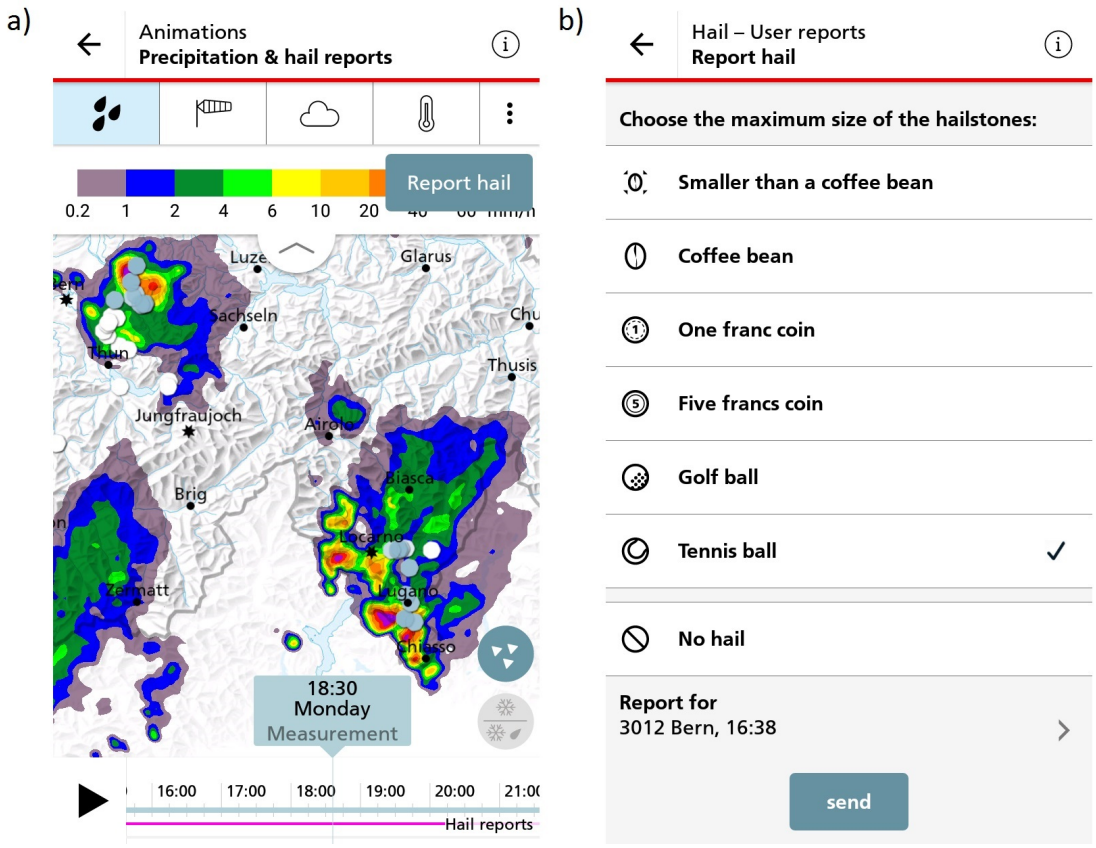

Figure 2.1: (a) Screenshot of the animation with radar-based precipitation rates (mm h^{-1}, colors) and the crowdsourced reports (blue and white dots) on 7 May 2018. The blue dots indicate hail reports for the shown time, white dots indicate past reports. (b) Screenshot of the hail size category scheme in spring 2018.

2.4.2 Crowdsourced data

The hail reporting function is part of the app of MeteoSwiss. It is included in the page that shows the radar precipitation fields in animated form, which is one of the most popular pages of the app (Fig. 2.1a). After passing a simple plausibility check, the hail reports are displayed seconds after they are submitted, overlaid on the radar echoes, and can be animated in time over the past 24 h. Users who observe hail can submit information on the time, location, and size of the hailstones. When a user submits a report, the current time and location of the phone are suggested as default input values, but both parameters can be adapted manually. The location information stems from position tracking by the smartphone. A manual adaptation of the location name and/or ZIP code will reduce the spatial accuracy by several hundred meters (depending on the size of the ZIP code area). The user can manually adapt the time by choosing the minute of the event. Knowing if the location and/or time were reported manually is important for filtering the reports. The user then chooses a size from a predefined hailstone size category scheme (Fig. 2.1b). Between May 2015 and September 2017, users could choose between the size

Table 2.1: Original and current crowdsourcing hail report size category scheme, the corresponding approximate diameters, and range of diameters they cover.

Size category	Diameter (mm)	Diameter range (mm)
Original		
Coffee bean	5-8	>0–15
One Swiss Franc coin	23	15-27
Five Swiss Franc coin	32	27-32
Larger than five Swiss francs coin	>32	>32
Current		
Smaller than a coffee bean	>0-5	>0-5
Coffee bean	5-8	5-15
One Swiss Franc coin	23	15-27
Five Swiss Franc coin	32	27-37
Golf ball	43	37-55
Tennis ball	68	>55

categories "no hail," "coffee bean," "1 Swiss Franc coin (CHF)," "5 CHF," and ">5 CHF" (see Table 2.1 for the corresponding diameters in millimeters). This original size category scheme was updated in September 2017 to include a "smaller than coffee bean" category, and the ">5 CHF" size was replaced with two categories, "golf ball" and "tennis ball" (see Table 2.1). The "smaller than a coffee bean" category was added to differentiate between graupel (<5 mm) and hail (≥5 mm). The other two categories extend the range of categories to one that replaces ">5 CHF" and another larger size that mainly serves to catch suspicious reports. In spring 2018, an instruction was added requesting the users to report the largest hailstone size that they see. In addition to the location, the event time [time indicated by the user; in CEST (UTC + 2 h) in summer and CET (UTC + 1 h) in winter], and the hailstone size, the app stores the submission time (time at which the user presses "send"; Fig. 2.1b) and an anonymous user ID.
Note that users can also report "no hail." The "no hail" reports provide valuable information in close proximity of a thunderstorm to delineate hail from no-hail areas. However, we do not include the "no hail" reports in this statistical analysis because we cannot use it to count false alarms, since we cannot dismiss the possibility that hail did occur within the radar grid box (1 x 1 km^2) and 5-min time step corresponding to a "no hail" report. To simplify reading the article hereafter, we will refer to the reported categories in terms of hailstone diameter (Table 2.1). Note that each category spans a wider range of diameters that varies according to the chosen category scheme.

2.5 Successful data acquisition

From 1 May 2015 to 31 October 2018, 59,020 MeteoSwiss crowdsourced hail reports were submitted by 39,733 different user IDs on 1,203 days over an area of 12,375 km^2 (with at least one report per square kilometer), which corresponds to a quarter of the Swiss territory. The dataset

has 17,739 reports in the "no hail" category and 41,281 reports that indicate the presence of hail. More than 10 reports were submitted each day on 718 days, and more than 100 reports were submitted each day on 140 days. These are impressive numbers when compared to the small size and population of Switzerland. Crowdsourcing hail with the MeteoSwiss mobile app has been successful for several reasons. First, hail is a rare natural phenomenon that fascinates many people, is easy to recognize, and often interrupts people's activities. Second, the crowdsourcing function is embedded in the radar animation of the MeteoSwiss weather app. This app is widely used, with an average of about 500,000 active users per day (of a population of approximately 8 million). Third, the MeteoSwiss mobile app has been downloaded more than 8 million times and is therefore the most popular weather app in Switzerland. The high number of users provides an unprecedented spatial and temporal observational coverage that could not be acquired differently given today's observational methods, knowledge, and monetary restrictions. Last, MeteoSwiss deliberately publishes blog posts about the reporting function in the spring, at the beginning of the hail season, to encourage usage. The blog is part of the app and is popular.

An example of the crowdsourced data submitted on 31 May 2017 in the region of Thun is shown in Fig. 2.2. The 86 reports were mainly submitted from densely populated areas. Almost all of the 86 reports are located inside the POH >80% area. There are some reports of the largest size category collocated with MESHS values >60 mm. On this day, several reports were submitted within the same 1-km^2 grid box. From one grid box (see coordinate 614.5/178.5 in Fig. 2.2), 17 reports were submitted within 26 min; the maximum number of reports per individual grid box ever recorded within 1 h. There is quite some variability in the reported hailstone sizes among the 17 reports, which points to subgrid-scale variability in the hailstone size, as well as to the uncertainty of the size estimates submitted by the app users.

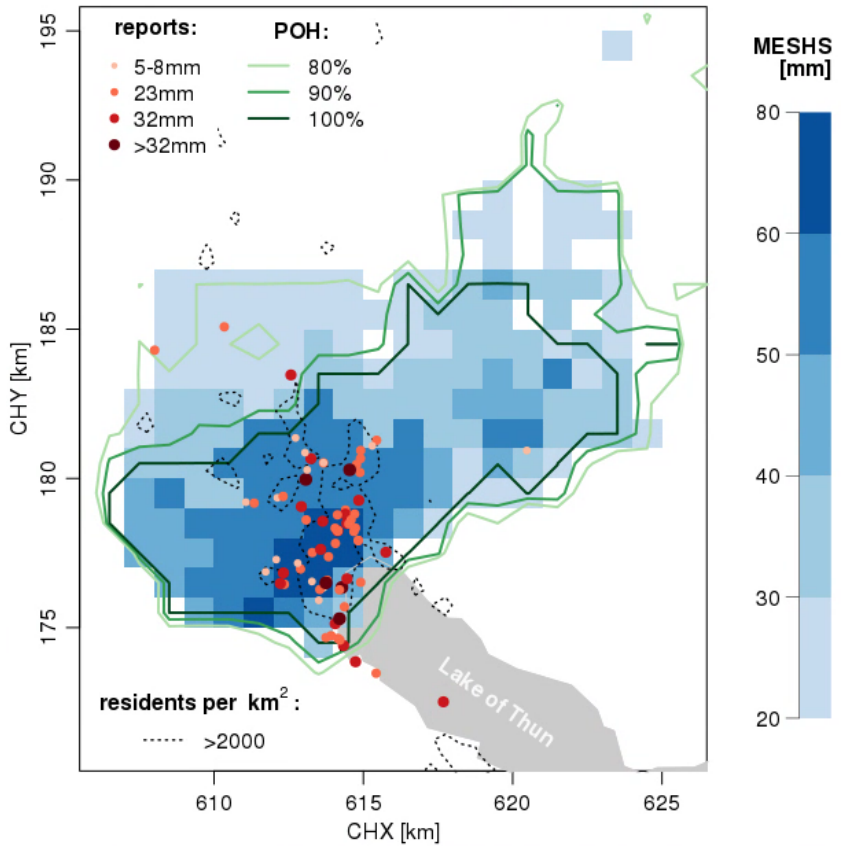

Figure 2.2: Maximum values of MESHS (blue grid boxes, mm), POH (green contours, %), and the crowdsourced hail reports (red dots, see also Table 2.1) for 31 May 2017 in the region of Thun, Switzerland. The dashed dark gray lines contour the areas with more than 2,000 residents per 1 km^2 in the year 2017 (Federal Statistical Office of Switzerland, 2017). The light gray area shows the Lake of Thun. The 32-mm report on the lake was probably submitted from a boat. The axes indicate the Swiss Coordinate System (km, tick marks every 5 km). The image is centered at approximately 46.802° N, 7.655° E.

2.6 Crowdsourced data acquisition using a government app versus a custom app

While the wide distribution of the MeteoSwiss app is a huge advantage for the dissemination of the app and hence the number of reports, working with the government weather app has implications for the hail reporting options. The app is one main warning channel for the Swiss authorities and, therefore, the stability of the app has precedence over the reporting function. Every additional function imperils the stability and has to meet strict requirements. Hence, working with a custom app (e.g., mPING, EWOB) dedicated solely to gathering information on hail (or thunderstorms) would have the advantage of a substantial extension of the reporting options. The MeteoSwiss app is continuously being updated and improved, and one next step

will be to provide an official Internet page informing on the hail reporting function. Suggestions for expanding the hail reporting function include the option of submitting photos and reporting the hail cover thickness, hail shape, hail size distributions, hail density, hailstone temperatures, event duration, or the damage caused. Such information would be very valuable; for example, Brimelow and Taylor (2017) verified the MESH algorithm with hail sizes estimated from photos posted in social media. In addition, quality control measures could be included in a custom app, such as the option to submit an email address to contact people who submit reports for later verification.

2.7 Quality control of the crowdsourced reports

2.7.1 Plausibility filters

The crowdsourced reports are influenced by human perception and sense of humor. This is why the crowdsourced data need to be quality controlled. Particularly for the comparison to the radar algorithms, erroneous reports need to be removed. We apply a multistep procedure to the 41,281 MeteoSwiss crowdsourced hail reports (excluding "no hail" reports) that is applicable in real time. First, we only keep the reports within an area that includes Switzerland and approximates the area that is well covered by the Swiss radar network (between 45.5°N, 5.6°E and 47.9°N, 10.7°E). This removes 479 (1%) reports. Second, any duplicate of the same anonymous ID, time (rounded to 5 min), coordinate (rounded to 1 km), and size is removed, in case the same user repeats the same report within a few seconds. This criterion accounts for 724 (2%) reports.

We then apply a time filter and discard reports with more than 30 min difference between the submission time and event time. The reasoning behind this filter being that when people report hail hours after the event happened, they might not remember the size of the hailstones and/or the time of the hail event very accurately. This is also one of the reasons why the app suggests a size category scheme rather than allowing people to directly estimate the size in centimeters. This removes 3,195 (8%) reports.
Next, reports that are implausible due to the meteorological conditions are removed. This reflectivity filter requires a minimum radar reflectivity of 35 dBZ, that is, a convective cell, to be located in the neighborhood of the report. The neighborhood method follows the so-called single observation neighborhood forecast verification (Ebert, 2008) and filters the reports as follows: we consider all radar grid boxes, for all time steps between 15 min before and 15 min after the reported time, whose centers are within a radius of 4 km from the exact report location. In most cases, this temporal space includes six time steps. Fig. 2.3) shows an example for two reports, using a radius of 2 km and for two time steps. Depending on the location of a report within its grid box, the spatial radius will include a different number of neighborhood grid boxes. The neighborhood accounts for the up to 2–4-km wind drift of hailstones (Schiesser, 1990; Schmid et al., 1992; Hohl et al., 2002) and for a margin of error in the reporting time. This filter is based on radar information, as are the POH and MESHS products, and hence not fully independent. Unfortunately, there is no fully independent validation information available in Switzerland.

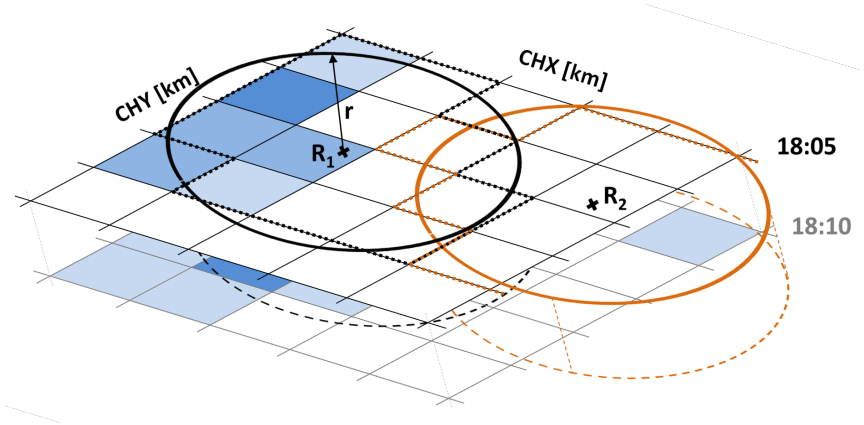

Figure 2.3: Example radar grids at two time steps with two crowdsourced reports (R1, R2). The blue grid boxes indicate nonzero radar values. The circles delimit the areas that are within a radius of 2 km around each report and the dotted lines show which grid boxes are within the neighborhoods defined by the circles.

However, the 35-dBZ threshold is smaller than the thresholds used to define POH and MESHS. This filter removes 16,892 reports, that is, 41% of the reports.
Next, reports by individual users with an unusual reporting pattern are removed. This includes reports from users with at least three reports of at least three different sizes, including the largest size category, within an hour. Furthermore, we filter reports if a user submits more than three reports on the same day and chooses a different, manually adapted location for each report. The last filter removes reports in which the same user submitted <5–8- or 5–8-mm reports and the largest size category within 2 min. These filters remove 327 (1%) reports.

Note that the quality control is not based on the number of reports for an individual event. Indeed, there are many cases in which single reports from lightly populated, remote areas are plausible and even confirmed by independent reports. There are also several cases in which the reflectivity filter identifies implausible reports clustered in a populated area.

Table 2.2: Number of matches between the filtered reports and POH and MESHS for each reported category, considering the radar gridbox value containing the report (A) and a 2-km and 5-min neighborhood window around the report (B). Numbers in rows 2–5 are absolute numbers (percentage of filtered reports).

	5-8mm	**23mm**	**32mm**	**>32mm**	**Total**
Number of filtered reports	12,136	3,171	653	610	16,570
Matches with POH (A)	4,506 (37%)	1,598 (50%)	263 (40%)	60 (10%)	6,427 (38%)
Matches with POH (B)	6,593 (54%)	1,971 (62%)	317 (49%)	101 (17%)	8,982 (54%)
Matches with MESHS (A)	1,139 (9%)	719 (23%)	157 (24%)	29 (5%)	2,044 (12%)
Matches with MESHS (B)	2,717 (22%)	1,306 (41%)	231 (35%)	49 (8%)	4,303 (26%)

The effect of the filters on the number of reports for each size category is shown in Fig. 2.4.

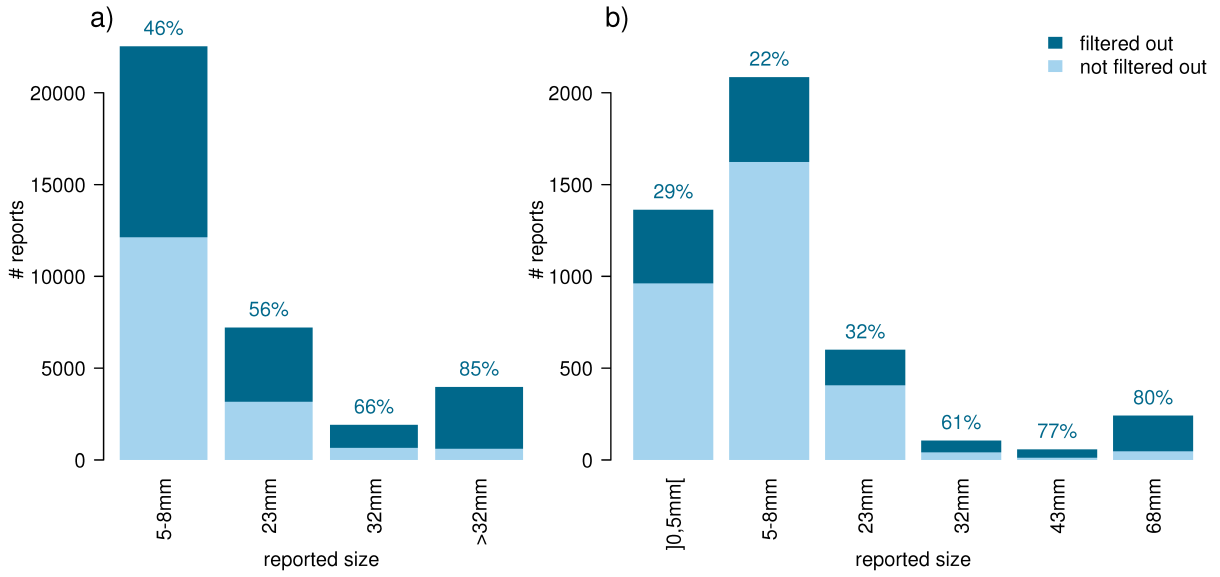

Figure 2.4: Number of reports per size category with (a) the original size category scheme that were valid from May 2015 to Aug 2017 and (b) the current size category scheme (valid since Sep 2017). "No hail" reports are excluded. The fraction of reports that were filtered out are given as percentages above the bars and indicated by the dark blue color. Note the different y-axis ranges.

Considering both size category schemes, 19,664 or 48% of all reports remain after filtering. For the original size category scheme (Fig. 2.4a), 16,570 or 40% of the reports remain. We refer to the remaining reports collected with the original size category scheme as the filtered reports. Of the 16,570 filtered reports, 12,136 (73%) are 5–8-mm, 3,171 (19%) are 23-mm, 653 (4%) are 32-mm, and 610 (4%) are >32-mm reports (see Table 2.2, top row, and Fig. 2.4a). The filters mainly reduce the number of >32-mm reports. This is expected, as the largest size category (until September 2017, >32 mm) might be chosen as a joke. Figure 4b shows the filter effects for reports submitted using the new size category scheme. Since the sample size is small and because of the change in category scheme, we do not compare the two histograms any further. The effects of altering hail reporting thresholds are discussed by Allen and Tippett (2015).

The large fraction of filtered reports for the large size categories (43 and 68 mm) suggests that the filters are efficient, particularly since these reports were mostly submitted during the winter half year, when such large hailstones are almost impossible. However, more tennis ball (68 mm) reports than golf ball (43 mm) reports remain in the sample after filtering, which indicates that the filters do not remove all untrustworthy reports.

Almost 81% (13,420) of the filtered reports were submitted on 100 hail days. The number of reports is greatest in the late afternoon and evening (Fig. 2.5), which reflects the typical thunderstorm diurnal cycle (e.g., Mandapaka et al., 2013; Nisi et al., 2018). The radar-based hail climatology (Nisi et al., 2016) indicates a second hail maximum at night that likely develops through down-valley winds and thunderstorm outflows converging in moist and unstable pre-Alpine air masses. This second maximum is not visible in the number of reports, most probably

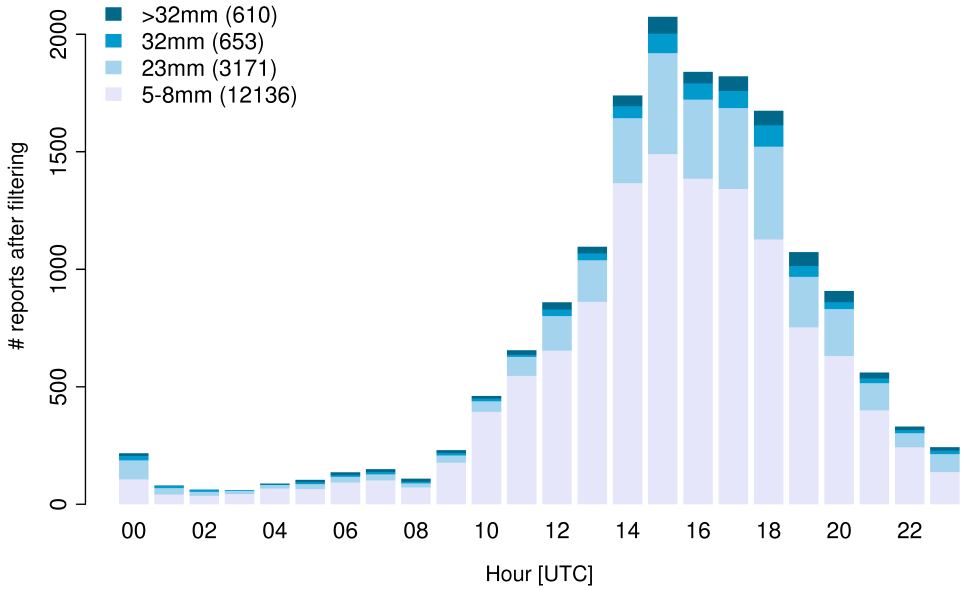

Figure 2.5: Number of crowdsourced hail reports per hour of the day and per size category after filtering. The legend also shows the number of reports per size category.

due to the general population being indoors. The spatial report density (not shown) primarily reflects the population density rather than the spatial hail frequency (see also Fig. 2.2) and therefore reflects the locations at which people and/or assets may be exposed, which is an advantage for hail risk studies.

2.7.2 Comparison of the MeteoSwiss crowdsourced hail reports with independent hail information

The network of hail sensors under construction already captured five events with graupel or very small hailstones and three events with maximum hail diameters >20 mm. During three of the five graupel/very small hail events, 1, 9, and 16 coffee bean reports (no reports of larger sizes) were recorded within a 2-km radius around the hail sensors. The MeteoSwiss crowdsourced hail reports submitted during the three hail events with hail diameters >20 mm, within 2 km of the sensors, were mostly equal to or larger than the diameters measured by automatic hail sensors. More events are needed to make a quantitative comparison.

Between May 2015 and July 2018, 110 filtered MeteoSwiss crowdsourced reports could be matched with 25 ESWD reports. For 21 cases, we found at least one MeteoSwiss report of the same size as the ESWD reports within the time uncertainty given by ESWD and within 2 km of the ESWD report. The MeteoSwiss crowdsourced reports for the remaining four cases indicated smaller hail diameters than the ESWD reports. In two cases, MeteoSwiss crowdsourced reports suggested larger hail sizes than indicated by ESWD. While the number of compared reports is too small for a conclusive statement, the results point toward the filtered MeteoSwiss crowdsourced reports being in good agreement with independent crowdsourced data.

2.8 Comparison with radar-based hail algorithms

2.8.1 Matching the reports to POH and MESHS

We match the filtered reports to nonzero POH and MESHS values. We use the 16,570 filtered reports received with the original size category scheme, as they constitute 84% of the sample. Again, we use a space and time neighborhood to match the reports with radar fields. Aside from the horizontal drift of hailstones, arguments can be made to allow for a margin of error in reporting time. Users might need to move themselves, a car, or flowers to safety before they report the hail fall, or they might not remember the time of the hail event. In addition, hail remains on the ground for some time before it melts and users might report hail on the ground rather than the hail fall. It therefore might be important to consider a spatial and temporal neighborhood to match the reports to the radar fields. To illustrate the sensitivity of the match to the chosen spatial and temporal neighborhoods, two methods are used to match the reports to the radar-based fields.

Method A assumes no spatial drift, an accurate reporting time, and uses the POH and MESHS values of the grid box and the 5-min time step closest to the reporting location and time. Method B uses the maximum POH or MESHS value within a spatial neighborhood radius of 2 km and a temporal neighborhood of 5 min centered around the exact reporting location and time. This neighborhood method is identical to the method applied in the reflectivity filter, but with a different neighborhood size (Fig. 2.3). Note that with the neighborhood method several reports might be matched with the same radar value. Of all filtered reports (including both size category schemes and excluding "no hail" reports), 86% (16,815 out of 19,664) are single reports within the respective grid box and 5-min time step. In 61% of cases with more than one report within the same grid box and 5-min time step, one unique size category was reported. Most of the cases (72%) where at least two sizes were reported within the same grid box and 5-min time step are combinations of <5–8-, 5–8-, and/or 23-mm reports. Repeating the analysis with only the maximum reported sizes does not significantly alter the results, which is why we conducted the analysis with all reports and not just the maxima. Table 2.2 shows the number of matches with POH and MESHS per size category for both methods. As expected, method B produces more matches than method A. The reports matched with method B but not with method A include cases in which hail drifted.

A sensitivity study that considered neighborhoods ranging between 2 and 6 km and between 5 and 30 min revealed little sensitivity of the results. The largest changes in the results occur when going from no neighborhood (method A) to a small neighborhood (method B; Table 2.2). Compared to the number of additional radar grid boxes that are considered with a larger neighborhood size, the increase in fraction of matches is relatively small (Table 2.3).

Only 9% of the filtered 5–8-mm reports are matched with MESHS using method A (Table 2.2). Since MESHS includes only hailstones ≥2 cm, a very low number of MESHS matches is expected for the smallest size class. POH estimates the probability of hail for hailstones of all sizes. Using

Table 2.3: Fraction of matches between the filtered reports and POH and MESHS considering different neighborhood sizes and the median number of grid boxes per neighborhood.

Neighborhood radii	**2 km and 5 min (method B)**	**4 km and 5 min**	**4 km and 15 min**
Matches with POH	54%	60%	67%
Matches with MESHS	26%	33%	41%
Median No. of grid boxes within the neighborhood	26	100	300

method A, 37% of the 5–8-mm reports are matched with a POH signal, and using method B, 54% are matched. For the 23-mm category, 23% of the reports are matched with MESHS using method A, and 24% are matched for the larger 32-mm class (41% and 35%, respectively, with method B). Interestingly, the fraction of matched reports decreases substantially for the largest size class (5%, method A and MESHS). This fraction is also very low for POH and when using method B, which suggests that there is still a significant number of reports in this category that are likely "joke" reports.

Between 46% (method B) and 72% (method A) of the filtered reports cannot be matched with POH larger than zero (74% and 88% for MESHS). There are several possible explanations for this. First, the neighborhood used to match the reports (i.e., 2 km and 5 min) is much more restrictive than the one used to filter the reports (i.e., 4 km and 15 min). However, increasing the matching neighborhood size does not greatly increase the number of matched reports (see Table 2.3). If the spatial neighborhood radius was doubled to 4 km, which quadruples the number of considered grid boxes, the total fraction of matched reports increases from 54% to 60% for POH and from 26% to 33% for MESHS. If the temporal neighborhood radius is additionally increased from 5 to 15 min, the fractions further increase to 67% (POH) and 41% (MESHS), which still leaves 33% (POH) and 59% (MESHS) unmatched filtered reports.

Second, recall that POH and MESHS are defined using reflectivity thresholds (45 dBZ for POH and 50 dBZ for MESHS). There is therefore a 10-dBZ difference between the minimum reflectivity of the filter (35 dBZ) and the required reflectivity for a POH signal. For 43% of the filtered reports that were not matched with POH using method B (39% for MESHS), the maximum reflectivity in the neighborhood was below the 45-dBZ threshold (50 dBZ for MESHS; not shown). It is therefore likely that hail (or graupel) can develop in Switzerland even if the radar reflectivity does not reach the threshold values of 45 or 50 dBZ. Third, the freezing-level height derived from the model influences the POH signal. The model may simulate a locally high freezing-level height stemming from the diabatic heating in a simulated thunderstorm cell. As a consequence, POH would be smaller or zero, since the distance between the freezing-level height and the maximum height with 45 dBZ would decrease. The same applies analogously to MESHS.

Fourth, the radar algorithms were fitted for convective thunderstorms happening during the summer season, and may miss events with graupel and/or small hail in the winter half year. The fraction of unmatched reports is much higher between October and April (88% for POH, 98% for MESHS with method B) than between May and September (38%, 71%). Another reason for the

large fraction of unmatched reports in the winter half year may be that users mistakenly report sleet. Finally, despite the filters, we likely still have an unknown number of erroneous reports in our sample (see Fig. 2.6).

2.8.2 Evaluation of the MeteoSwiss crowdsourced hail reports

The POH values of the matched reports increase with increasing reported size (Fig. 2.6a). Note that POH is not intended to provide any hailstone size information. In an ideal setting, POH would be independent of the hail size. However, given how POH is defined, we expect POH to be higher for large hail sizes and lower for smaller hail sizes. Figure 4b suggests that the original 5–8-mm category includes reports of graupel. In the original scheme, "coffee bean" (5–8 mm) was the smallest available size category; the large fraction of "smaller than a coffee bean" reports in the current category scheme strongly suggests the presence of graupel in the original "coffee bean" category. This is consistent with the POH values for this category being significantly lower than the POH values for the larger hailstone size categories (Fig. 2.6a). Since the notches of the POH boxplots do not overlap when comparing the 5–8-, 23-, and 32-mm size categories, the increase in median POH with the reported size is significant. There is a significant difference between the medians of the method A and method B POH values when comparing the 5–8-, 23-, and 32-mm categories (nonoverlapping notches and Mann–Whitney U test with p value of 0.05, not shown). Only a small fraction of reports are matched with small POH values. Using method A, only a quarter of the matched POH values are below 70% for the 23- and 32-mm reports. Using method B, only a quarter of the values are below 80%. Last, more than 50% of the matched POH values for the 23- and 32-mm reports and for method B have values >98%. The large interquartile and notch range of the POH values matched with >32-mm reports reflect the much smaller sample size and might indicate that this sample potentially still contains some incorrect reports despite the filtering.

MESHS values increase with increasing reported size (see Fig. 2.6b). This increase in the median is significant (Mann–Whitney U test, p value of 0.05) in all cases except when comparing the medians of 23 and 32 mm with >32 mm (for both methods A and B). This increase in median values (except for >32 mm) shows that the MESHS correctly recognizes the relative maximum expected size of hail above 2 cm. The interquartile ranges (IQRs) of MESHS span 1.5–2 cm. They approximate the size range that would be assigned to the reporting categories using the nearest neighbor (see Table 2.1). The constant IQRs suggest that the variance in MESHS is constant throughout the reported sizes.

When considering the 23- and 32-mm reports, MESHS is roughly 10–15 mm larger than the reported size, depending on whether method A or B is used for matching. The >32 mm and method B matching boxplot has lower quartile values than the boxplot of the MESHS values matched with 32-mm reports. As previously discussed, we assume that the matched sample likely still contains reports in which users exaggerated the reported size. However, the lack of additional fully independent data prohibits a definitive statement if the users systematically overestimate the hail size of the largest reporting category or MESHS systematically underestimates

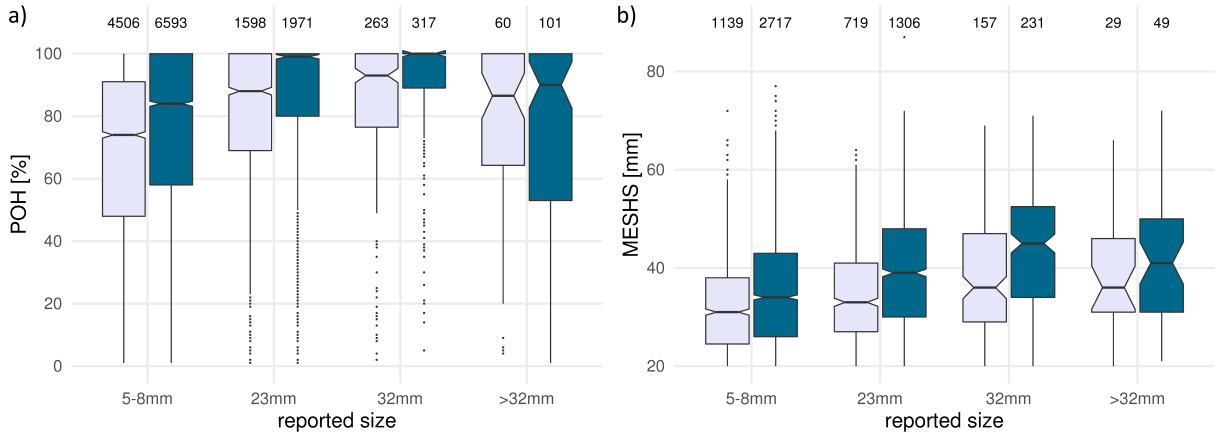

Figure 2.6: Boxplots of (a) POH vs reported size and (b) MESHS vs reported size considering no neighborhood (light) and a neighborhood radius of 2 km and 5 min (dark). The numbers at the top of the plots indicate the number of values (matches) that contribute to each boxplot. The boxplots are Tukey-style whiskers with notches showing the 95% confidence interval for the median m, given by m ±1.58 x IQR/$\sqrt{n}$ (McGill et al., 1978). IQR is the interquartile range and n is the sample size (see also Krzywinski and Altman, 2014).

the size. Since the sample of matched reports for >32 mm is very small in comparison with the other reporting categories, the incorrect reports have a larger influence. We therefore expect the quartiles to be larger once the sample size has reached several hundred reports. Once more 32-, 43-, and 68-mm reports are gathered, the IQRs for 32 mm and these larger categories can be meaningfully compared.

2.9 Summary and Conclusions

The crowdsourced hail reports gathered with the MeteoSwiss app constitute an extremely valuable observational dataset on the presence and approximate size of hail in Switzerland. This dataset has the advantage of unprecedented spatial and temporal coverage, and the automatic real-time processing and visualization is very convenient for nowcasting applications. Beside the scientific value of the dataset, we hope that the crowdsourcing function serves as a bridge between the general population and the world of research. This requires feedback from the scientists to the app users, which is currently provided through blog posts linked to the app and in newspaper articles. It will be extended in the future to include information on a dedicated website.

The reported hailstone sizes indicate that hail with a size close to the size of coffee beans is most abundant (note that this size category likely contains also reports of graupel). The number of reports follows the typical diurnal cycle of thunderstorm activity, with most reports being submitted in the early evening and evening. The spatial distribution of the reports primarily reflects the population density.

While the crowdsourced dataset dramatically increases the number of hail observations, they need

to be quality controlled. Our reflectivity filter requires reports to be close to a radar reflectivity area of at least 35 dBZ. Overall, the plausibility filters remove approximately half of the reports in the dataset.

Our analyses suggest that except in the largest size category, enough false reports are filtered out for them to not substantially influence statistical analyses. The dense spatial and temporal coverage of the filtered reports allowed us to carry out a systematic comparison to the two operational, single-polarization radar-based hail algorithms, probability of hail (POH) and maximum expected severe hail size (MESHS). The fraction of unmatched reports between May and September (38% for POH and 71% for MESHS; using method B) suggest that POH and MESHS are too restrictive in identifying hail areas. Of these unmatched reports, 43% (39% for MESHS) were submitted in an area with a maximum reflectivity between 35 and 45 (or 50 for MESHS) dBZ. Using a lower reflectivity threshold in the algorithms may therefore improve their quality. However, adapting the radar-based algorithms should entail a quantification of the false alarm rate, which cannot be achieved with the crowdsourced reports alone.

The positive correlation between reported sizes and the values of POH and MESHS suggest that the filters adequately separate plausible reports from improbable reports, except for the largest hail size category. Furthermore, the comparison of MESHS with the reported size shows that MESHS can be used as an estimate of the maximum size of hail >2 cm in terms of relative comparisons. Absolute MESHS values matched with the 23- and 32-mm categories exceed the reported hailstone size on average by 1.5 cm when a spatial neighborhood is considered to match the crowdsourced reports with MESHS values (method B). This difference merits further investigation using data from the hail sensor network. If the measurement campaign with the 80 new automatic hail sensors is successful, we will be able to test this conclusion and further improve the hail algorithms.

2.10 Acknowledgments

The implementation of the crowdsourcing function in the app was enthusiastically supported by Markus Aebischer and Bertrand Calpini (MeteoSwiss) and was funded by the Mobiliar Lab for Natural Risks of the University of Bern thanks to the help of Matthias Künzler. Pascal-Andreas Noti and Andrey Martynov carried out a pilot study of the MeteoSwiss crowdsourced hail reports and the comparison to the radar algorithms in the framework of Pascal Noti's master's thesis (http://occrdata.unibe.ch/students/theses/msc/192.pdf). We also thank Mattia Brughelli and Veronika Roethlisberger for their advice on the population data. Last but not least, we thank the three reviewers for their encouraging comments and their great suggestions for improving our article.

Chapter 3

Update on the MeteoSwiss crowdsourced hail reports until September 2020

Motivated by two additional summers since the analyses presented in Barras et al. (2019), this short chapter gives an update on the characteristics of the MeteoSwiss crowdsourced hail reports until September 2020. By then, a total of 119'549 reports were registered. The years 2015–2017 each had > 17'000 reports. This number dropped to a few thousands in 2018 and 2019 (Fig. 3.1, see also Table 3.1). This drop is most likely due to an update in the MeteoSwiss app in September 2017. In a successful attempt to reduce the fraction of false reports, the hail reporting function was concealed by an additional button. While less reports were registered, the fraction of filtered reports increased to > 70 % in 2018 and 2019 (Table 3.1; columns "after filtering") and the total number of filtered reports did, therefore, not decrease as drastically as the total number of reports.

The opposite idea of not concealing, but revealing the reporting function is being tested since July 2020. Since the recent update the "report hail" button and previous reports done in the previous 24 hours are immediately visible when opening the subpage leading to the hail reporting function (equivalent to (Fig. 2.1a)). The number of reports in 2020 exploded to 54'084, of which almost 31'000 were sent only in July (Fig. 3.1). Within almost only three months, the previous number of filtered reports increased by 70 % to a total of 41'191 filtered reports.

The right columns in Table 3.1 present the contribution of each of the three main filter criteria to filtering the reports each year. The fraction of reports filtered out by the reflectivity filter (>= 35 dBZ, see Barras et al., 2019) has decreased considerably in 2015–2019, to only a third of the initial value. The reason why much less reports were filtered out in 2018 and 2019 is probably the concealment of the reporting function. Since a large fraction of the general user community did not know about the reporting function, only people that were shown the reporting function or who found it by chance would send reports. The other two filter criteria limiting the temporal distance between event and submission time to less than 30 minutes ("< 30 min") and detecting strange reporting behaviours ("blacklist") do not contribute as much to filtering the reports. The

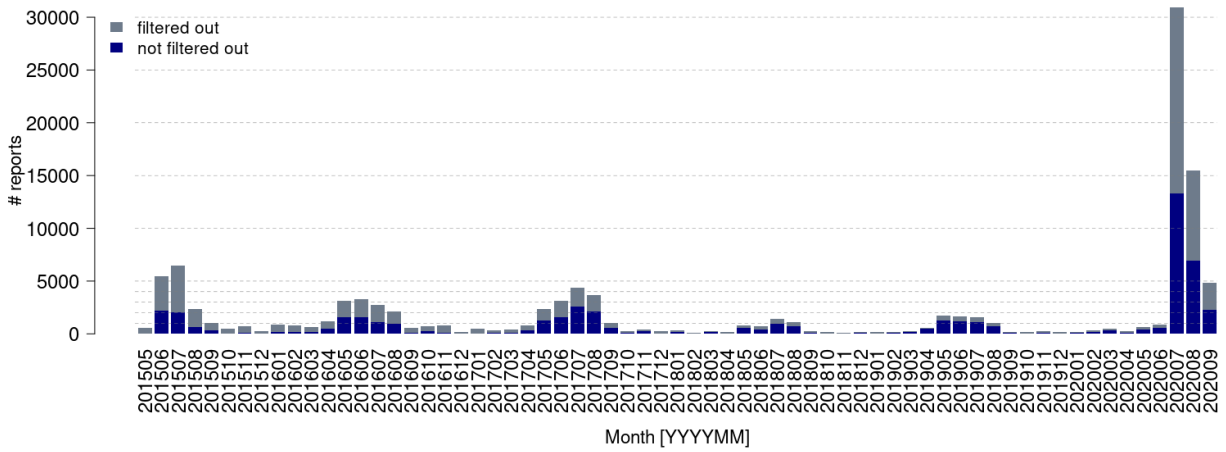

Figure 3.1: Number of MeteoSwiss crowdsourced hail reports per month before and after applying the filtering criteria, between May 2015 and September 2020.

fraction of reports filtered out by these criteria have decreased from 2015 to 2020. The numbers in the brackets in the right columns in Table 3.1 indicate which fraction of reports would not have been filtered out if not for that specific filter criterion. For the $<$ 30 min criterion, the reports counted in the fraction within the brackets have a bigger likelihood of being correct after all, since the reflectivity filter did not filter them out. For the blacklist criterion, the numbers in the brackets show which reports with strange reporting patterns were likely made during a hail event. The differences between the number outside and within the bracket for the latter two filtering criteria in Table 3.1 suggest that in the initial years, users were more likely to trigger these filters during weather conditions that did not produce hail.

Table 3.1: Total number of MeteoSwiss crowdsourced hail reports, excluding the size "no hail"; the total number as well as the fraction of these reports remaining after filtering each year; the fraction of "total" that each filter criterion would filter out if applied independently of the other filter criteria. The numbers in brackets indicate the fraction of reports in which the given filter criterion is solely responsible for filtering out a report.

	total	**after filtering**		**filter criteria [%]**		
year	**#**	**#**	**% of total**	**>35dBZ**	**<30min**	**blacklist**
2015	10754	3692	34	61 (54)	9 (3)	4 (1.2)
2016	12030	5421	45	50 (42)	9 (4)	4 (0.7)
2017	13245	7705	58	37 (32)	7 (4)	2 (0.7)
2018	4236	2950	70	25 (22)	6 (4)	2 (0.6)
2019	6163	4521	73	22 (19)	5 (4)	2 (0.8)
2020	31729	16902	53	44 (42)	4 (3)	1 (0.4)

The remaining paragraphs of this chapter repeat the matching and comparison of crowdsourced hail reports with POH and MESHS using the two neighbourhood methods A and B presented in section 2.8.1. Method A matches the reports with the radar grid box closest to the report and Method B considers the maximum radar value within 2 km and within 5 minutes of the report (more details in Barras et al., 2019). The number of matches between the radar-based hail algorithms and the reports from the current size category scheme is shown in Table 3.2 (equivalent to Table 2.2). Compared to Table 2.2, a 4 % larger fraction of reports is matched with POH and a 1 % smaller fraction of reports is matched with MESHS (both matching methods). However, the fraction of <5–8 mm reports matched with MESHS is only 2 % (A) and 9 % (B) (Table 3.2). If <5–8 mm reports are ignored in the calculation, then the total fractions of reports matched with MESHS are 14 % (A) and 31 % (B), 2 % (A) and 5 % (B) larger than the fraction of reports matched with MESHS in Table 2.2.

Table 3.2: Number of matches between the filtered reports from the current category scheme and POH and MESHS for each reported category, considering the radar gridbox value containing the report (A) and a 2-km and 5-min neighborhood window around the report (B). Numbers in rows 2–8 are absolute numbers (percentage of filtered reports).

	<5–8 mm	5–8 mm	23 mm	32 mm	43 mm	68 mm	Total
# filtered reports	7285	11728	3699	805	307	762	24586
POH (A)	1793 (25%)	5882 (50%)	2104 (57%)	397 (49%)	109 (36%)	29 (4%)	10314 (42%)
POH (B)	3016 (41%)	8049 (69%)	2539 (69%)	466 (58%)	135 (44%)	68 (9%)	14273 (58%)
MESHS (A)	171 (2%)	1267 (11%)	895 (24%)	231 (29%)	70 (23%)	10 (1%)	2644 (11%)
MESHS (B)	671 (9%)	3288 (28%)	1632 (44%)	359 (45%)	102 (33%)	17 (2%)	6069 (25%)

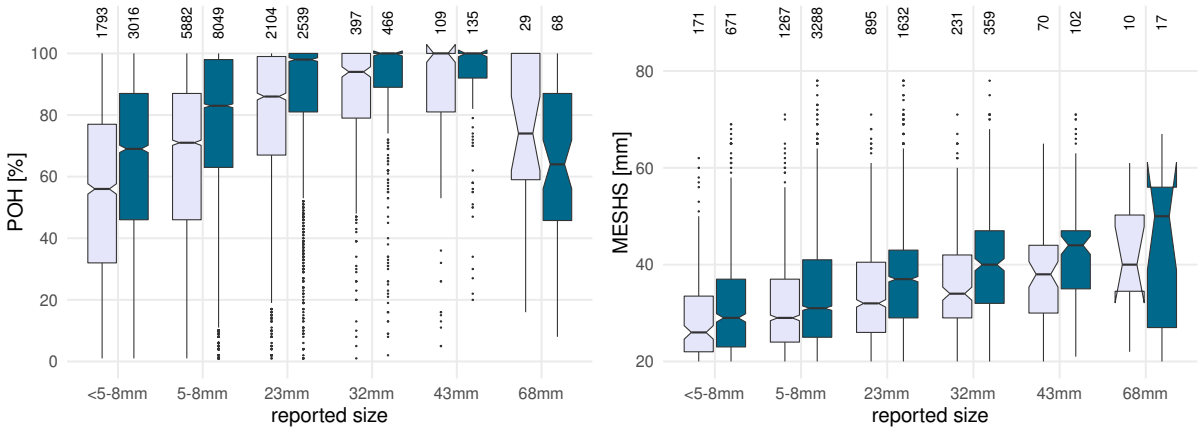

Figure 3.2: Boxplots of (a) POH vs reported size and (b) MESHS vs reported size considering no neighborhood (light, in text called Method A) and a neighborhood radius of 2 km and 5 min (dark, in text called Method B) with the reports from the current size category scheme. The numbers at the top of the plots indicate the number of values (matches) that contribute to each boxplot. See Fig. 2.6 for further details.

Similarly as in Fig. 2.6, matched POH and MESHS values in Figure 3.2 increase with an increasing reported size and results mostly confirm the findings in Barras et al. (2019). The additional categories in the current scheme give more detail on the ranges of POH and MESHS associated with graupel (< 5–8 mm) vs. hail (5–8 mm and larger size categories). Furthermore, the additional categories for bigger hail as well as the larger number of reports increases the robustness of the results for the categories 23 mm, 32 mm and 43 mm.

Figure 3.2a suggests that at least 75 % of POH values associated with < 5–8 mm and 5–8 mm are smaller than 100 % and at least half of the POH values matched with < 5–8 mm are smaller than 60 % (Method A, Figure 3.2a). The category 68 mm has few reports and behaves similarly like the category > 32 mm in Fig. 2.6; the large interquartile and notch range reflect the smaller sample size and are a potential indicator for incorrect reports despite the filtering. The median of MESHS values matched with 32 mm are roughly 3–8 mm larger than the reported size (depending on the matching method). Median values matched with 43 mm are roughly 5 mm smaller (Method A) or equal (Method B) to the reported size.

Chapter 4

Nowcasting of hail with XGBoost

4.1 Abstract

In this chapter, extreme gradient boosted tree (XGBoost) models predict the maximum probability of hail (POH) and the maximum expected severe hail size (MESHS) in steps of 5 minutes, for lead-times of 5 to 45 minutes and for individual thunderstorms in Switzerland. Thunderstorm and storm environmental variables from the summer of 2018 are extracted along cell paths, up to 45 minutes before each thunderstorm tracked location. Data sources are the Swiss radars, the numerical weather prediction model COSMO-1, Meteosat satellites, lightning, topographical information and other meta data. For each past cell position, 12 statistics (mean, median, standard deviation, sum and 7 percentiles) of the predictor variables are calculated within circles of 23 km around the cell positions. The statistics serve as features (also called predictors) with which models predicting the maximum POH and MESHS are trained and tested. Tow types of XGBoost models are created for each lead-time and target variable. The binary XGBoost models predict the occurrence of hail (POH $\geq$ 10 %, MESHS $\geq$ 2 cm) and the linear XGBoost models predict the non-zero POH and MESHS values. This project focuses on determining the effect of hyper-parameter tuning and on the effect of reducing the number of features to the top 5 to 1000 most important features. Furthermore, the models are interpreted with the Shapley additive explanations (SHAP) method.

Binary XGBoost models successfully predict the occurrence of hail equally well as the Lagrangian persistence for lead times of 5 minutes. For larger lead-times, XGBoost models perform better. Out of >10'000 features, 500-1000 top features are necessary to reach the same model performance as models using all features. The top 100 features contain variables from all data sources. The most frequently used sources are radar data, followed by COSMO-1 and satellite data. The number of radar-based features in the list of top 100 features decreases with an increasing lead-time. For all lead-times, the majority of the top 100 features are statistics from the most recent observation (t_0 = t). Time steps between t-5' and t-45' are, however, used as well. The SHAP model interpretation method suggests that feature values indicating intense thunderstorm activity at t_0 increase the probability of hail occurring. This chapter shows the effectiveness of using machine learning to predict hail and may lead to an operational hail warning system in the

future.

4.2 Introduction

In Switzerland, the currently operational thunderstorm radar tracking algorithm (TRT, Hering et al., 2008; Rotach et al., 2009) detects thunderstorm cell by searching for local maxima in the radar reflectivity. Simultaneously, TRT ranks the thunderstorm severity using a fuzzy logic scheme. The future location of each thunderstorm cell is determined through forward extrapolating past cell movements and the severity is assumed constant (see Fig. 1.1). Using this information, automatic warnings are issued to the Swiss population (Panziera et al., 2016). While the TRT severity ranking provides a rough indication for the probability of hail, Switzerland does not have an operational, automatic nowcasting estimate for the future presence or size of hail up to 45 minutes in advance yet. Today's available data and statistical methods allow to improve the currently operational procedure.

Numerical weather prediction models, diverse observational systems, statistical methods as well as rising computational powers have increased the skill in nowcasting hail in the past decades. Initially, hail was not predicted directly, nowcasts predicted larger-scale convective phenomena. Browning et al. (1982) and Wilson et al. (1998) summarizes nowcasting techniques used between the 1960s and 90s, such as extrapolations of radar data, knowledge based expert systems and numerical forecasting models that are initialized with radar data. In the 2000s a multitude of nowcasting models were developed, some of which were tested in Forecast Demonstration Projects (FDP) e.g. during the Olympic games in Sydney in 2000 (Keenan et al., 2003; Ebert et al., 2004) and in Beijing in 2008 (Wilson et al., 2010). Nowcasting models included new data sources, such as lightning, satellite and surface data (e.g., Eilts et al., 1996; Pierce et al., 2000; Bonelli and Marcacci, 2008; Kober and Tafferner, 2009), tracked objects in radar data, such as convective storms (e.g., Dixon and Wiener, 1993; Hering et al., 2004; Mueller et al., 2000; Kober and Tafferner, 2009) and boundary layer convergence lines (Mueller et al., 2000) and combined characteristics of thunderstorm cells recognized in radars with a life cycle model (Pierce et al., 2000). Blending techniques combined radar and numerical weather prediction (NWP) models (e.g., Golding, 1998; Liang et al., 2010; Li and Lai, 2004; Wong and Lai, 2006). In NWP models with convection-allowing resolutions, new components simulate convective hazards, such as the hail diagnostic HAILCAST(Brimelow et al., 2002) implemented in the Weather Research and Forecasting (WRF) model by Adams-Selin and Ziegler (2016) (WRF-HAILCAST). This hail diagnostic has been implemented in the Swiss NWP model COSMO-1 and its operational use will be tested in summer 2021.

Another promising nowcasting method is machine learning (ML). Instead of resolving the exact processes in a complex model, ML methods can diagnose linear and non-linear relationships between variables describing the storm and its environment (predictors) and the thunderstorm hazards (predicted variables). These relationships may help to understand better which condi-

tions and interactions of environmental processes are responsible for which size of hail at the ground, at any location and time. At MeteoSwiss, this idea lead to a project called Coalition-3, which had the aim of nowcasting storm severity using ML and which was extended to this part of this doctoral thesis. In this chapter, ML models are developed to predict the likelihood and maximum size of hail at lead-times from 5 to 45 minutes for any existing thunderstorm within the range of the Swiss radar network.

Previous research projects have demonstrated the potential for ML-based hail prediction. Marzban and Witt (2001) successfully used neural networks to predict the maximum size of severe hail per storm. Manzato (2013) applied an ensemble of neural networks to predict the occurrence and size of hail in northeast Italy. The US annual Hazardous Weather Testbed Spring Experiment tests every year the newest technology in an operational setting with forecasters (Gallo et al., 2017). The WRF-HAILCAST diagnostic (Adams-Selin and Ziegler, 2016) has been compared to the Thompson hail size diagnostic (Thompson et al., 2004, 2008 in Gagne et al., 2019), the Gagne Machine Learning Method (Gagne et al., 2017, 2018) and storm surrogate variables extracted from the WRF and Data Assimilation and Research Testbed (DART; Anderson et al., 2009) models, such as the updraft helicity (Sobash et al., 2016; see McGovern et al., 2017 for more details). Results showed that ML based models had higher skill to nowcast hail than the other methods (McGovern et al., 2017). Gagne et al. (2015) initially predicted the daily maximum hail size up to one day ahead using statistics of variables from ensembles of WRF models. Three ML methods were tested, random forests, a combination of logistic classification model and ridge regression and gradient boosting regression trees. Out of the three, the gradient boosted regression trees performed statistically significantly better for most of the model ensembles. Building on that model, Gagne et al. (2017) matched storms simulated in convection allowing models with radar-observed storms and then predicted the spatial distribution of hail sizes by synthesizing storm properties and pre-storm environmental variables. Hill et al. (2020) generated probabilistic predictions of severe weather for day 1–3, including hail for day 1, using random forests (RF; Breiman, 2001). Models were created for different regions in continental United States, using atmospheric fields from the NOAA Second Generation Global Ensemble Forecast System Reforecast (GEFS/R) dataset as predictors and Storm Prediction Center (SPC) storm reports as targets. A weighted blend of SPC and RF outlook was shown to have the highest predictive skill. Zhou et al. (2019) used convolutional neural networks (CNN; LeCun et al., 1990) to infer different thunderstorm hazards, including hail, in Global Forecast System (GFS) forecasts. Finally, Flora et al. (2020) predicted the occurrence of tornadoes, severe hail (hail diameter > 1 in (1 in=2.54 cm)) and severe winds with ML models, using the Warn-on-Forecast System (WoFS) ensembles of the WRF model. The ensembles predicted the occurrence of severe hail with a normalized critical success index (NCSI[1]) of 0.3 (0.2) for the first (second) hour (Flora et al., 2020). Not predicting hail, but also applying ML to nowcast short lead-times, Lagerquist et al. (2017) combined several ML techniques to nowcast damaging straight-line convective winds and Lagerquist et al. (2020) predicted the next-hour tornado occurrence probability using CNN.

[1] CSI normalized by the CSI of a no-skill system; see Flora et al. (2020)

ML applications in weather and climate sciences have been criticized in the past for generating black-box models that do not improve the understanding of the involved physical and dynamical processes. However, model interpretation methods are giving insight into the patterns discovered by ML models. McGovern et al. (2019b) have synthesized and analyzed multiple approaches to model interpretation and visualization in meteorology. Gagne et al. (2019), for example, demonstrate it through the CNN based discrimination of different storm morphologies being associated with extreme hail or no hail. I use another method that is based on game theory (Shapley, 1953), the Shapley Additive explanation (SHAP, Lundberg and Lee, 2017), which is able to determine in which way each predicting feature contributes to each predicted value. While this method does not guarantee causality, it can help with designing targeted hypotheses that could be tested in the future.
This project aims at answering the following questions:

- For which lead-times between 5 and 45 minutes do machine-learning models predict the probability and maximum size of hail in Switzerland better than the Lagrangian persistence?
- How many features are necessary for a ML model to perform well?
- Which data sources do the ML models use and which features are most important?
- Which information on thunderstorm environments can we gain using the SHAP interpretation method?

This chapter continues with a presentation of Coalition-3 and its connection to this part of the doctoral thesis (section 4.3). Some ML terminology is clarified in section 4.4). Section 4.5) describes the used data sources. The methods giving details on this projects choices (section 4.6)) are followed by the results (section 4.7)) which are further discussed in section 4.8). Finally, this chapter ends with some conclusions (section 4.9)) and an outlook (section 4.10)).

4.3 This projects connection with Coalition-3

This ML project is partially embedded in a MeteoSwiss project called Coalition-3 (Hamann et al., 2019). Coalition-3 developed from two earlier projects, Coalition-2 (MeteoSwiss, 2020a) and Coalition-1 (Nisi et al., 2014). Coalition stands for Context and Scale Oriented Thunderstorm Satellite Predictors Development and was initially an operational 'expert system' that provided cell-based 0-60 min nowcasts of thunderstorm severity. Coalition-1 predicted the storm severity using the concept of energy conservation, knowing that storm intensification is indicated by rapid cooling of cloud tops and increasing vertical column liquid content. A conceptual Eulerian model and a Hamilton's equations based formulation integrated information from radar, satellites, numerical weather prediction models, climatological data and a digital terrain model. The variables were used in pairs, so called modules, e.g. the evolution of the cloud top temperature was used as predictor for the liquid water content. Coalition-1 has 8 such modules, 5 of which predict the vertically integrated liquid content and 3 predict the cloud top temperature.

However, Coalition-1 could not skilfully predict the storm severity for lead-times greater than 30 minutes (Nisi et al., 2014). The Coalition-2 project focused on exploiting satellite observations to detect the early stages of thunderstorms, before the onset of rain. The infra-red channels of the Meteosat SEVIRI instrument were used to determine the cloud top glaciation, cloud top lifting and cloud optical thickness.

In Coalition-3, the conceptual, Eulerian model approach was replaced with a Lagrangian ML approach. The aim of Coalition-3 was to develop a ML-based nowcast of storm severity (in Table 4.1 called "TRT Rank") for the next 45 minutes with an update cycle of 5 minutes, starting from the TRT cell locations. Data from several data sources (see Table 4.1) describing historical thunderstorm cell properties and environments for the period April-September 2018 were extracted and assembled into a data table of samples and features (see also section 4.6.1). The prediction models were created using the extreme gradient boosted tree algorithm XGBoost (Chen and Guestrin, 2016, see section 4.6.2). This same dataset and ML method were used in this hail nowcasting project, except that the target variables were replaced. Furthermore, while the data extraction and models predicting storm severity were written in python 2.6 (see https://github.com/meteoswiss-mdr/coalition-3), this author's ML models, accompanying analyses and pre- and postprocessing scripts were written in R.

A new project, Coalition-4 (MeteoSwiss, 2021) is ongoing since autumn 2020 and is planned to end in September 2023. This project will take a step further in using deep learning methods to predict the onset and development of thunderstorms and their hazards, such as lightning, hail, and heavy precipitation. It uses data from the new geostationary satellite generation GOES-R to prepare a European version using Meteosat Third Generation. Furthermore, plans are to issue automatic warnings directly to the public, clients in aviation and civil protection.

4.4 Machine Learning terminology

This section introduces some basic terminology related to ML that was used throughout this and the Coalition-3 projects. Specific choices of parameters and model setup procedures are explained in the methods.

Machine learning: Machine learning is a group of methods aiming to discover and learn previously unknown patterns in data "without being explicitly programmed" to do so (Samuel, 1959 in Koza et al., 1996). To build a ML model, a data sample of observations or instances, in which patterns can be recognized, is required. These observations have attributes, also called features or predictors, used as inputs to the models. The target variable (also called reference or predictand) is the object that the model tries to predict. Finally, the model itself is a complex mathematical operator that estimates the relationship between predictors and target variable in the dataset (Quinto, 2020).

Supervised machine learning: A supervised ML model detects patterns in a dataset in which the "right answer" (target variable) is already provided (Shavlik et al., 1990). For example,

a supervised model may classify new data into known classes of which the model has previously learned to recognize distinctive patterns.

Data splitting: Typically, a full dataset is split into a training, validation and test dataset. The training data set usually comprises the largest fraction of the data and is the base on which the model is fitted. The validation data set is optional and is used to evaluate the model and tune hyper-parameters during the fitting process. The test dataset assesses the performance of the final model. When using k-fold cross-validation (Anthony and Holden, 1998) to tune hyper-parameters, the validation data set is not defined beforehand. Instead, the training dataset is split into k folds and during each training iteration, each fold is held out and used once as a validation data set.

Decision trees: In ML, decision trees (Shavlik et al., 1990) are a series of threshold tests, seeking to divide the dataset into sections. If the target data set contains classes, one typically uses classification trees. For these trees, the sections contain samples belonging to a class. If the decision tree is meant to predict a continuous target variable, regression trees are applied. For regression trees, the sections contain samples, with which a linear regression with the target is well fitting. The series of threshold tests is typically depicted as a tree drawn upside down, with the roots at the top. Like a natural tree, the trunk splits into branches, based on a condition that splits the data. The splitting point is also called node. Each branch can either lead to another node or, if the branch does not split anymore, a leaf. A sample of the dataset is therefore associated with one leaf or another, depending on the feature values of that sample. The complexity of trees can be further defined using hyper-parameters.

Hyper-parameter tuning for decision trees: Hyper-parameters determine the characteristics of the ML model and its training procedure. A difficulty of ML models is that the hyper-parameters leading to the best model solution are often unknown. This is why hyper-parameters are tuned during the training process. For decision trees, hyper-parameters define among others the complexity of trees, such as for example the maximum number of possible consecutive nodes (max_depth) or the degree of target purity in a leaf at which the tree will stop splitting further (min_child_weight). The target purity in a leaf describes the fraction of samples that belong to the same class (for classification) or simply the number of samples in a leaf (for regressions). Other hyper-parameters aim at preventing overfitting (subsample, etc. see below).

Over- and Underfitting: Initial training iterations are typically characterized by an improvement of the fit with both the training and testing dataset. However, once the fit with the validation dataset proceeds to deteriorate, the model starts to overfit. Overfitting can be avoided by tuning hyper-parameters and/or by using k-fold cross-validation. If the models are trained with too few iterations, then the model underfits, meaning that an improvement of the model performance is still possible.

(Stochastic) gradient boosted trees: This algorithm builds ensembles of decision trees such

that each new tree corrects the errors made by the previous one (Friedman, 2001). The error is measured using a loss metric. The learning rate (a hyper-parameter called eta) determines how fast one tries to move in the direction of lowering the loss. The objective function is the quantity being optimized (Kuhn and Johnson, 2013); in a binary problem, it can be a logistic regression function and the output of a prediction therefore a probability. The stochastic part of gradient boosting consists of not using the entire data sample in each iteration, but to randomly sample a fraction of the data (Friedman, 2001 in Kuhn and Johnson, 2013). The hyper-parameter defining the size of that fraction is called subsample. Another similar hyper-parameter (colsample_bytree) does not reduce the number of data samples (as subsample) but the number of features by a specific fraction.

Extreme gradient boosted trees (XGBoost): Compared to stochastic gradient boosted trees, XGBoost has the capability of building several nodes within each depth of each tree in parallel and, therefore, is a faster algorithm. This parallelization is further supported by sparsity aware algorithms. Finally yet importantly, the objective function in XGBoost includes two regularization terms that penalizes the complexity of a model and further helps to avoid overfitting (Chen and Guestrin, 2016).

4.5 Data

Several data sources that capture properties of thunderstorms and their environments are used to train the models for each of the 30'675 time steps and locations of hailstorms that occurred during 37 days between May and September 2018. The choice of time range is related to the close collaboration of this project with MeteoSwiss' project Coalition-3. The diurnal and seasonal distribution of the analyzed time steps are shown in Fig. 4.1. Observations of radar and satellite instruments, NWP forecasts and other auxiliary data (see Table 4.1) are used. Radar and satellite data (using the rapid scan service) are observed with a temporal resolution of 5 min. COSMO data is available hourly and interpolated to the desired time. The plan was to test different methodical approaches on the 2018 data set. Final nowcasting models would then be calculated, once a larger data set was available. Unfortunately, due to reasons that were outside of my influence, it was impossible to extend the training dataset to the entire convective seasons of 2018, 2019 and 2020. That is the reason why the models are trained with data from the 2018 convective season only.

4.5.1 Radar

Single-polarisation radar variables from the Swiss radar network (Germann et al., 2015) provide the information on location and intensity of thunderstorms. These variables are available every 5 minutes on a Cartesian 1 x 1 km^2 grid. The thunderstorm radar tracking algorithm TRT uses the radar reflectivity to detect, track and nowcast thunderstorms. The target variables of the statistical models in this work are two radar-based hail products: the "probability of hail" (POH) and the "maximum expected severe hail size" (MESHS). POH is derived from the

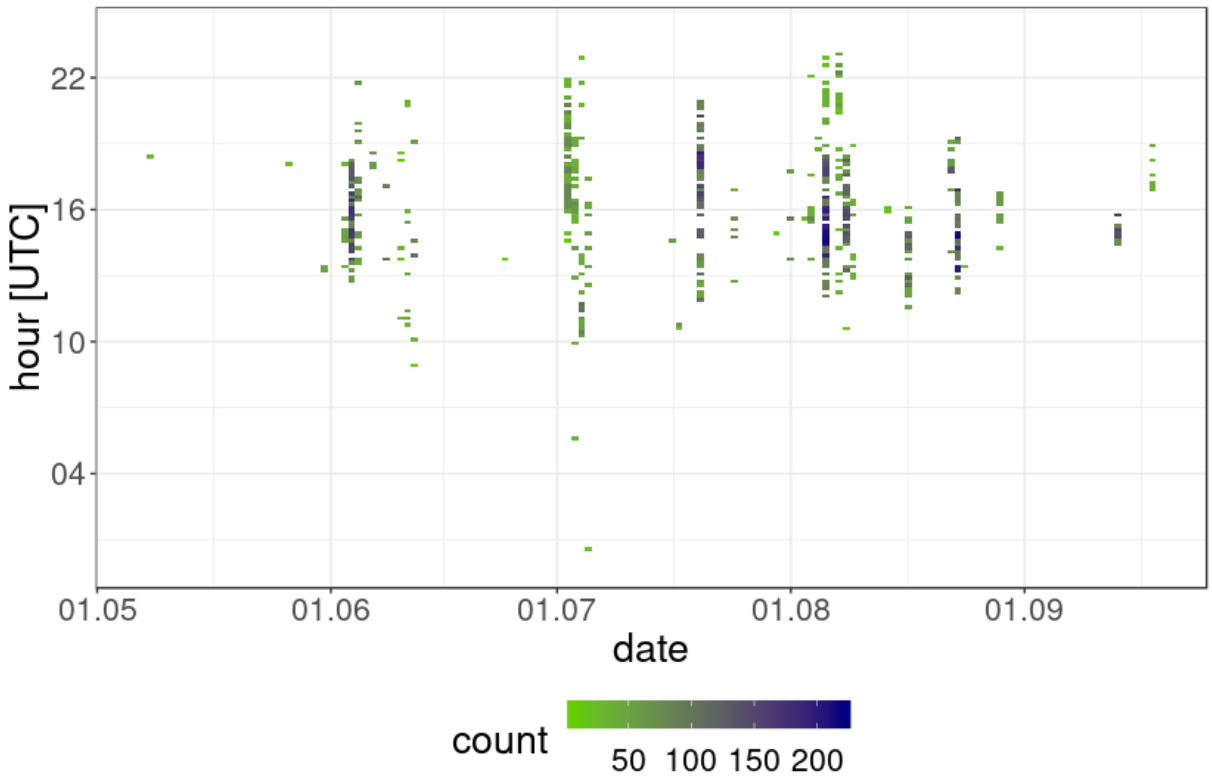

Figure 4.1: Number of data samples per day and time of the day in the year 2018, used to train and test the XGBoost hail nowcasting models.

height difference between the maximum altitude with a 45 dBZ reflectivity (echo top 45) and the freezing level height. The freezing level height information stems from the operational NWP model COSMO-CH. POH values vary between 0 % and 100 % and indicate the probability at ground-level (Foote et al., 2005b based on Waldvogel et al., 1979). MESHS is derived from the difference between the freezing level height and the 50 dBZ echo top. It provides estimates of the maximum hail diameter per km^2 at the ground for hailstones $\geq$ 2 cm (Joe et al., 2004 based on Treloar, 1998).
Additional radar variables were used as input for the statistical model. These include gridded 2-D radar variables such as e.g., the precipitation intensity and storm cell-averaged characteristics obtained using the TRT algorithm (see Table 4.1 for a full list of all variables).

4.5.2 Satellite

Several infrared and water vapor channels from the Meteosat Second Generation Satellites (MSG; Schmetz et al., 2002) provide information on cloud properties and upper-level dynamics (see Table 4.1 for details). For this work, only channels without a solar component were used, to guarantee a consistent performance during day and night. The so called window channels, e.g. the 10.8 μm channel, are absorbed little by the atmosphere; hence they are useful to observe the surface or cloud top temperatures and their temporal derivative represents the cloud top cooling. Channel

combinations with differences in absorption of water and ice are used to estimate the cloud top phase or the cloud optical thickness, particularly the channel difference 12.0 μm – 10.8 μm (CG3). The water vapour channels (6.2 μm and 7.3 μm) provide information about the water vapour concentration and dynamics in the middle and upper atmosphere. The satellite data has a subsatellite resolution of 3 km, which translates into 3 x 5 km in Europe. The data was regridded to 1 km resolution. While the satellite data has a lower spatial resolution, developing thunderstorms can be detected up to ten minutes before they are detected by the radar (Nisi et al., 2014).

4.5.3 Numerical weather prediction model COSMO-1

The Consortium for Small-Scale Modeling (COSMO)-Model is a non-hydrostatic, limited-area, numeric weather prediction model that is based on primitive, thermo-hydrodynamical equations describing compressible flow in a moist atmosphere (see http://cosmo-model.org/). The basic version of the model was developed by the German Weather Service and further adapted for Switzerland by MeteoSwiss. COSMO-1 is a deterministic model simulated every 3 hours, with a spatial grid-resolution of 1 km and model output is available hourly (MeteoSwiss, 2020b). Convective variables, such as CAPE, the surface lifted index (SLI) or the level of free convection (LFC) and other characteristics of the atmosphere at different vertical levels are extracted from the analysis of COSMO-1 (see Table 4.1).

4.5.4 Lightning

Lightning measurements from the Météorage lightning detection network (EUCLID, see www.euclid.org) exploit possible increases in predictability for example from lightning jumps (see e.g., Nisi et al., 2020, for more details). All types of lightning (cloud to ground, intra-cloud) are considered (see Table 4.1).

4.5.5 Topographical information and other data

Topographical data from the COSMO-1 model describing the complex topography and measuring the effect of thunderstorms passing over hills and mountains are measured through the aspects, slopes and altitudes. Other data such as the sunshine angle, optical motion speed in u and v directions, the cell locations and propagation speed are taken into account (see Table 4.1).

Table 4.1: List of variables as well as their source, abbreviations and processing method (more about last column in section 4.6.1). The bold words in the "source" column are the names of the data sources when mentioned in the results.

sources	**abbreviations**	**names**	**units**	**statistics**
radar	POH	probability of hail	%	all + nonmin + pixc
	VIL	vertically integrated liquid	kgm^{-2}	
	MESHS	maximum expected severe hail size	mm	
	ET15, ET20, ET45, ET50	Altitude of maximum reflectivity for 15 dBZ, 20 dBZ, 45 dBZ and 50 dBZ	m	
	precipitation	5-minute precipitation intensity	mmh^{-1}	
	MaxEcho	maximum column reflectivity	dBZ	all + nonmin + pixc
	RADAR_FREQ_QUAL	radar frequency quality	dimensionless	all
radar + Cosmo-1	TRT Rank	thunderstorm radar tracking severity	dimensionless	single values
	TRT Rank diff	thunderstorm radar tracking severity difference to the rank at t_0	dimensionless	
satellite (MSG)	IR_087	MSG IR channel 7 (8.7 μm); helps to discriminate ice and water clouds	K	all
	IR_097	MSG IR channel 8 (9.7 μm; ozone absorption band; indicates areas with tropopause folding)		
	IR_108	MSG IR channel 9 (10.8 μm; particularly sensitive to high thin Cirrus clouds)		
	IR_120	Infrared channel 10 (12 μm)		
	IR_134	Infrared channel 11 (13.4 μm; CO2)		
	WV_062	Water vapour channel 5 (6.2 μm; high altitude (∼350hPa) water vapor content of atmosphere)		
	WV_073	Water vapour channel 5 (7.3 μm; mid altitude (∼500hPa) water vapor content of atmosphere)		
	CD1	cloud depth indicator 1 (WV_062-IR_108)		

continued . . .

... continued

sources	abbreviations	names	units	statistics
	CD2	cloud depth indicator 2 (WV_062-WV_073)		
	CD4	cloud depth indicator 4 (WV_073-IR_134)		
	CD5	cloud depth indicator 5 (WV_062-IR_097)		
	CG1	cloud glaciation indicator 1 (IR_087-IR_120)-(IR_120-IR_108)		
	CG2	cloud glaciation indicator 2 (IR_087-IR_108)		
	CG3	cloud glaciation indicator 3 (IR_120-IR_108)		
Cosmo-1 **model**	TWATER	total column water content	kgm^{-2}	all
	tropopause height	tropopause height	m	
	tropopause temperature	tropopause temperature	K	
	tropopause pressure	tropopause pressure	Pa	
	FF_10M	wind speed 10 m above ground	ms^{-1}	
	VMAX_10M	wind gust speed 10 m above ground	ms^{-1}	
	CAPE_MU	most unstable convective available potential energy (CAPE)	Jkg^{-1}	
	CAPE_ML	mean surface layer CAPE	Jkg^{-1}	
	CIN_MU	most unstable convective inhibition (CIN)	Jkg^{-1}	
	CIN_ML	mean surface layer CIN	Jkg^{-1}	
	SLI	surface lifted index	K	
	LCL_ML	lifting condensation level (mixed layer)	m	
	LFC_ML	level of free convection (mixed layer)	m	
	T_SO	soil temperature	K	
	T_2M	air temperature 2 m above ground	K	
	TD_2M	dew point temperature 2 m above ground	K	
	GLOB	global radiation	Wm^{-2}	
	PS	surface pressure (not reduced)	Pa	

continued ...

... continued

sources	abbreviations	names	units	statistics
	MSLP	surface pressure reduced to msl	Pa	
	MSLP tendency	sea surface pressure tendency	Pa h^{-1}	
	HZEROCL	freezing level height	m	
	WSHEAR_0-3km	bulk wind shear (surface - 3 km)	ms^{-1}	
	WSHEAR_0-6km	bulk wind shear (surface - 6 km)	ms^{-1}	
	PV 300 hPa, 500 hPa, 700 hPa	ertel potential vorticity at 300 hPa, 500 hPa and 700 hPa	$Km^2kg^{-1}s^{-1}$	
	THETAE 300 hPa, 500 hPa, 700 hPa	equivalent potential temperature at 300 hPa, 500 hPa and 700 hPa	K	
	MCONV 300 hPa, 500 hPa, 700 hPa	moisture convection at 300 hPa, 500 hPa and 700 hPa	gpm	
	RELHUM 300 hPa, 500 hPa, 700 hPa	relative humidity at 300 hPa, 500 hPa and 700 hPa	%	
	U_OFLOW, V_OFLOW	u and v components of optical cell motion	ms^{-1}	
lightning	THX_densIC	lightning density inter/intra-cloud	km^{-2}	
	THX_densCG	lightning density cloud to ground	km^{-2}	
	THX_curr_abs	lightning absolute current	kA	
	THX_curr_neg	lightning negative current	kA	
	THX_curr_pos	lightning positive current	kA	
	THX_dens	total lightning density	km^{-2}	
topo-graphical data	Topo_ Aspect	topographic aspect (negative dot product of local aspect and optical motion vector)		
	Topo_Altitude	topographic altitude	m	
	Topo_Slope	topographic slope	dimensionless	
Cosmo-1 model	SOLAR_ TIME_SIN	sine component of solar declination angle	dimensionless	

continued ...

. . . continued

sources	**abbreviations**	**names**	**units**	**statistics**
	SOLAR_ TIME_COS	cosine component of solar declination angle	dimensionless	
TRT **cell** properties	CG	number of cloud to ground (CG) lightning in cell; past 5 min	dimensionless	
	CG_minus	number of negative CG lightning in cell; past 5 min	dimensionless	
	Dvel_x	spread of cell velocity in East West direction	kmh^{-1}	
	Dvel_y	spread of cell velocity in South North direction	kmh^{-1}	
	ET15	maximum ET15 in cell	km	
	ET15m	mean ET15 in cell	km	
	ET45	maximum ET45 in cell	km	
	ET45m	mean ET45 in cell	km	
	POH	maximum POH in cell	km	
	RANK	TRT rank [0:4]	dimensionless	
	RANKr	detailed TRT rank (0:40)	dimensionless	
	VIL	maximum vertically integrated liquid	kgm^{-2}	
	angle	ellipse angle of ellipse delimitating TRT cell	degrees	
	area	area of the TRT cell	km^2	
	det	detection threshold of TRT cell	dBZ	
	ell_L	ellipse, semi-major axis	km	
	ell_S	ellipse, semi-minor axis	km	
	ich	zonal cell position in Swiss coordinate system	dimensionless	
	jch	meridional cell position in Swiss coordinate system	dimensionless	
	lat	latitude of cell center	degree	
	lon	longitude of cell center	degree	
	maxH	maximum MaxEcho within TRT cell	km	
	maxHm	mean MaxEcho within TRT cell	km	
	perc_CG_plus	percentage of positive CG lightning in cell, past 5 min	%	
	vel_x	velocity of cell in East West direction	kmh^{-1}	
	vel_y	velocity of cell in South North direction	kmh^{-1}	

4.6 Methods

4.6.1 Data retrieval and preprocessing

The following data retrieval methods were implemented in the course of the Coalition-3 project. The thunderstorms that pass through the radar coverage area of the Swiss radar network are tracked using the TRT algorithm. While in an operational setting this distinction would not be made, in our dataset we regard only the cells with a minimum lifetime of 15 minutes. Each 5-minute TRT cell position is treated using the following procedure: First, a circle with a diameter of 23 km is placed on top of the TRT cell at t_0 (t_0 = t; t_0 being the moment of the last measurement; Fig. 4.2). The size 23 km is chosen such that there is no influence from changing cell sizes along their lifetimes and such that potentially important processes from thunderstorm inflow regions are taken into account (Gagne et al., 2017 and Lagerquist et al., 2017 in Gagne et al., 2019). Second, the likely position of this circle is extrapolated backward and forward every 5 minutes until t-45min and t+45min. The motion vectors are determined using the software package pySTEPS (Pulkkinen et al., 2019). The algorithm is applied only to the strongest 1% reflectivities (over all timesteps and the entire radar space) to avoid any influence of weak precipitation areas and to concentrate on the movement of the cell rather than the movement of the background flow. Then, well suited targets are chosen for tracking with the algorithm of Shi and Tomasi (1994). These movements are translated into 1 x 1 km^2 gridded motion vectors using the Lucas and Kanade (1981) algorithm. Each motion vector represents the optical cell motion between two consecutive 5-minute radar time steps. Finally, the three motion vectors between t-15min and t_0 (t_0 — t+15min) associated with the current cell position are averaged to yield the vector $\overrightarrow{v_p}$ ($\overrightarrow{v_f}$), in which direction the initial circle at t_0 is translated backward (forward). In an operational setting, $\overrightarrow{v_f}$ would be unknown and the last derived $\overrightarrow{v_p}$ would be used instead. This deviation from the potential operational setting was made to increase the likelihood of the cells staying within the circles of the future positions and thus to decrease the chance of a good model skill being attributed to the future cells not truly being within the circles. Furthermore, the reason for not using known past cell positions as the centers of circles is to avoid issues related to cells splitting and merging.

Within all circles, 12 statistics are calculated for each variable describing the thunderstorm and its environment: sum, mean, standard deviation, minimum, maximum, the 1st, 5th, 25th, 50th, 75th, 95th and 99th percentile. Each statistic of each variable of each time step between t-45 and t_0 is a feature and is used to predict the target variables. Some ML methods require the features to be further preprocessed, e.g. through normalization (e.g., Manzato, 2013), however this is not necessary for XGBoost.

The radar fields may contain many zeros or low values. Therefore, for all nine radar variables (Table 1), statistics are calculated once on all values (all) and again on all values excluding the minimum (nonmin). In addition, the number of grid-boxes with non-minimum values ("pixc") of each radar variable are used. For the maximum radar reflectivity (CZC), the number of grid-boxes

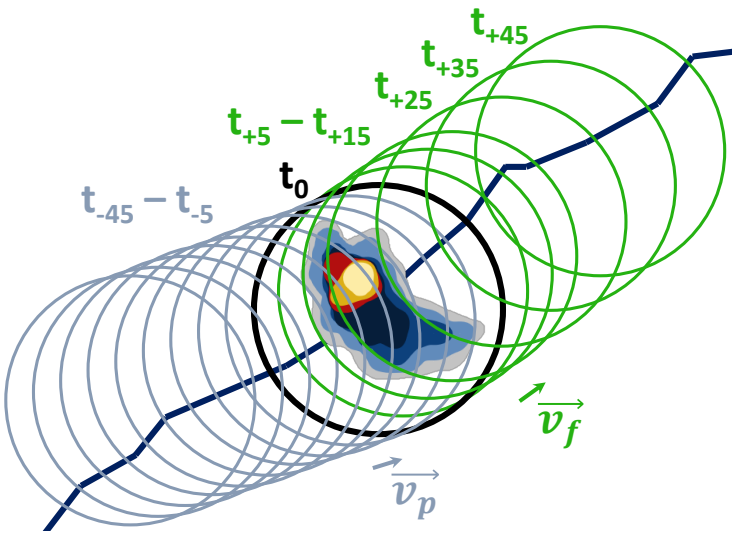

Figure 4.2: Visualization of past (grey) and future (green) cell positions for one thunderstorm, of which the features are extracted and for which nowcasts are produced. The example thunderstorm position is shown within the black circle at t_0. The dark blue line indicates this thunderstorms' example TRT track and the vectors $\overrightarrow{v_p}$ and $\overrightarrow{v_f}$ indicate the average optical cell motion that determines the forward and backward extrapolated circle positions. The circles have a diameter of 23 km.

with values larger than 57 dBZ is also counted. For all count features, only the sum is used as a statistic. The 12 statistics are calculated on all the other variables with grid-data as well. The TRT cell properties are calculated within TRT cell contours (see Hering et al. 2008). Overall, the total number of features is 10'610. In total 30'675 samples of cell histories are calculated in this way.

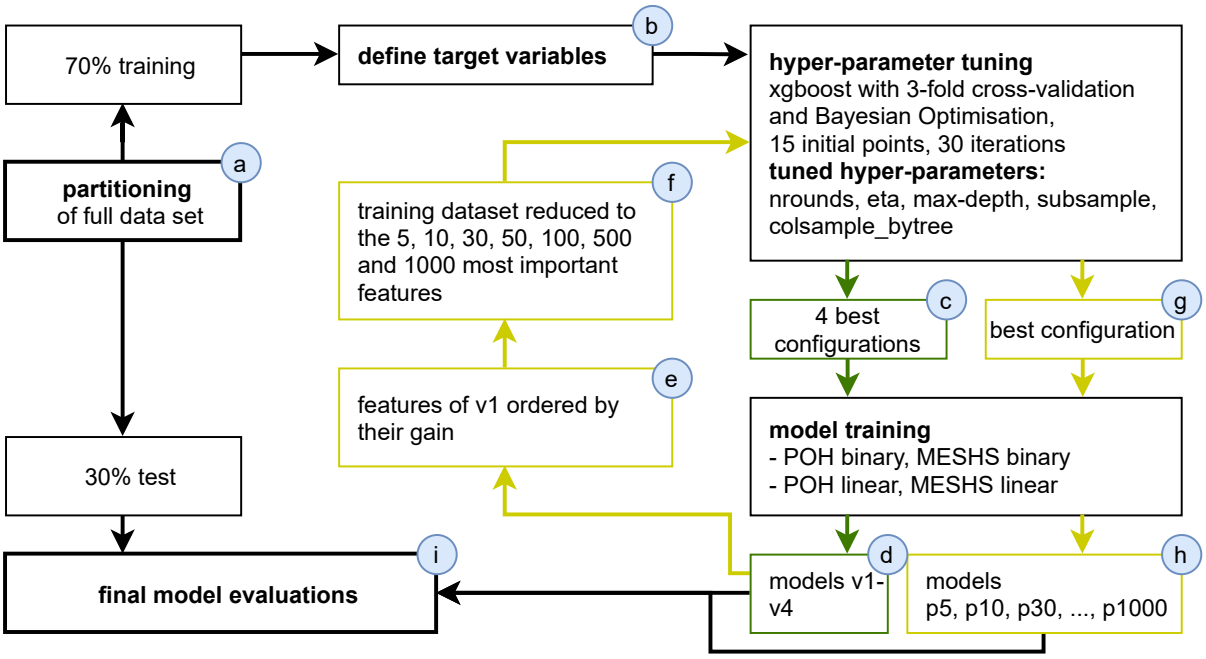

Figure 4.3: Chart presenting the XGBoost model building process, starting at a) and ending at i).

4.6.2 XGBoost: models and configurations

Figure 4.3 summarizes the steps taken for creating the nowcasting models from the initial assembled data table to the evaluation. These choices were made independently from Coalition-3. After initially splitting the full data set into 70% training and 30% test data set (Fig. 4.3a), the target data sets are defined (Fig. 4.3b). The ML algorithm XGBoost is trained to nowcast the time steps t+5', t+10', t+15', t+25', t+35' and t+45' after t_0[2]. For all lead-times, a first run uses K-fold cross-validation and BO to determine the optimal hyper-parameters for four versions of models that use all features (Fig. 4.3c). Subsequently, the models are trained again (without cross-validation) with the previously determined optimal hyper-parameter configurations (Fig. 4.3d). The tuning and training process is repeated with a reduced number of top features, as ranked by the first initial model (Fig. 4.3e-h). Finally, all models are evaluated using the test data set (Fig. 4.3i).

All XGBoost models are created in R using the R-packages "xgboost" (Chen et al., 2020), "ml-BayesOpt" (Matsumura, 2019) and "rBayesianOptimization" (Yan, 2016). Unless specified differently, the configurations are set to default.

[2]If approximately 5 minutes were needed to extract all data and run the models, then a t+10' nowcast would truly be a t+5' nowcast.

4.6.2.1 Defining the target data sets and model configurations

The target variables are the maximum POH and MESHS within the forward extrapolated positions of the circles. The POH dataset is strongly skewed, with half the values being zero and a third of the remaining values being greater than 80 %. Other studies have used postprocessing methods, e.g. an isotonic regression to correct highly uncalibrated probabilistic output (Niculescu-Mizil and Caruana, 2005; Lagerquist et al., 2017; McGovern et al., 2019a; Burke et al., 2020; Flora et al., 2020). I deal with this issue by splitting the task of predicting whether POH is < 10 % or ≥ 10 % and predicting the actual POH value between two prediction models. The binary XGBoost model has the task of predicting the probability of POH being <10 % or $\geq$ 10 % with a logistic regression as the objective function. The POH values ≥ 10 % are then predicted using the linear XGBoost model that has a linear regression as the objective function. MESHS, having the particularity of not being defined for 0 – 2 cm, is also predicted using the same configuration as POH with two models applied consecutively. The binary XGBoost model predicts whether MESHS will have a value ≥ 2 cm or $= 0$ cm and the linear XGBoost model predicts the ≥ 2 cm MESHS values. Before training the linear XGBoost models, the samples with zero values in their target variable are therefore removed such as to decrease the influence of zero values on the resulting prediction. For binary models the loss is measured by the negative log-likelihood function (logloss) and for linear models by the root-mean-squared-error (RMSE).

4.6.2.2 Hyper-parameter tuning

Tuning hyper-parameters of a model has two main purposes. The one is discovering the best model configuration as efficiently as possible and the other is to avoid over- or underfitting the models. I tune the following set of hyper-parameters. Initial tests (not shown) have framed the final range of values indicated in the brackets.

- eta (0.1–1): the step size shrinkage, which controls the learning rate
- max_depth (4–6): the maximum depth of a tree, which controls the number of consecutive nodes
- subsample (0.1–1): Which fraction of the original sample of data points should be used at each iteration
- colsample_bytree (0.4–1): Which fraction of features should be used at each iteration
- nrounds (200–750): Optimal number of iterations to reach the smallest testing error; it is also the number of trees

XGBoost further allows tuning alpha, lambda and gamma. Alpha removes unimportant features, a process also known as the L1 regularization. Lambda limits the possibility of few, very important features being too dominating, also called the L2 regularization. Gamma, also known as the Lagrangian multiplier, defines the minimum loss reduction needed to make a further partition on a leaf node of the tree ((Chen and Guestrin, 2016)). The XGBoost models predicting storm

severity in Coalition-3 were tuned for these parameters, lowering the RMSE by 1–8 % for the different lead-times (Ulrich Hamann, personal communication).

To further prevent overfitting, I apply a 3-fold cross validation with early stopping. To profit from early stopping, the model is tested against the folded validation data sets after each iteration. Thanks to early stopping, the model stops adding more trees as soon as the test performance has continuously deteriorated for more than a specified number of iterations that is subjectively set in advance (called early stopping rounds; in this project subjectively set to 5).

Which hyper-parameter combinations are tested? Default methods to choose configurations are grid-search, i.e. testing all combinations of a set of parameter values, or random search, i.e. random parameter combinations are sampled. In this project, I use the Bayesian Optimisation method (BO, Mockus et al., 1978 in Snoek et al., 2012). BO is more efficient than the previously mentioned methods, because it pays attention to the information given by previously tested hyper-parameter configurations to choose the configuration that will be tested next. The models are first trained with a randomly sampled set of 15 initial hyper-parameter configurations. Subsequently, 30 additional iterations determine the ideal set of hyper-parameters. See more details on the applied BO method in the Appendix section B.1.

4.6.2.3 Final training and reducing the number of features

Noticing that BO suggests several different hyper-parameter configurations with very similar results (see example in Appendix Fig. B.1), not only the best configuration but the four best configurations modeled with all features are retrained. Hereafter, these four models are called v1-v4. These models indicate the robustness of results, with respect to the choice of hyper-parameters.

In an operational setting, the available computation time is limited. A reduction in number of input features reduces computing time. Furthermore, fewer features make the algorithm less susceptible to failures in data delivery. To understand how a reduced number of features affects the model performance, models with only the most important 5, 10, 30, 50, 100, 500 and 1000 features are trained again (called p5, p10, etc.). The importance of features is measured by the v1 model gain. The gain measures the average gain of all splits of a feature, divided by the amount of information in the split itself (see Kuhn and Johnson, 2013, p. 378). For the p-models, hyper-parameters are also first tuned using K-fold CV and BO and later used to retrain the actual final models.

In summary, this project created for each of the 6 lead-times and 4 model types (POH binary, POH linear, MESHS binary and MESHS linear) 11 models with distinct configurations (v1-v4, p5-p1000).

4.6.3 Model evaluation

Finally, the quality of the models is evaluated using the 30 % of the dataset that had been set aside as a hold-out test dataset. For the linear XGBoost models predicting non-zero POH, any prediction smaller than zero or greater than 100 is set to 0 or to 100, respectively. For the linear XGBoost models predicting MESHS, any predicted value < 2 cm is set to 2 cm. The models are evaluated with binary, probabilistic and linear verification scores (see Appendix sections B.2 and B.3 for equations). In all cases, the model performances are compared to Lagrangian persistence nowcasts and, for the binary models, to the climatological nowcasts. The persistence nowcast extrapolates the target values at t_0 forward in space to the future positions. The climatological nowcast predicts hail using a randomly perturbed target dataset.
Binary verification methods require probabilistic predictions to be binarized. Unless clearly stated otherwise, any probability > 0 is converted to 1 (POH $\geq$ 10 %, MESHS $\geq$ 2 cm). The dotted lines in the performance diagrams (Fig. 4.6 and 4.12) explore the effect of altering this threshold to higher values in decimal steps.

4.6.3.1 Binary and probabilistic verification scores

The binary scores are calculated using the contingency table (Table B.1), which counts the number of hits, false alarms, misses and correct rejections. The applied binary skill scores include the symmetric extreme dependency index (SEDI; Ferro and Stephenson, 2011). This index is base rate independent, complement symmetric and is an index that is commonly used to deal with the verification of rare events. The SEDI can range between -1 and 1 with 1 being the best result and 0 indicating no skill. The approximate 95 % confidence interval of the SEDI gives an estimate of uncertainty of the SEDI based on the number of values (Ferro and Stephenson, 2011).

The probability of detection (POD, also called Hit Rate) measures the fraction of correctly predicted observed events out of the total events observed. The false alarm ratio (FAR) is the ratio of the false alarms to the total events forecasted. Note that the SEDI is calculated from the POD and false alarm rate, which should not be confused with the FAR. The POD and FAR are used to create a performance diagram (Roebber, 2009). In the performance diagram, the FAR is hidden in the success ratio, which is calculated with 1 – FAR. A perfect forecast will have a success ratio and a POD of 1. A point on the performance diagram that is not on the diagonal between (0,0) and (1,1) indicates a frequency bias. The frequency bias is the fraction of observed events divided by the predicted events. A point above (below) the diagonal is a sign of overforecasting (underforecasting); it means that there are more (less) false alarms than misses. Finally, the critical success index (CSI, also called threat score) is sensitive to both misses and false alarms and is calculated by dividing the hits by the sum of the number of hits, false alarms and misses (Roebber, 2009, ; see also Appendix section B.2).

4.6.3.2 Linear verification scores

Linear verification scores (see also Appendix section B.3) include the mean error (ME, also called linear bias), the mean absolute error (MAE), which measures the magnitude of the error, and the root-mean squared error (RMSE), which is similar to the MAE except that it strongly penalizes the greater errors. Furthermore, the linear correlation measures the linear association between the prediction and target. The Taylor diagram (Taylor, 2001) graphically represents the cosine-relationship between the centered RMSE, the correlation and the standard deviations of the prediction and target data set. Compared to the RMSE, the centered RMSE is debiased as shown in equation 4.1 with y as the predicted values, x as the target values and N as the lengths of the vectors x and y.

$$cRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(y-\overline{y})-(x-\overline{x})]^2} \tag{4.1}$$

In the Taylor diagrams in Fig. 4.7 and 4.13, the centered RMSE and the standard deviations are normalized by the standard deviation of the target data set such that different models with different target data sets are comparable (Kärnä and Baptista, 2016). The Taylor diagrams compare the XGBoost model performances to the persistence in two ways: The first evaluation shows the combined error of nowcasting the occurrence of POH/MESHS and its value (e.g., Fig. 4.7a). The second shows the conditional error, provided that POH/MESHS is non-zero (e.g., Fig. 4.7b). For the second evaluation, the focus lies on evaluating the linear model, independent of the binary models' skill.

4.6.4 Post-processing with Probability Matching (PM)

Results of linear prediction models show that when the difficulty of predicting a target increases, the standard deviation of the prediction decreases and the model tends to predict the value with lowest cost. This difference in standard deviation and any additive bias can be corrected using probability matching (Ebert, 2001). This method consists of displacing the predicted values, as fitted on an empirical cumulative density function (ecdf), to its target ecdf counterpart (see e.g., Fig. 4.4). This procedure removes any difference in standard deviation and therefore makes a comparison with the persistence nowcast easier. The results of the probability-matched predictions have a slightly larger RMSE and lower correlation (see e.g., Fig. 4.7).

4.6.5 Model interpretation

One major criticism of machine learning based weather prediction models is their apparent lack of meteorological interpretability. While the detected linear or non-linear relationships between variables may truly have no direct meteorologically explainable background, observing the feature importance may lead to targeted hypotheses. The SHAP value (Shapley Additive explanation; Lundberg and Lee, 2017), a method that is based on game theory (Shapley, 1953), derives from the model the individual contribution of each feature to a particular prediction. The SHAP value measures the average of all permutations of marginal contributions of each individual feature to

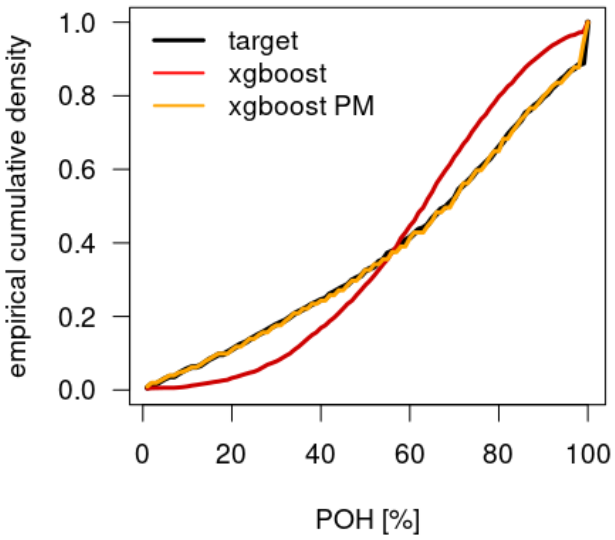

Figure 4.4: Example of probability matching for the prediction using the t+25' POH linear model v1. The XGBoost prediction values (ecdf in red) are corrected such that the new probability matched ecdf (XGBoost PM; in orange) fits the ecdf of the target data set (in black).

the final model prediction. If we consider N features with S being a subset of N and $\nu(S)$ the contribution of the S features to the model prediction, then feature i's marginal contribution is $\nu(S \cup \{i\}) - \nu(S)$. The average of all permutations leads to a SHAP value of:

$$\phi_i(N, \nu) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [\nu(S \cup \{i\}) - \nu(S)] \tag{4.2}$$

I recommend reading more on SHAP and other model interpretation methods in e.g. (Molnar, 2021).

For binary predictions of POH or MESHS, a positive (negative) SHAP value increases (decreases) the probability of POH $\geq$ 10 % or MESHS $\geq$ 2 cm. For the linear models, a positive (negative) SHAP value increases (decreases) the predicted values of POH or MESHS. The graphs related to SHAP values in the results section were made using minimally modified functions from the R package "SHAPforxgboost" by Liu and Just (2020). The data sources of all features and of the top 100 features (according to mean absolute SHAP values) of the p1000 models are counted by lead-time and by which time steps (t-45',..., t_0) the statistics of variables were taken at (Fig. 4.8). SHAP summary plots show SHAP values and their absolute mean, for the 30 most important features of the p1000 models.

4.7 Results

4.7.1 Probability of hail

4.7.1.1 Model evaluation

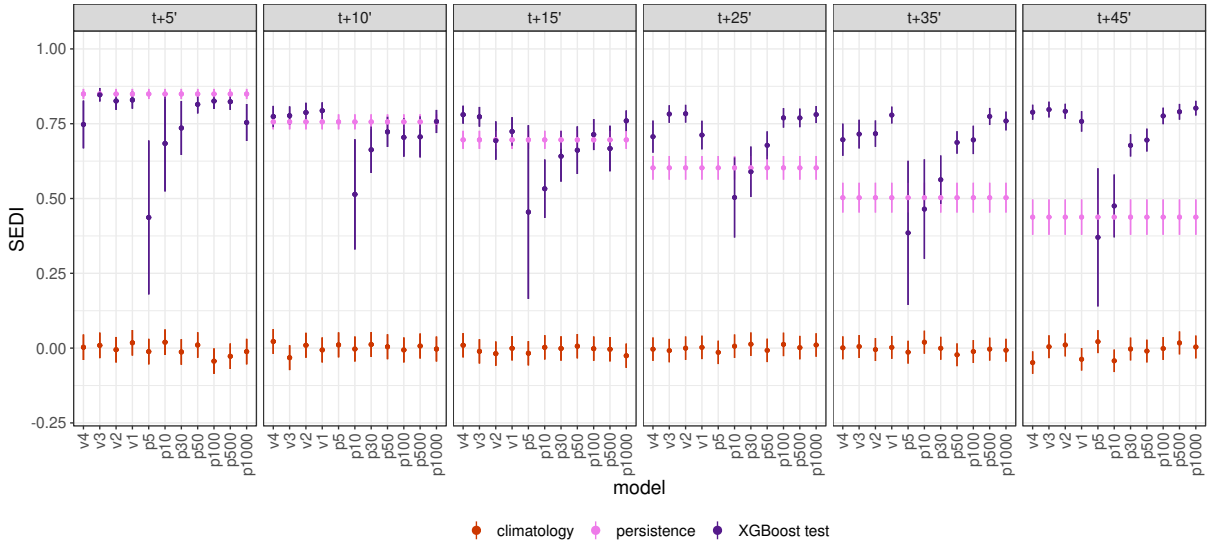

Figure 4.5: SEDI (dots) and their 95 % confidence intervals (vertical bars; see eq. B.4) of POH binary predictions for each model version (x-axis) and lead-time, with the XGBoost test dataset (dark purple), the climatology (red) and persistence (light purple).

Figure 4.5 shows the SEDI and their 95 % confidence intervals for each model and each lead-time. As the lead-time increases, the SEDI of the Lagrangian persistence decreases from 0.8 at t+5' to 0.45 at t+45'. At t+5', the persistence fares better than all binary XGBoost models, except model v2. At all larger lead-times, the best XGBoost models produce better nowcasts than the persistence. The gap between model and persistence increases with increasing lead-time. The SEDI of the best models stay mostly at 0.76 and the best model for t+45' has a higher SEDI than t+35' or t+25', with a SEDI of 0.78. The reason why SEDI remains so high, despite the increasing lead-time is the simultaneous decrease in number of hits and increase in number of correct rejections. The number of false alarms always remains low and the number of misses increases slowly up to t+25' and then decreases again for higher lead-times (see Fig. B.2 and B.3 in the Appendix). The number of POH < 10 % cases increases with lead-time, since the likelihood that a thunderstorm has dissolved at t+45' is higher than at shorter lead-times. While according to BO, v1 should be the most skilled prediction model, sometimes v2, v3 or v4 yield higher SEDI scores against the test dataset. For each lead-time, reducing the number of features of v1 to its top 5, 10 or 30 features has a strongly detrimental effect on the model skill. An ideal number of features seems between 50–500 features for t+5' and between 500–1000 features for larger lead-times.

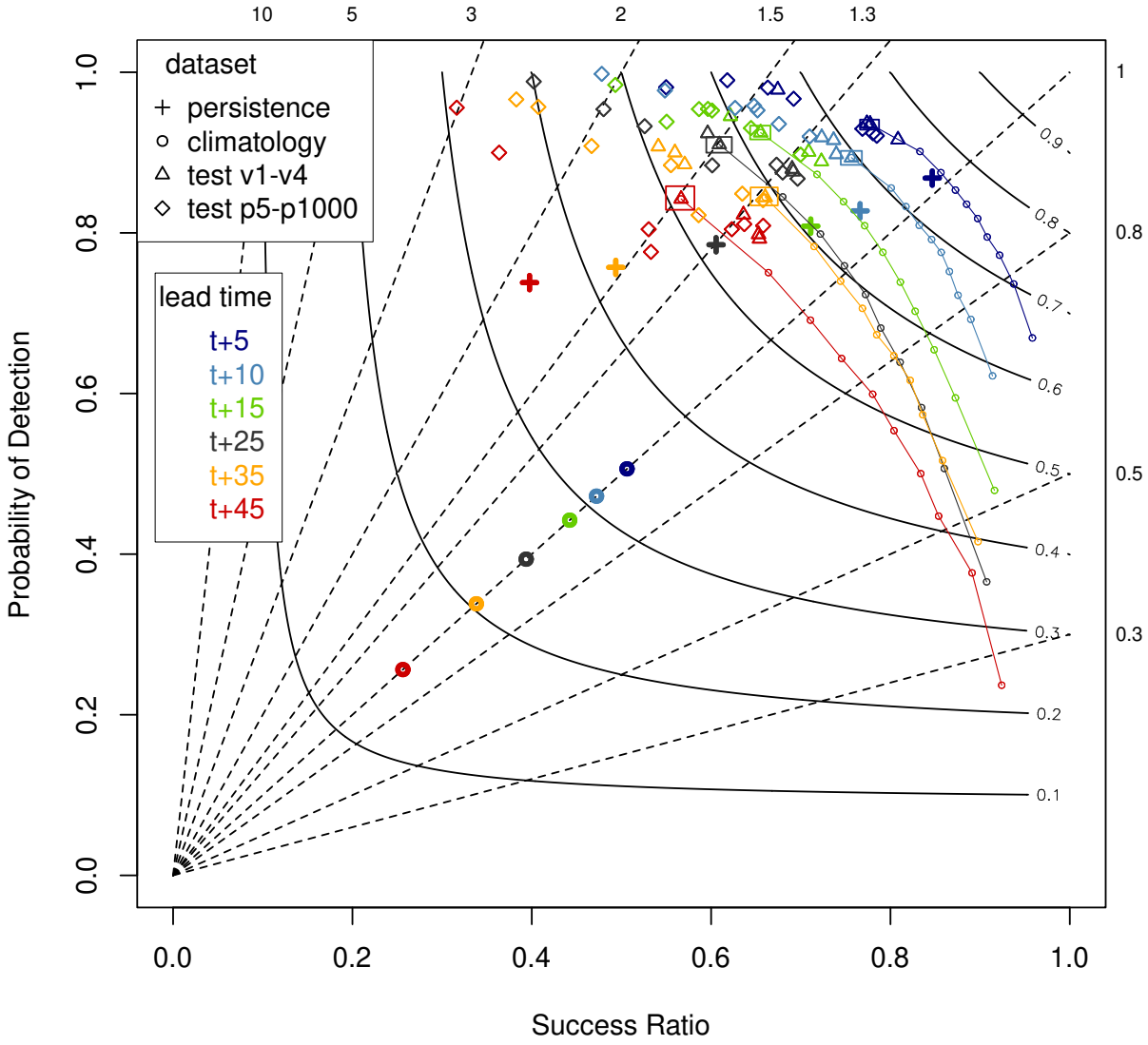

Figure 4.6: Performance diagram for binary predictions of POH at different lead-times (colors) and for the different configurations (crosses = Lagrangian persistence, triangles and diamonds = all binary XGBoost models v1-v4 and p5-p1000, circles = climatology). The y-axis shows the POD, the x-axis the SR (1-FAR), the curved lines indicate the CSI and the dashed lines the frequency bias. The colored boxes show the confidence interval (CI) calculated by percentile bootstrapping the observations and predictions for the XGBoost models v1. The threshold (th) that separates predicted binary values (0 is POH $<$ 10 % vs. $x >$ th = POH $\geq$ 10 %) is 0.0. For v1, the dotted lines indicate the performance values for th between 0.0 and 0.9 in steps of 0.1. Other models of the same color have similar CI and different binary thresholds behave similarly to the examples shown.

The performance diagram in Fig. 4.6 explores the quality of models and persistence according to the POD, Success Ratio, CSI and frequency bias. While Fig. 4.5 clarified the effect of different model versions and different numbers of features in a model, this diagram focuses more on why

the model quality decreases with increasing lead-time and reduced number of features. As in Fig. 4.5, the performance of the best binary XGBoost model is equal (t+5') or better than the persistence (other lead-times; Fig. 4.6). For t+5', the best model has a CSI of 0.75, a POD between 0.8 and 0.9 and a Success Ratio of 0.85. The persistence (climatology) has a CSI of 0.75 (0.34). As the lead-time increases, the CSI values of the best model versions decrease in steps of approximately 0.02-0.07. To maximize the CSI, a slightly higher binarization threshold of about 0.2-0.6 needs to be chosen, which shifts the performance of the model to a higher CSI and closer to the diagonal in this diagram. If the probability threshold is set even higher, the frequency bias becomes < 1 and the number of misses surpasses the number of false alarms. Despite their POD rising to 1, models predicting with a lower number of features tend to have a strongly positive frequency bias between 1.3 and 3, predicting more false alarms than misses (Fig. 4.6).

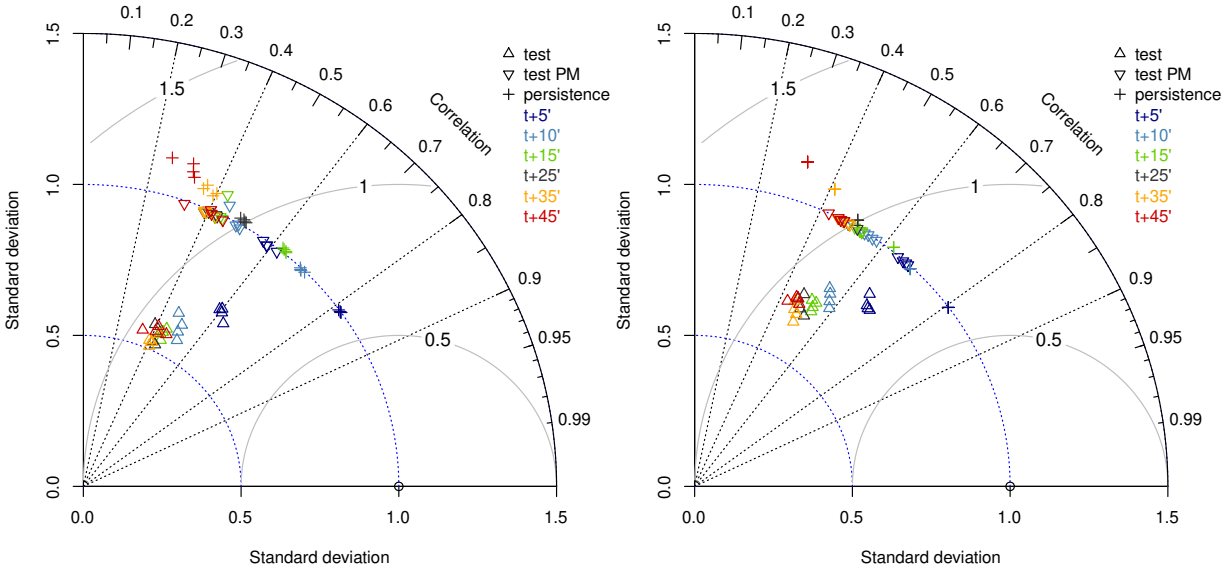

Figure 4.7: Standardized Taylor diagram to evaluate POH predictions at different lead times (colors) comparing the test data set (triangles) and persistence (crosses). Different triangles of the same color are the v1-v4 models for the same lead-time. a) evaluates the combined error of the binary and linear models and b) shows the linear model performance assuming 100 % correct binary predictions.

Figure 4.7 shows the Taylor diagram for the linear models predicting the POH values ≥ 10 %, with and without probability matching. Fig. 4.7b shows what the performance would be if the binary XGBoost model had perfect skill. For visualization's sake, only the v1-v4 models and their corresponding persistences are shown. The reason why the persistence does not have only one result per lead-time in Fig. 4.7a is due to the binary model influencing which target-prediction pair is evaluated in the linear predictions. The following statements are made based on Fig. 4.7a, although similar results can be said for Fig. 4.7b. At all lead-times, the models standard deviations are always 25-50 % smaller than the target standard deviation. Probability matching (PM) removes this difference in standard deviation, however the correlation decreases marginally

and the stcRMSE increases by 15-50 %. For the lead-times t+5', t+10', t+15' and t+25', the persistence clearly suggests more correct POH values than the models. The standardized cRMSE (stcRMSE) of the persistence for these lead-times are 0.6, 0.8, 0.9 and 1. For t+25', the persistence fares slightly better (stcRMSE = 1) than the probability matched predictions (stcRMSE = 1.1). Only for t+35' and t+45', the XGBoost models predict values that are closer to the truth than the persistence (stcRMSE 0.9 or with PM 1.1, vs. the persistence stcRMSE 1.2).

4.7.1.2 Model interpretation of binary models

The following XGBoost model interpretations are conducted for the p1000 binary XGBoost models predicting POH. Despite having 1000 features available, the model for t+5' uses only 500 features (Fig. 4.8a). The remaining features all have mean absolute SHAP values equal to zero. The models for larger lead-times use almost all available features. While the top 1000 features stem primarily from the COSMO model (49 %), radar (21 %) and satellite variables (20 %), all data sources are present (Fig. 4.8a). With increasing lead-time, the fraction of radar variables reduces and the fraction of model variables increases (Fig. 4.8a). For the other data sources, the number of variables per lead-time is close to constant. Fig. 4.8b indicates that radar variables are most relevant for short lead-times (see also Fig. 4.9). Lightning variables constitute only 1.3 % of the top 100 variables. Of these, half are used for t+5' predictions and none are used for t+15' and t+35' predictions. Topographical variables slightly increase in number in the top 100 features from lead-times t+15' to t+45' (Fig. 4.8b).

The XGBoost models use predominantly the most recent measurements (Fig. 4.8c and d). Other time steps of the cell histories are used as well, with the time steps t-20' to t-30' being used the least. The radar-based features are dominant at t_0, and their number decreases the older they are (i.e. from t-5' to t-45'). For satellite and model variables, after initially decreasing from t_0 to t-25', the number of features increases again from t-25' to t-45'. Compared to the top 1000 features, there are 2 times more statistics taken at t_0 vs earlier time steps within the top 100 features (Fig. 4.8c vs. d). The distributions shown in Fig. 4.8c and d look similar if created separately for each lead-time (see Fig. B.4 in Appendix).

The standard deviation provides $\geq$ 20 % of all statistics within the top 1000 features at all lead-times, which is often more than double the frequency of other statistics. The frequency of the remaining statistics fall close to the range of random sampling (8 %), except for the mean, which tends to have a very low count (see Appendix Fig. B.5). Within the top 100 features, short lead-times have strongly varying frequencies of statistics, led by the 75th percentile, the standard deviation and the sum. As the lead-time increases, the standard deviation becomes more frequent than other statistics.
The models are further interpreted using SHAP summary plots, showing SHAP values for the 30 most important features of the p1000 models. Large statistical values of most radar variables at t_0 increase the probability of POH $\geq$ 10 % at t+5' (Fig. 4.9). The SHAP values of the most

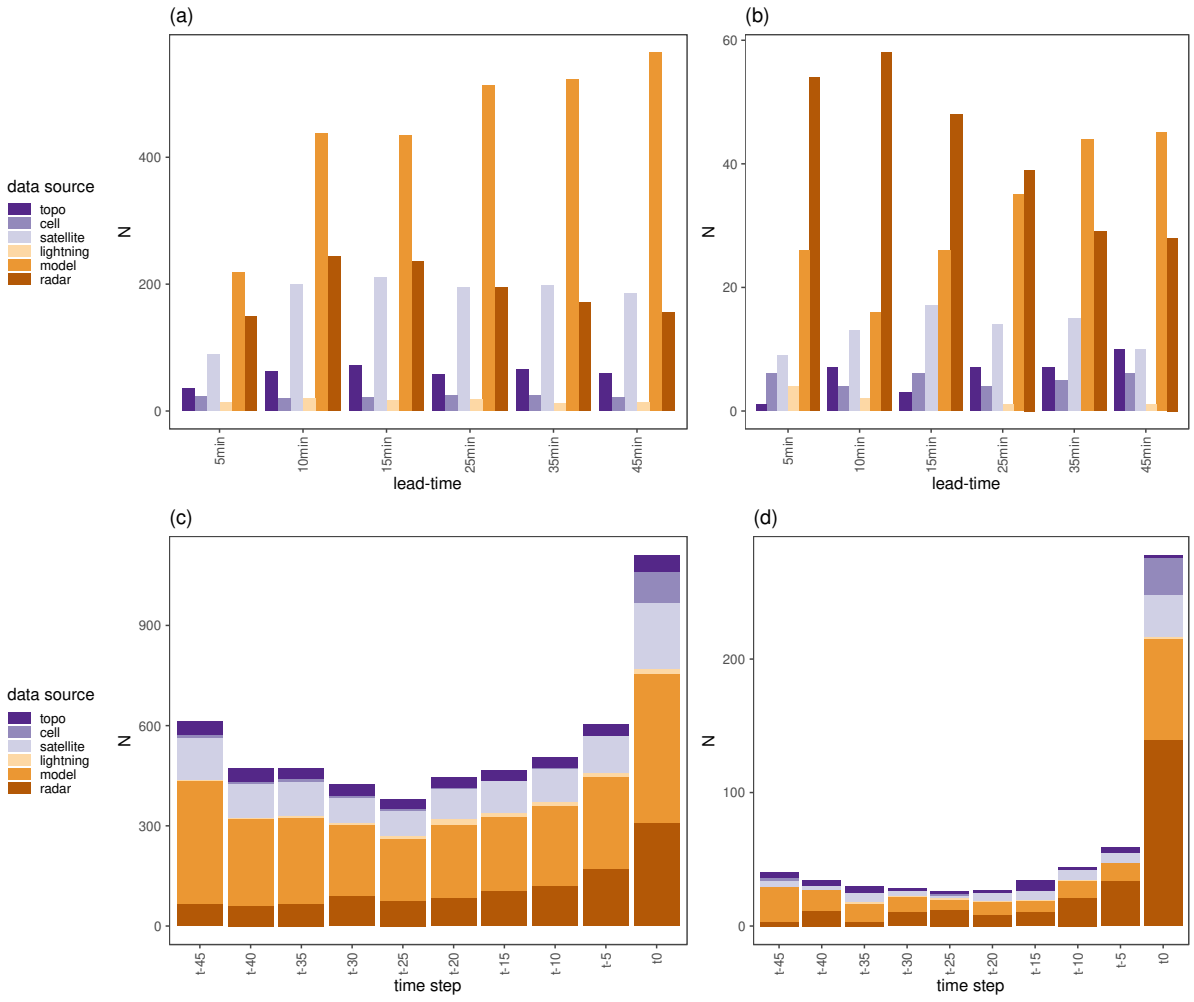

Figure 4.8: For all 1000 features (a and c) and the top 100 features (based on the mean absolute SHAP value; b and d) of all the binary XGBoost models predicting POH with 1000 features (p1000), the number of variables per lead-time (a and b) and time step (c and d) at which the statistics were taken. Colors indicate different data sources (see Table 4.1).

influential feature, the sum of all POH values at t_0 (POH (sum) t0), suggest that the larger the area with high POH values, the more likely the maximum POH will be $\geq$ 10 % at t+5'. These observations suggest that the model recognizes that nowcasting the most recent state of the atmosphere has a high skill, in particular for very short lead times. However, there are also some unintuitive exceptions, e.g. if at t-5' the sum of ET50 values is high, it decreases the probability of POH $\geq$ 10 % (see 16th row "ET50 (sum) t-5'" in Fig. 4.9). For t+10' and t+15', the SHAP summary plots look similar to Fig. 4.9(see Appendix Fig. B.9 and B.10). For these lead-times, top features include other statistics but still mostly radar variables and the mean absolute SHAP values do not decrease as steeply as for t+5'.

At t+45', feature values suggesting an intense (weak) thunderstorm activity at t_0 also increase

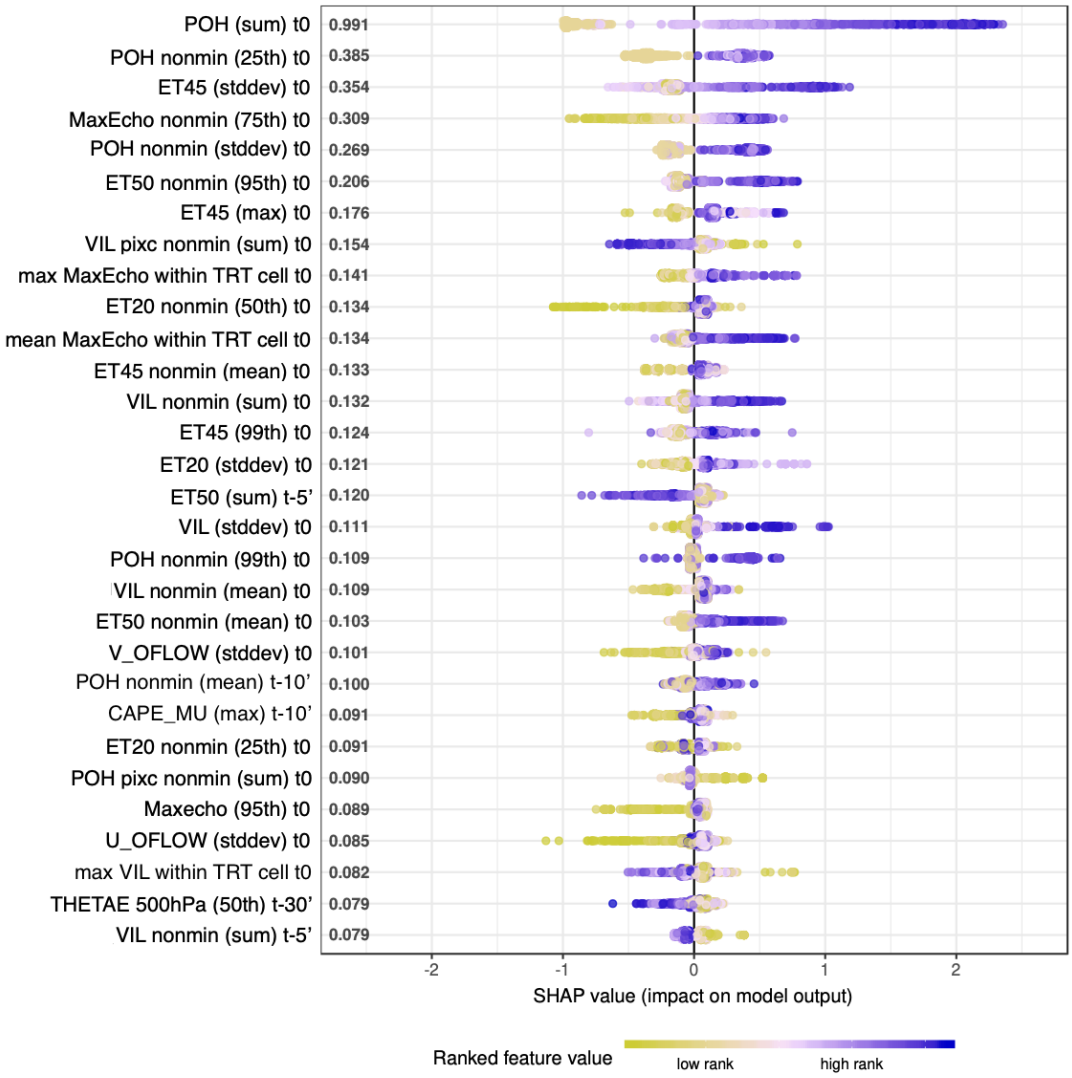

Figure 4.9: SHAP summary plot for the 30 most important features of the binary XGBoost model predicting POH at a lead-time of 5 min using 1000 features. Numbers on the left side in the graph are the mean absolute SHAP values. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). The vertical spread of dots mirrors the probability density function. Negative (positive) SHAP values indicate that the feature has a decreasing (augmenting) effect on the probability that POH $\geq$ 10 %.

(decrease) the probability of POH $\geq$ 10 % (see first five rows in Fig. 4.10). Next to radar variables, the 30 most important features for a t+45' prediction include more higher altitude atmospheric properties (WV_062, PV, CG3, CD2), topographic features (Topo_Aspect, Topo_Altitude) and features describing the vertical temperature profile and energy content of the atmosphere (CAPE_ML, CIN_MU, TD_2m). The SHAP values for PV 300 hPa (min) t-35' (16th row in Fig. 4.10) are positive for very high values and very low PV values. For moderate values, the SHAP values are negative. It is a good example indicating a non-linear influence of a feature on the

predicted value.

Next to radar-based features, several variables appear repeatedly within the top 30 features of XGBoost models predicting POH. These are, for example, the water vapour channel 5 (WV_062; t+15', t+25', t+35', t+45'), the mean and/or maximum MaxEcho within the TRT cell at t0 (all lead-times), the u and/or v components of the optical cell motion (U_OFLOW, V_OFLOW; all lead-times except t+15') and the topographic aspect (t+15', t+25', t+45'; see Fig. 4.9, B.9, B.10, B.11, B.12 and 4.10).

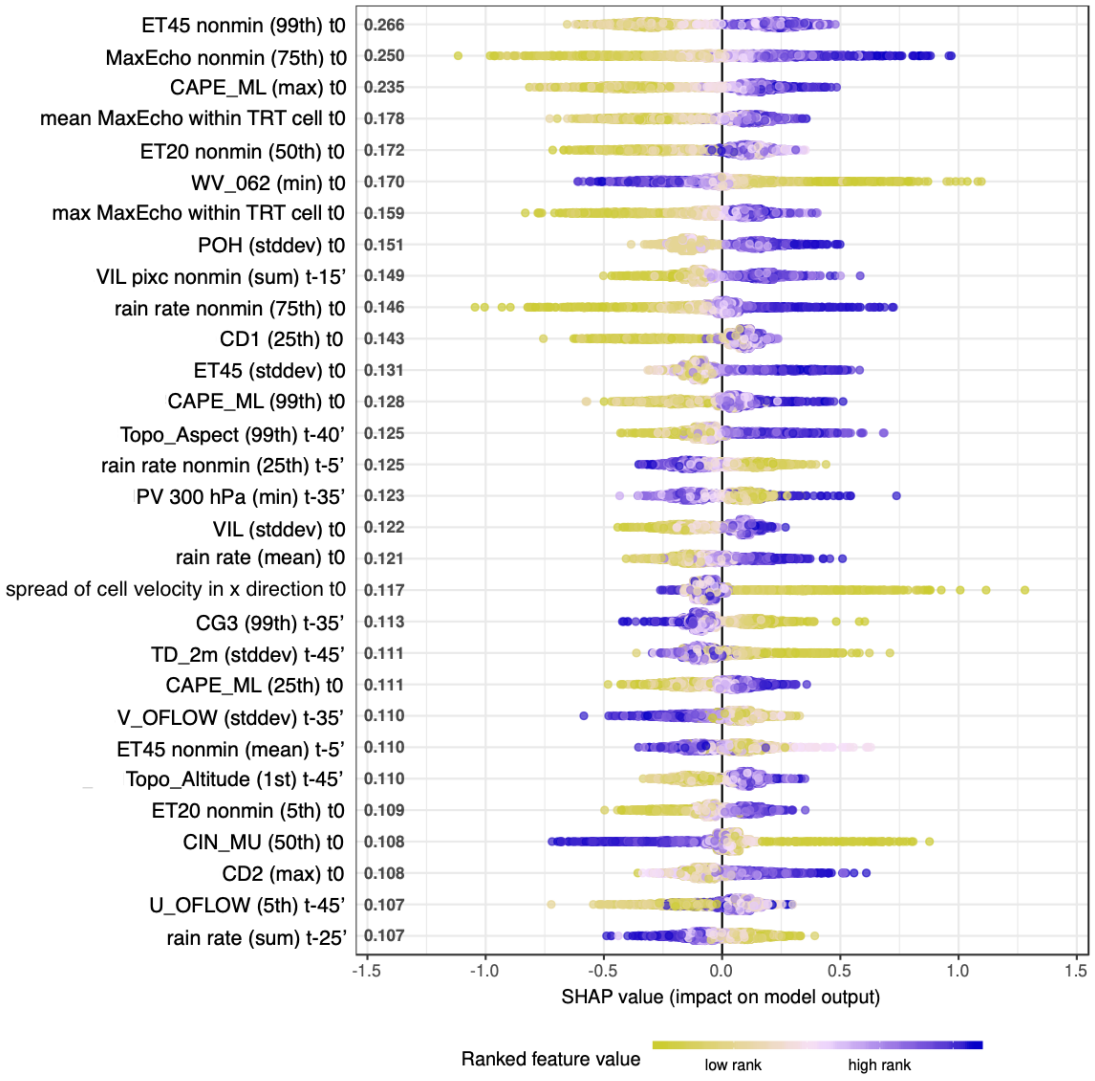

Figure 4.10: SHAP summary plot for the 30 most important features of the binary XGBoost model predicting POH at a lead-time of 45 min using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). See caption of Fig. 4.9 for more details.

4.7.2 Maximum expected severe hail size

4.7.2.1 Model evaluation

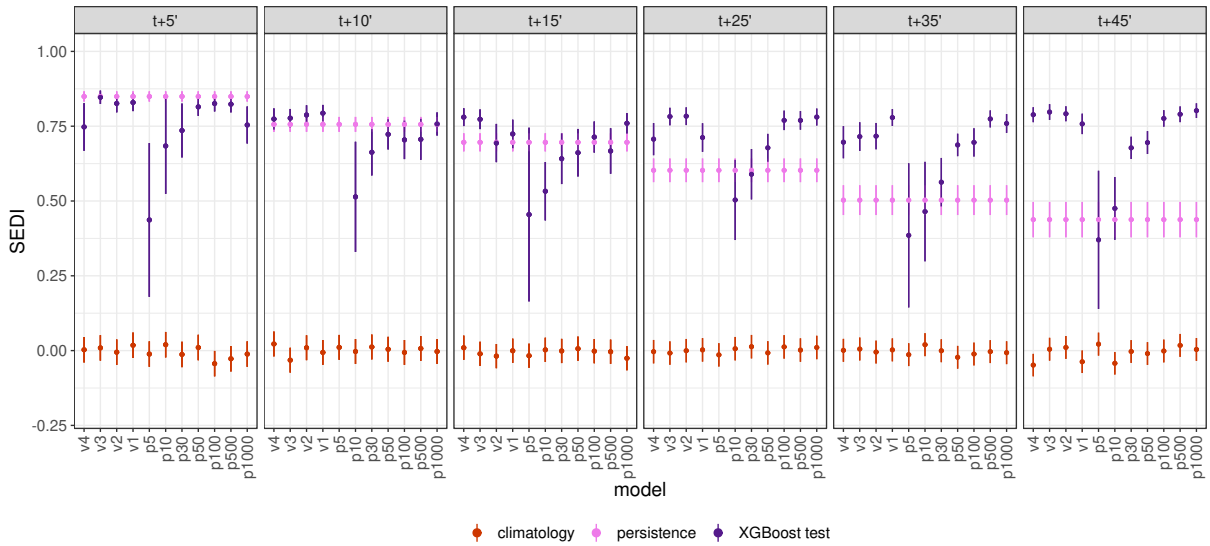

Figure 4.11: SEDI (dots) and their 95 % confidence intervals (vertical bars; see eq. B.4) of MESHS binary predictions for each model version (x-axis) and lead-time, with the XGBoost test dataset (dark purple), the climatology (red) and persistence (light purple).

We observe a decrease in SEDI with increasing lead-time for both binary XGBoost models and persistence predicting MESHS (Fig. 4.11). For t+5', the XGBoost models and the persistence have SEDI values close to 0.8. These values decrease to 0.73 for XGBoost models and to 0.3 for the persistence at t+45'. The decrease in SEDI value of the binary XGBoost models is slower than of the persistence and the best XGBoost model SEDI values for t+35' and t+45' do not statistically significantly differ. The best predictions are made either by the models v1-v4 (t+10', t+25'), or the p100 models (other lead-times). The higher the lead-time, the worse p5 and p10 models fare, compared to models with a higher number of features. Considering computing time optimization being easier with fewer features, an ideal number of features is likely between 50-100 features for the larger than t+5' lead-times. For t+5', depending on the relative importance of the probability of detection and false alarm ratio (see Fig. 4.12) one may prefer using either the XGBoost model or the persistence to nowcast MESHS. For binary MESHS predictions, different hyper-parameter configurations (models v1-v4) do not exhibit statistically significant differences in SEDI.

Figure 4.12 shows the performance diagram for the binary XGBoost models predicting MESHS (triangles v1-v4 and diamonds p5-p1000) as well as the persistence (crosses) for the different lead-times (colors). The XGBoost models predict the occurrence of hail with a CSI of 0.5 for t+5', of 0.4 for t+10' and t+15', 0.38 for t+25' and 0.34 for t+35' and t+45'. The persistence CSI is equal for t+5', 0.38 for t+10' and worse than 0.3 for larger lead-times. The climatology

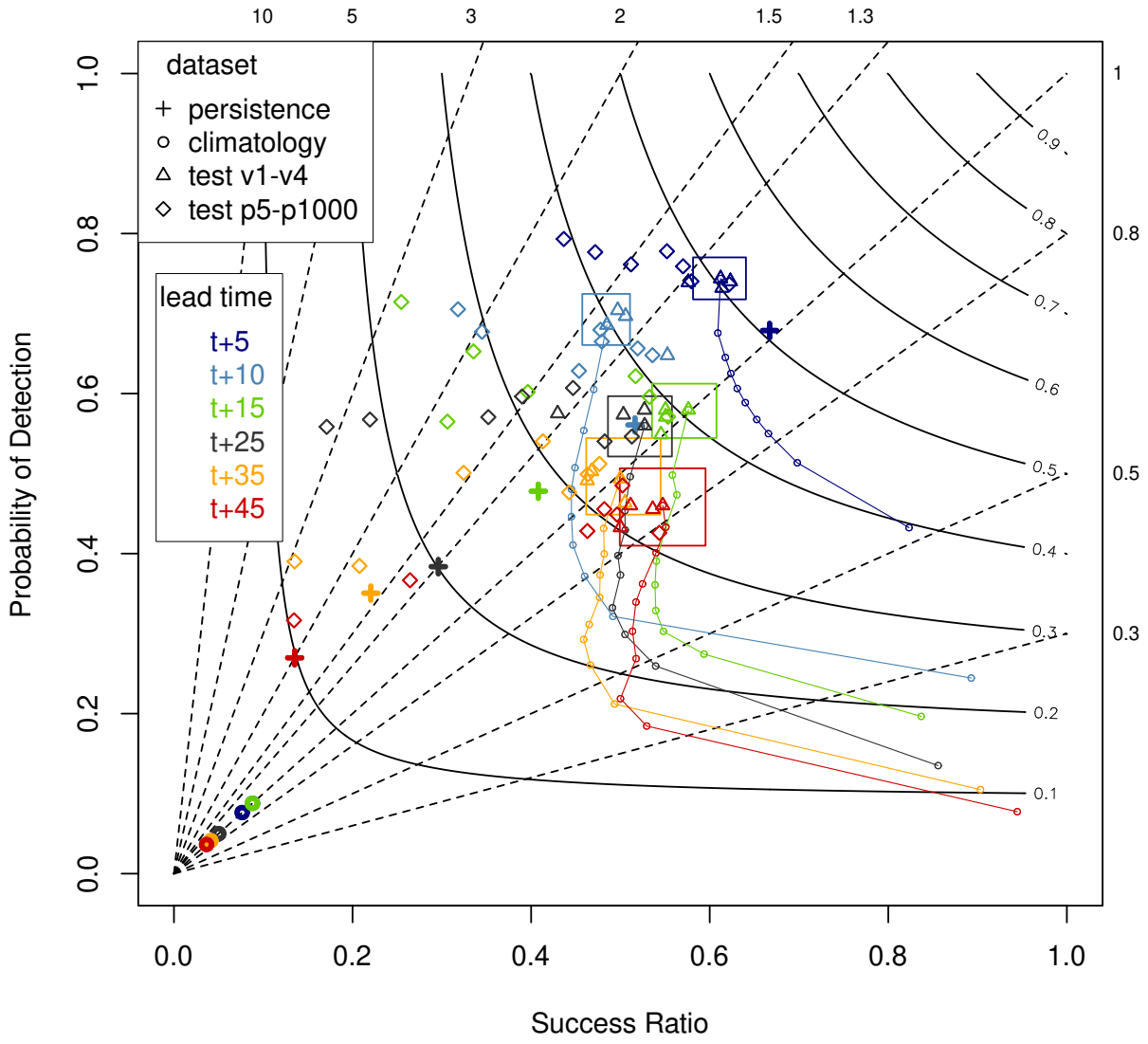

Figure 4.12: Performance diagram for binary predictions of MESHS ≥ 2 cm vs. MESHS = 0 cm at different lead-times (colors) and for the different configurations (crosses = persistence, triangles and diamonds = XGBoost models v1-v4 and p5-p1000, circles = climatology. See caption of Fig. 4.6 for more details.

CSI is smaller than 0.05, likely because the observed non-events in the target dataset outnumber the observed events by a factor > 6 (see Fig. B.7 in the Appendix). If the number of false alarms and the number of misses should be balanced in XGBoost model predictions, the binarization threshold assigning the probabilistic predictions to MESHS ≥ 2 cm should be larger than 0.0 for lead-times up to t+25', and 0.0 for the higher lead-times (lines with dots in Fig. 4.12). The best t+5' model has an equal CSI as the persistence (CSI 0.51). For t+10' and higher lead-times, the best binary XGBoost models always show better CSI values than the persistence. Compared to the 10-minute XGBoost models, the XGBoost models predicting higher lead-times mostly lose skill in POD, while the Success Ratio stays mostly equal.

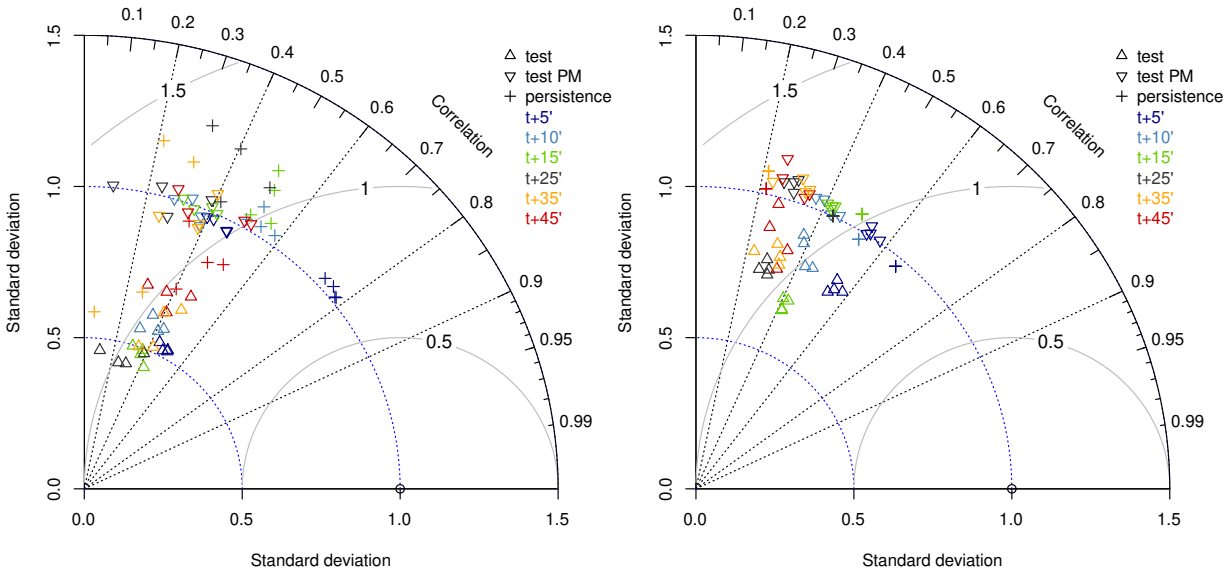

Figure 4.13: Standardized Taylor diagram for linear MESHS predictions (with and without probability matching) for the models v1-v4, compared to the persistence. See caption of Fig. 4.7 for more details.

The difference between Fig. 4.13a and b shows the impact of the MESHS binary model prediction: If the linear model is evaluated after applying the binary model, The MESHS linear models for lead-times up to t+25' predict values with larger stcRMSE than the persistence. For all lead-times XGBoost models predict MESHS values $\geq$ 2 cm with a stcRMSE of 0.8 or greater and correlations between 0.2 and 0.5 (Fig. 4.13a). The stcRMSE of the persistence nowcast for t+5' is 0.7 for the test sample which XGBoost binary models predicted to be $\geq$ 2 cm (Fig. 4.13a) and 0.8 for the test sample containing all non-zero values of the target dataset (Fig. 4.13b). Particularly for large lead-times, the performances of different model versions differ strongly, probably because of the small sample size (e.g., Fig. B.7).

4.7.2.2 Model interpretation of binary models

The data sources in binary XGBoost models predicting MESHS distribute almost identically as for POH (Fig. 4.14). The difference between the numbers of radar-based vs. model-based features per lead-time in the top 100 features is smaller (Fig. 4.14b vs. Fig. 4.8b). Furthermore, the fraction of satellite-based features per lead-time is larger and varies more. For binary predictions of MESHS, t_0 is the most used time step. However, the difference in number of features describing t_0 compared to the other time steps is smaller than for POH. The frequency of statistics in MESHS models is almost equal to POH, with 20 % of the statistics of the top 1000 features being the standard deviation. Compared to POH models, the statistics of the top 100 features of MESHS models vary less and follow a similar distribution as the top 1000 features (Fig. B.8).

According to the SHAP values, the 30 most important features to nowcast binary MESHS at

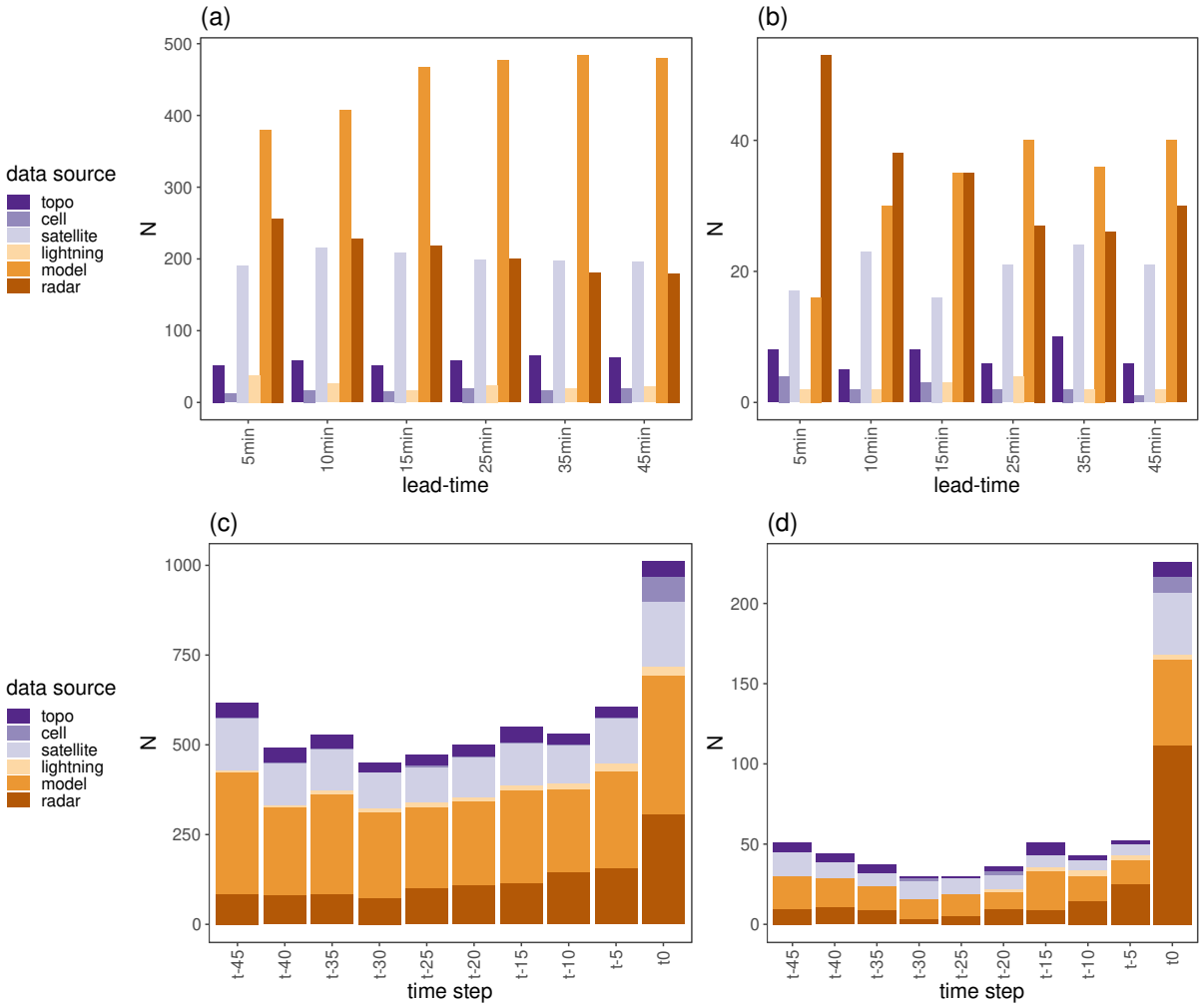

Figure 4.14: For all 1000 features (a and c) and the top 100 features (based on the mean SHAP value; b and d) of all the binary XGBoost models predicting MESHS with 1000 features (p1000), the number of variables per lead-time (a and b) and time step (c and d) at which the statistics were taken. Colors indicate different data sources (see Table 4.1).

a lead-time of 5 minutes are radar variables (22 out of 30 features), mostly taken at t_0 (Fig. 4.15). Satellite-based features are on rank 10 (IR_108), 18 (CG2), 21 (CD2) and 26 (WV_073). Same as POH, the SHAP values for most features support the indication that the present storm intensity is likely to persist. For example, a high maximum POH (POH (max) t_0 in 4th row) and non-minimum mean POH at t_0 (POH nonmin (mean) t_0; 7th row) increase the probability of MESHS $\geq$ 2 cm at t+5'.

The top 30 features for the t+10' XGBoost model using 1000 features are dominated by radar variables, however with more statistics from VIL and less from rain rates than for the model predicting t+5'. Other variables, such as the relative humidity (RELHUM), the LCL and the lightning density (THX_dens) are within the top 30 already at t+10' (Fig. 4.16). As the lead-time rises, the top 30 feature variables vary more in source and time step and the inclination towards

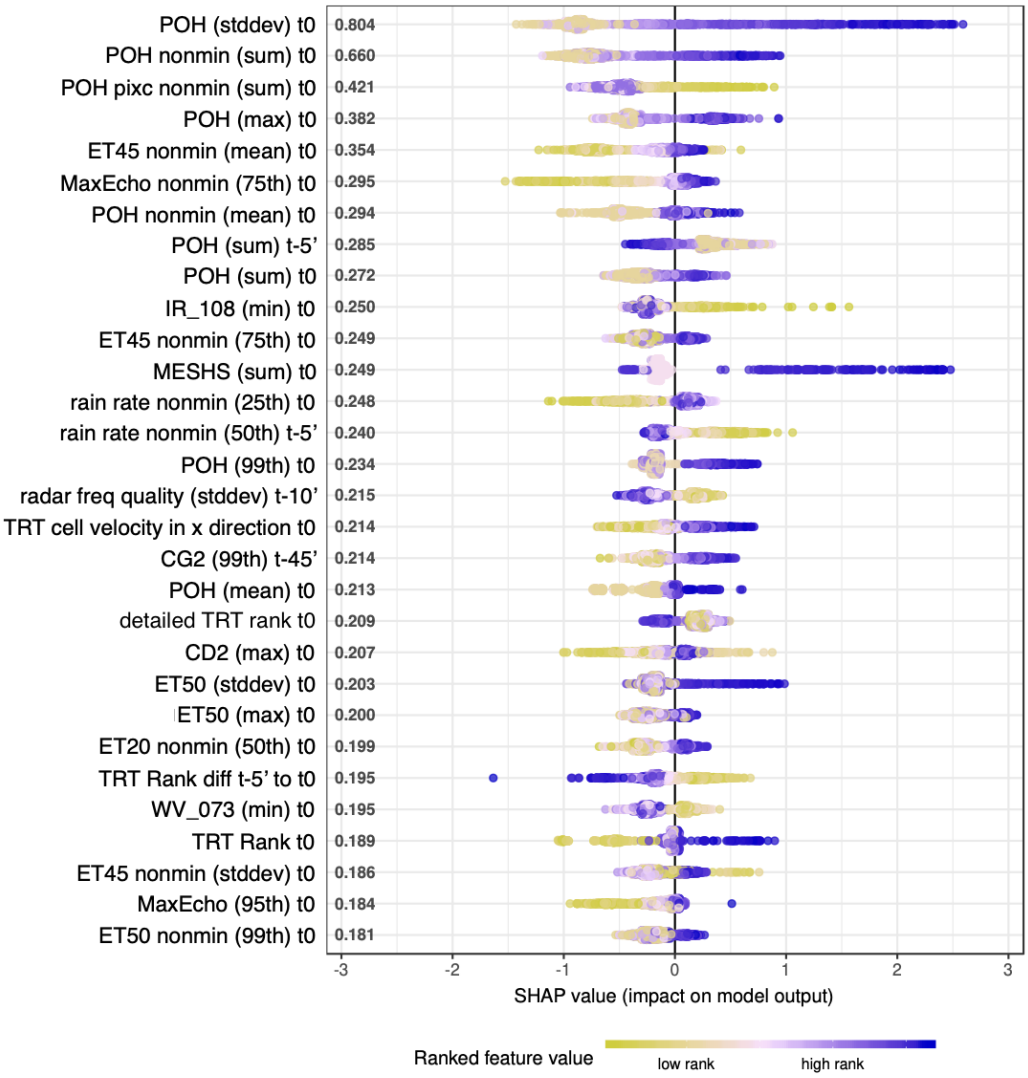

Figure 4.15: SHAP summary plot for the top 30 features of the binary model predicting MESHS for t+5', using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). Negative (positive) SHAP values indicate a decreased (increased) probability of MESHS $\geq$ 2 cm.

being mostly on the positive or negative side of the SHAP = 0 line decreases – the predictions loses in strength of conviction (see Fig. B.13 – B.16 in the Appendix).

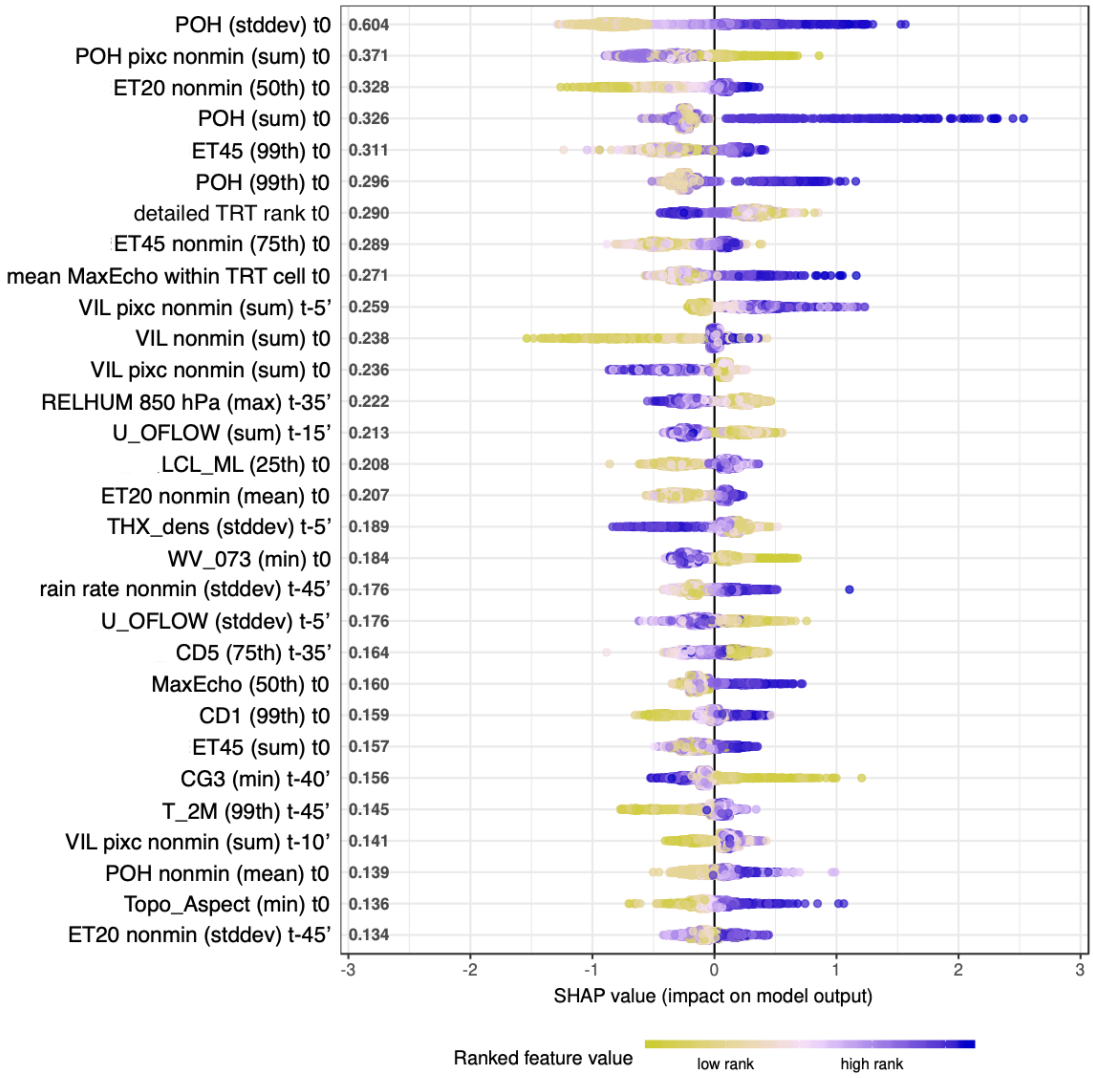

Figure 4.16: SHAP summary plot for the top 30 features of the binary model predicting MESHS at a lead-time of 10 min using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). See caption of Fig. 4.15 for more details.

4.8 Discussion

4.8.1 Discussion of methods

Using all available environmental variables and all 12 statistics of all past time steps between t_0 and t-45' results in 10'610 features. For the presented XGBoost models the results suggest that at least 100 top features were needed to reach a similar performance as models using all features. Even 100 features are quite a large number and are probably a challenge in an operational real-time nowcasting setting, particularly considering that the XGBoost models are (at the moment) incapable of creating a prediction if any feature misses. Furthermore, the large

number of features renders model interpretations more challenging. The SHAP method provides a mean to interpret the XGBoost models, however given the many models, features and possible interactions, it is difficult to gain a comprehensive overview of important model choices. One might prefer to create another ML model that has the task of helping with understanding these models' choices in the future.

The question arises, whether the number of variables and statistics are large enough to capture all interesting thunderstorm environmental processes, despite some thunderstorm environmental variables missing. These missing variables are, for example, the directional wind shear, the cloud top divergence, the storm-based helicity and three-dimensional distribution of water and ice contents in the atmosphere. Furthermore, radars in Switzerland provide dual polarization variables that may give more details on updraft strengths and widths, for example through the differential reflectivity (ZDR) and specific differential phase (KDP) columns. These dual polarization variables lead to a hydrometeor classification (Besic et al., 2016), which may also improve the prediction. At MeteoSwiss, they were not included in nowcasting models yet, because they do not yet exist in a Cartesian format.

The storms were tracked 45 minutes forward and backward starting at each TRT time step. At the earliest position (t-45') before a first TRT position and at the latest position (t+45') after the last TRT time step, the hailstorms might not actually exist. The models do not seem to struggle with that aspect of cell life-time, because the number of false alarms stays relatively low and the false alarm rate does not rise drastically with an increasing lead-time (see e.g., Fig. B.2, B.3, B.6 and B.7).

The past and future cell positions were determined using the average of three optical cell motion vectors close to t_0. This method accounts for short-time changes in direction and speed, but does not detect changes in direction or speed at time steps beyond 15 minutes of t_0. Circle diameters of 17 km, 23 km and 33 km had been tested during the development of the cell motion tracking method. The diameter 23 km was chosen mainly for two reasons. First, cells were not as likely to move out of the extrapolated positions at t-45' and t+45' compared to using a diameter of 17 km. Second, compared to 33 km, the statistics would not be influenced by neighboring cells as much.

The Coalition-2 algorithms locating thunderstorms in satellite images could validated the backward extrapolation method beyond TRT. Even though the extrapolation method solves the problem of thunderstorm cells splitting and merging, knowing about previous splits and merges might have some predictive value.

In this project, the top features going into the p5-p1000 models were determined with the v1 model, despite the test results suggesting that v1 models do not always yield the best performances compared to v2-v4 (see Fig. 4.5 and 4.11). It would have been more elegant to apply

the feature reduction analysis on the best v1-v4 model available. Other solutions deal with strongly skewed target variables. Instead of creating two separate models, predictions could be post processed for example with an isotonic regression (Niculescu-Mizil and Caruana, 2005; see e.g., Lagerquist et al., 2017; McGovern et al., 2019a; Burke et al., 2020).

4.8.2 Discussion of results

Compared to other studies predicting extreme weather events with samples sizes ranging between 10^5 to 10^6 (e.g., Flora et al., 2020; Lagerquist et al., 2017; Burke et al., 2020),a sample size of 3 x 10^4 seems relatively small. While for binary models, the different model versions v1-v4 and p5-p1000 suggest that the sample size was sufficient, the linear XGBoost models suffered from the small sample size. Splitting the task of predicting the targets into two (as a reaction to the skew of the target variables), with one binary model being followed by a linear model, was, however, a relative success.

Lead-time steps of 5 minutes represent a relatively high temporal resolution. The model quality is quite high, considering that no temporal error was allowed in the evaluation. The likely reason why, compared to the climatology, the MESHS binary models performed better than the POH binary model, is the distribution of the target variable. POH is calculated using the ET45. MESHS uses ET50, a more extreme reflectivity value. Furthermore, MESHS $\geq$ 2 cm has been likened to POH $>$ 80 % (Nisi et al., 2016), a much higher threshold than POH $\geq$ 10 %. POH models thus predict events that occur more frequently. For POH, the number of observed events (hits+misses) is larger than the number of observed non-events (correct rejections+false alarms), except at t+45' (see Fig. B.3). For MESHS and all lead-times, the number of observed non-events outnumber the observed events by a factor greater than 6 (Fig. B.7). This imbalance in number of events vs. non-events affects the effect of the random perturbation of the target dataset, through which the climatological prediction is defined. The much smaller number of observed events in the MESHS dataset strongly reduces the number of samples used to train and evaluate linear XGBoost models predicting MESHS. I hypothesize that these could be improved by increasing the size of the data set.

To which degree these models are applicable on samples from other years has yet to be tested. The year 2018 was characterized by thunderstorms that had quite atypical tracks; they did not as often travel across the Swiss plateau in long, straight tracks from the south-west to the north-east in the same way as, in other years (see Schroeer et al., 2019). Therefore, nowcasting models that were trained on samples from different years may result in different model performances and different SHAP feature rankings.

4.9 Summary and Conclusions

In this project, XGBoost models predicted the maximum probability and size of hail (POH and MESHS), every 5 minutes up to 45 minutes in advance (excluding t+20', t+30' and t+40'), us-

ing > 10'000 statistics of environmental and thunderstorm variables. These variables stem from radar, COSMO-1, satellite, lightning and stationary data sets such as topographical information. Cell histories are extracted, starting at thunderstorm positions determined by the TRT algoritm, up to 45 minutes before the time of the latest observation. Thunderstorm cells are detected with a temporal resolution of 5 min, which is the operational resolution of the Swiss radar network. The past and future thunderstorm positions are estimated using vectors of optical cell motion. Around each past and future position, circles with a diameter of 23 km define the area in which 12 different statistics of features (input variables) are calculated. Two target variables were defined, the maximum POH and the maximum MESHS within the 23 km circles. For each target variable, two types of models were tuned and trained. The two models are applied consecutively. The two binary XGBoost models predict the probability of POH $\geq$ 10 % and MESHS $\geq$ 2 cm. The two linear XGBoost models predicts the maximum value of POH and MESHS (POH linear, MESHS linear) once the binary models have determined the presence of POH $\geq$ 10 % and MESHS $\geq$ 2cm.

To better understand the effect of hyper-parameter tuning, the four best hyper-parameter configurations are retrained to yield the binary and linear models v1-v4 for each lead-time. Furthermore, a sensitivity study on the necessary number of features per model is conducted by retuning and retraining models, which use only the top 5, 10, 30, 50, 100, 500 and 1000 features of the v1 model. Finally, the effect of input variables on the nowcasting result was discussed using the SHAP method, providing some insight into the choices made by the XGBoost models.

Results show that the binary POH models successfully predict the occurrence of hail with a CSI of 0.75 for t+5' and 0.58 for t+45' (Fig. 4.6). For t+5', the performance is as good as a Lagrangian persistence nowcast. For large lead-times XGBoost models fare better (at t+45' CSI of persistence is 0.35). The linear XGBoost models for POH predict actual values with a greater standardized centered RMSE (stcRMSE; RMSE that is debiased and standardized by the target standard deviation) than the persistence up to t+25'. For t+35' and t+45', the XGBoost models provide a better nowcasts, although with correlation coefficients of 0.4 and stcRMSE's close to 1.

MESHS binary models predict the occurrence of maximum MESHS $\geq$ 2 cm with a CSI of 0.5 for t+5' and 0.35 for t+45' (Fig. 4.12). Compared to POH, binary XGBoost models for MESHS perform much better at larger lead-times than the persistence and climatology (persistence and climatology CSI at t+45' are 0.1 and 0.05). The reason is that in the target data set, the data set contains six times more MESHS = 0 cases than MESHS $\geq$ 2 cm cases. Because of that imbalance, the sample used to train and test the MESHS linear models was small and both the persistence and XGBoost models yielded relatively poor predictions (Fig. 4.13). Furthermore, it is more challenging to predict an event that occurs even more infrequently.

According to SHAP values, the most important features to predict POH and MESHS are radar-based, followed by COSMO-1-based and satellite-based features. As the lead-time increases, the number of radar-based features in the top 100 features decreases and the number of COSMO-

1-based features increase. However, all data sources are present within the top 100 features for binary XGBoost models predicting POH and MESHS. Although the statistics are taken predominantly at the latest observation time, the other time steps between t-5' and t-45' are used too (Fig. 4.8 and Fig. 4.14). Approximately 100–500 top features are necessary to reach the same performance as the models using all features for both binary XGBoost models predicting POH and MESHS (Fig. 4.5, 4.11). The larger the lead-time, the more important is a larger number of features. The standard deviation is the most frequently used (20 %) of all 12 statistics within the top 1000 and top 100 features.

SHAP summary plots (Fig. 4.9, 4.10, 4.15, 4.16; see also Appendix section B.6) suggest that the models recognize the advantage of persistently predicting the latest observed state. Feature values characterizing intense hailstorm activity at the latest observation time increase the probability of POH $\geq$ 10 % and MESHS $\geq$ 2 cm at all lead-times. However, XGBoost models are capable of recognizing other patterns in features, which provide a better prediction than the Lagrangian persistence.

4.10 Outlook

During the implementation and writing of this project, many open questions and ideas for future research ventures emerged.
In future projects, hail nowcasting using machine learning could be developed and improved as follows:

- Extend the full dataset to include data from other years
- Add to the features list dual-polarization radar variables, including width and height of ZDR and KDP columns. von Matt (2020) has found that ZDR-Columns are very likely to increase the performance of hail nowcasting models in Switzerland. Furthermore, the width of thunderstorm updrafts was found to have a strong connection to the amount of hail growth (Nelson, 1983 and Foote, 1984 in Gagne et al., 2019).
- In this project, the models created with a smaller number of features took the top 5, 10, etc., 1000 features as ranked by the v1 model. Another solution could first create the p1000 model with the top 1000 features of v1. The p1000 model may rank these 1000 features differently than the v1 model. The top 500 features going into the p500 model would then be taken from the ranked features of the p1000 model. This iterative process could be repeated to produce the remaining p-models.
- Tune the XGBoost hyper-parameters alpha, lambda and gamma (see section 4.6.2).
- Use SHAP values, instead of the gain, to determine the rank of features, determining which features go into the p5-p1000 models
- Create spatial maps of model performance (e.g. Hill et al., 2020)

- Analyze the SHAP values in depth for seasonal and diurnal cycles, create spatial maps of SHAP values to determine regional differences in feature importance and influence, explore SHAP values by groups of similar features and observe interactions between features. These analyses could provide more detailed information on hail predictability and help understand strength and weaknesses of different features.
- Use for example generative adversarial networks (Goodfellow et al., 2014) to create thunderstorms that produce the largest hail for the longest duration as case studies. The case studies could evolve into story lines to help simulate the hail risk.
- Test models that only use features taken at the latest observation time. If the model performance were sufficient, we would not need to determine past positions to nowcast hail.
- Try using predicted target variables as input features for longer lead-times

Further ideas:

- Include more features describing the atmosphere at varying altitudes: Gagne et al. (2019) demonstrated that incorporating both vertical profiles and spatial information into a deep learning hail size diagnostic model could provide both increased hail size analysis skill and insight into important factors for hail growth. Idealized modeling studies of supercells found that small changes in moisture and wind profile could alter storm morphology and hail growth(Gagne et al., 2019).
- Move towards seamless forecasting by combining the nowcasting result with extended NWP model prediction e.g. from the module hailcast and extending the nowcast/forecast lead-time to one or several hours. An example of an ongoing larger project aiming at developing a seamless prediction system is the SINFONY project by the German Weather Service (see DWD, 2021 and Blahak et al., 2018).

The MeteoSwiss Coalition-4 project is investigating these further ideas:

- Adapt the available input variable list to the next generation of geostationary satellites, which are equipped with advanced imagers with increase spatial, spectral and temporal resolution. Furthermore, space born lightning imagers could complement ground based lightning observation or provide lightning observations for remote locations. The space born hyperspectral sounder could contribute to improving the accuracy of temperature and humidity profiles and the derived atmospheric instability.
- Extent the prediction to other thunderstorm specific hazards, such as heavy precipitation, lightning, and wind gusts.
- Deep learning algorithms could be applied for the nowcasting of hail. Convolutional neural networks accept images as input and could therefore interpret spatial structure. Recurrent neural networks are promising for the interpretation of temporal developments. Generative Adversarial Networks are capable of creating realistic ensembles with coherent spatial and temporal statistics.

4.11 Acknowledgements

I would like to thank Ulrich Hamann and Joël Zeder for developing and implementing the data retrieval and preprocessing methods. Loris Foresti and Daniele Nerini provided and explained the Pysteps algorithm. Many thanks also go to Ulrich Hamann, Alessandro Hering and Urs Germann, with whom I discussed initial methods and preliminary results during my stay at MeteoSwiss Locarno-Monti. I am extremely grateful to Olivia Martius and Uli Hamann for reading this manuscript and for their invaluable comments and questions. Finally, yet importantly, many thanks go to the R community for providing R packages, with which I could create, verify and analyze the machine learning models.

Chapter 5

Multi-day hail clusters and isolated hail days in Switzerland – large-scale flow conditions and precursors

This chapter contains a manuscript that has been written together with Olivia Martius, Alessandro Hering, Luca Nisi, Katharina Schroeer and Urs Germann. The manuscript has been submitted under the title "Multi-day hail events and isolated hail days in Switzerland – large-scale flow conditions and precursors" in Weather and Climate Dynamics (Barras et al., 2021 (in review)).

5.1 Abstract

In Switzerland, hail regularly occurs in multi-day hail clusters. The atmospheric conditions prior to and during multi-day hail clusters are described and contrasted to the conditions prior to and during isolated hail days. The analysis focuses on hail days that occurred between April and September 2002—2019 within 140 km of the Swiss radar network. Hail days north and south of the Alps are defined using a minimum area threshold of a radar-based hail product. Multi-day clusters are defined as 5-day windows containing 4 or 5 hail days and isolated hail days as 5-day windows containing a single hail day. The reanalysis ERA-5 is used to study the large-scale flow in combination with objectively identified cold fronts, atmospheric blocking events, and a weather type classification. Both north and south of the Alps, isolated hail days have frequency maxima in May and August-September whereas clustered hail days occur mostly in July and August. Composites of atmospheric variables indicate a more stationary and meridionally amplified atmospheric flow both north and south of the Alps during multi-day hail clusters. On clustered hail days north of the Alps, blocks are more frequent over the North Sea, and surface fronts are located farther from Switzerland than on isolated hail days. Clustered hail days north of the Alps are also characterized by significantly higher convective available potential energy (CAPE) values, warmer daily maximum surface temperatures, and higher atmospheric moisture content than isolated hail days. Hence, both stationary flow conditions and anomalous amounts of moisture are necessary for multi-day hail clusters on the north side. In contrast, differences in

CAPE on the south side between clustered hail days and isolated hail days are small. The mean sea level pressure south of the Alps is significantly deeper, the maximum temperature is colder, and local moisture is significantly lower on isolated hail days. Both north and south of the Alps, the upper-level atmospheric flow over the eastern Atlantic is meridionally more amplified three days prior to clustered hail days than prior to isolated days. Moreover, blocking occurs prior to more than 10 % of clustered hail days over Scandinavia, but no blocks occur prior to isolated hail days. Half of the clustered hail days south of the Alps are also clustered north of the Alps. On hail days clustering only south of the Alps, fronts are more frequently located on the Alpine ridge, and local low- level winds are stronger. The temporal clustering of hail days is coupled to specific synoptic- and local- scale flow conditions, this information may be exploited for short to medium-range forecasts of hail in Switzerland.

5.2 Introduction

In Switzerland, hail days can occur several days in a row. Such multi-day clusters of hail days can cause substantial damage in a short time. Multi-day clusters of severe weather and associated high impacts have also been reported from North America (Shafer, C., Doswell III, 2012; Trapp, 2014; Schroder and Elsner, 2020; Gensini et al., 2019). Although the atmospheric conditions associated with hail in Switzerland and central Europe are well studied (e.g., Huntrieser et al., 1997; Madonna et al., 2018; Taszarek et al., 2017; Púčik et al., 2015; Brooks, 2009; Púčik et al., 01 Nov. 2019; Kunz et al., 2020), little is known about the large-scale weather conditions that lead to multi-day clusters of hail days. Such multi-day clusters are likely the result of the extended longevity or repeated re-establishment of particular features of weather situations over Europe. Addressing this research gap is relevant for insurance and forecasting applications. For insurance companies, an important question is whether hail events can be considered as independent or not. For forecasting, it is relevant to know whether processes and weather situations leading to isolated hail events and multi-day clusters of hail events differ substantially from each other.

Large-scale flow patterns have been linked to the spatial and annual variability of thunderstorms in Europe (Piper et al., 2019; Mohr et al., 2019), and two case studies highlight two particularly long-lasting sequences of consecutive thunderstorm days in central Europe (Piper et al., 2016; Mohr et al., 2020). Piper et al. (2016) compare a 15-day episode of thunderstorms in Germany in May–June 2016 with the period 1960–2014 and find that this event was exceptional for its number of days with prevailing extreme precipitation or convection-favoring conditions. Mohr et al. (2020) investigate a series of severe thunderstorms in May–June 2018 in central Europe and find a blocking anticyclone that trapped moist and warm air over western and central Europe and several cut-offs on the block's southern fringe to provide exceptionally persistent low-stability conditions. Madonna et al. (2018) compared large-scale conditions during June 2006, a month with above-average hail days (12, average 9.2) with June 2004, where only 2 hail days occurred in northern Switzerland. June 2006 saw warmer surface temperatures over most of Europe, higher CAPE values, more moisture, and more unstable local conditions than the climatology. During

June 2004, reanalysis data indicates increased blocking frequency south and west of Greenland, less moisture, and more frequent lows and fronts in the Alpine region and north of the Alps. Whereas these investigations have studied individual cases in detail, an analysis of the synoptic and large-scale conditions during and leading to multi-day hail clusters in Switzerland has yet to be conducted.

The first objective of this study is to quantify the occurrence of multi-day hail clusters in Switzerland and northern Italy in the period from 2002–2019. The second objective is to identify the main features of large-scale circulation over Europe during and prior to multi-day hail clusters and contrast these with those of the circulation on isolated hail days.

More specifically we aim to answer two questions:

- Which atmospheric conditions are associated with and differentiate multi-day hail clusters and isolated hail days in Switzerland north and south of the Alps during 2002—2019?
- Which atmospheric conditions occur on days before multi-day and isolated hail events?

The paper is structured as follows: Sect. 5.3 presents the data we use in this study. Sect. 5.4 focuses on the methods: how we defined clustered and isolated hail days and the method for determining the statistical significance of the difference between composites. Sect. 5.5 describes the results, which are discussed and summarized in Sect. 5.6. The article ends with the conclusions and outlook in Sect. 5.7.

5.3 Data

5.3.1 Probability of hail (POH)

This study uses the radar- and model-based probability of hail product (POH, Foote et al., 2005a based on Waldvogel et al., 1979) to identify hail days between April and September 2002—2019 in the Swiss radar domain. POH is an operational product that indicates the likelihood of hail at the ground (zero to 100 %) on a 1 x 1 km Cartesian grid, with radar hail data quality generally assumed to be highest within a 160 km radius around the five Swiss weather radar stations (Nisi et al., 2016). Car insurance loss data has verified a threshold of POH $\geq$ 80 % to indicate the presence of hail locally (Nisi et al., 2016; Madonna et al., 2018). The extent of the daily area of POH $\geq$ 80 % is extracted separately for a domain north of the Alps (Fig. 1, blue area) and another south of the Alps (Fig. 5.1 green area). The domain south of the main Alpine ridge contains southern Switzerland and a region of Northern Italy within a 140 km radius of the weather radar stations (Fig. 5.1).

5.3.2 Car insurance loss reports

Area thresholds for the identification of hail days are defined with hail-related car insurance loss reports provided by the Swiss Mobiliar insurance company. The insurance loss reports are

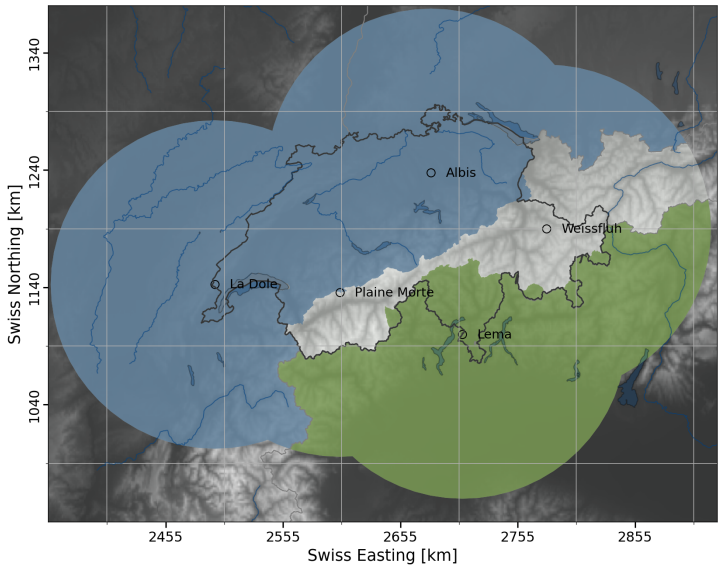

Figure 5.1: The investigation areas south of the Alps (green) and north of the Alps (blue) overlaid on a topographical map (gray shading). Light grey shading indicates the area within a 140 km radius of the five Swiss weather radar stations with the highest radar quality (Nisi et al., 2016). Switzerland is centered at 46.8 °N 8.2 °E.

available for the years 2003—2012 and are described in detail in Morel (2014) and Nisi et al. (2016). Morel (2014) shows that some car insurance loss dates had to be corrected because of human error. To increase the robustness of the car insurance loss information, we consider only days with at least five car insurance loss reports.

5.3.3 Weather Type Classification

This study uses an automatic daily weather type classification (WTC) of the synoptic situation over Central Europe (Weusthoff, 2011). The WTC has ten classes: eight classes for the eight main wind directions and two classes for low- and high-pressure situations based on the geopotential height at 500 hPa. The wind and geopotential data are taken from the ERA-Interim reanalysis.

5.3.4 Reanalyses

The two reanalysis data sets used in this study (ERA-Interim, see Dee et al., 2011, and ERA-5, see Hersbach et al., 2020) are produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). Our analysis considers the period April 2002–September 2019. We extracted ERA-5 variables at a 6-hourly temporal resolution and a spatial resolution of 0.5°. The large-scale dynamics are described through the Ertel potential vorticity (PV; in PV units (PVU)), calculated and interpolated to the 335 K isentrope, and horizontal wind components at 250 hPa (in m s^{-1}). Daily atmospheric blocking events were calculated as in Rohrer et al. (2018), following the algorithm developed by Schwierz et al. (2004). Synoptic and local conditions are represented

by low-level winds (at 850hPa, in m s^{-1}), the daily maximum surface temperature (T2M, in degrees Celsius), the daily maximum convective available potential energy (CAPE, in J kg^{-1}) and the daily mean sea-level pressure (MSLP, in hPa). The daily statistics for these last three variables are calculated from hourly values. Bulk wind shear values are obtained by subtracting the horizontal wind components at 850 hPa from the wind at 500 hPa. The total precipitable water (TPW, in mm) provides information on the moisture content of the atmosphere. The front data stem from the ERA-Interim reanalysis that has been interpolated to a spatial grid of 1°and has a temporal resolution of 6 hours (see Schemm et al., 2015, for details). These fronts have a minimum gradient of equivalent potential temperature of at least 4 K per 100 km at 850 hPa and a minimum length of 500 km. The composites show the percentage of all time steps with fronts.

5.4 Methods

5.4.1 Definition of hail days

We identify hail days by denoting the area where POH equals or exceeds 80 % during a day as the daily POH footprint. To determine hail days, we need to define a minimum footprint area. This is because, despite rigorous data quality control in the Swiss operational radar data processing, some data points still have residual radar artefacts not related to hail. The number of data points affected is small considering the amount of ground clutter in the raw radar data for an Alpine country, but we have to take them into account when identifying hail days using daily POH footprints. We tested minimum footprint area thresholds between the 70th and the 95th percentile of the daily footprint area distribution in the northern and southern domains. We found that the 80th percentile of the area distribution is best suited to identifying hail days. This threshold best corresponds to days with car damage reported across Switzerland over 2003–2012 (Table C.1 in the Appendix). If we use the 80th percentile to define hail days, most days with $\geq$ 5 car insurance losses occur on hail days, and the number of days with $\geq$ 5 car insurance losses occurring on nonhail days are minimized. As a result, we define a hail day as a day with a footprint greater than the 80th percentile of the POH $\geq$ 80 % area distribution. This corresponds to a daily maximum POH $\geq$ 80 % over an area greater than 580 km^2 in the northern domain and greater than 499 km^2 in the southern domain.

This definition produces an average of 26 hail days per hail season north of the Alps; a minimum of 16 hail days occurred in 2014 and a maximum of 43 hail days in 2009. South of the Alps, it produces an average of 25 hail days per hail season; a minimum of 15 hail days occurred in 2004 and 2007 and a maximum of 38 hail days in 2019.

5.4.2 Selection of serially clustered versus isolated hail days

To define the clustered hail periods and isolated hail days, we use a counting approach similar to Pinto et al. (2014) and Kopp et al. (2021). In a period of 5 days we require multi-day clustered hail periods to have at least 4 hail days and isolated hail periods to have only 1 hail day. To

ensure independence, all isolated hail days must have a period of at least 3 nonhail days to the next hail day.

All hail days in 2002–2019 and their assignment to the clustered or isolated hail day category are shown in Fig. 5.2 and counted in Table 2.1. In total, 308 hail days are identified north of the Alps and 294 hail days south of the Alps. North of the Alps, the period with clustered hail days starts on day of the year (DOY) 129 (mid-May) and ends on DOY 238 (end of August). South of the Alps, the period with clustered hail days starts a month later on DOY 160 (mid-June) and ends on DOY 230 (mid-August). Isolated hail days occur earlier and later in the season than clustered hail days (Fig. 5.2). To avoid seasonality effects, all isolated hail days outside the seasonal range of the clustered hail days are excluded, leaving 69 isolated hail days north of the Alps and 42 isolated hail days south of the Alps. For significance testing, we separate the clustered hail days into independent clustering periods of 5 days. We define independent periods as 5-day periods separated by at least 2 days. In addition, series of hail days that cluster for more than 11 days, for example in 2003, are split into independent 5-day clustering periods that each contain at least 4 hail days. This results in 32 independent 5-day periods with a total of 135 clustered hail days north of the Alps and 21 of these 5-day periods with 89 hail days south of the Alps. About half of all clustered hail days south of the Alps are also clustered hail days north of the Alps (Fig.5.2). In summary, this article analyzes 204 (135+69) of a total of 308 hail days north of the Alps and 131 (89+42) of 294 hail days south of the Alps (Table 5.1).

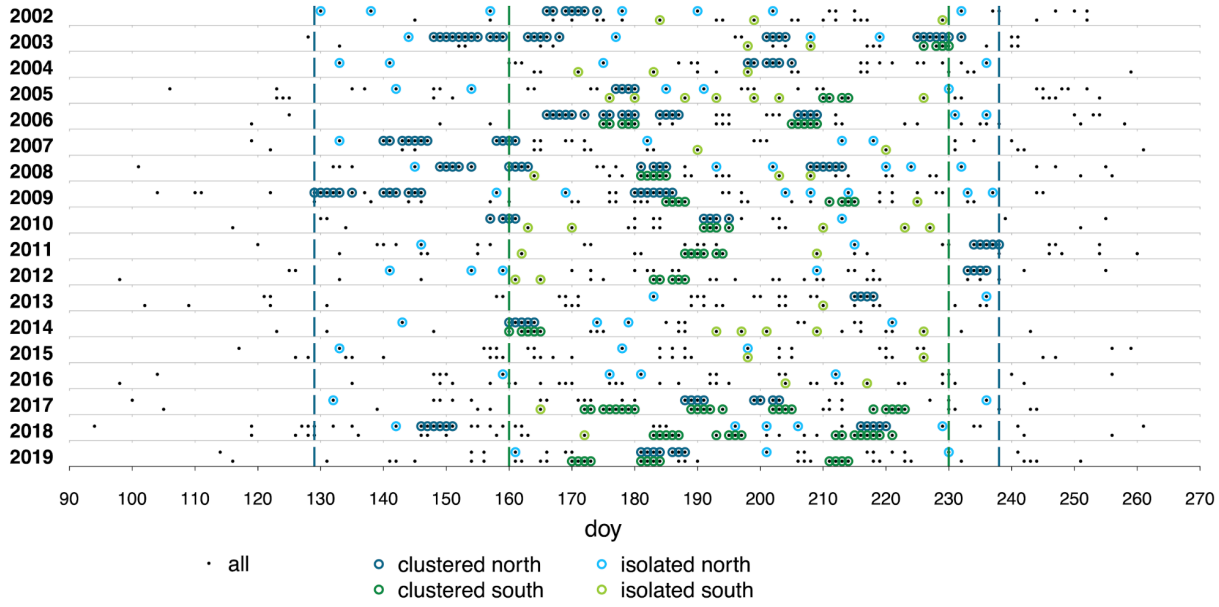

Figure 5.2: All hail days for each year between 2002 and 2019, north of the Alps (top rows, blue colors) and south of the Alps (bottom rows, green colors); the clustered hail days have dark circles, and the isolated hail days have light colored circles. Note that isolated hail days are only considered in the analysis if they occur no earlier or later in the year than the earliest or latest clustered hail day (marked by vertical dashed lines).

Table 5.1: Number of hail days north and south of the Alps and the number of days leading to them. The bold columns indicate how many days enter the composites of the large-scale flow.

		clustered events			isolated events		
	total number of hail days	clustered hail days	**days before clustered hail days (d-1, d-2, d-3)**	**clustered hail days in independent 5-day periods**	isolated hail days	**days before isolated hail days (d-1, d-2, d-3)**	**isolated hail days in the DOY range of clustered hail days**
North of the Alps	308	164	**31**	**135**	96	**69**	**69**
South of the Alps	294	102	**21**	**89**	99	**42**	**42**

5.4.3 Composites of large-scale flow

Differences in large-scale conditions during clustered and isolated hail days are assessed using composites of the large-scale flow from reanalysis data. The composites are built for all 135 clustered and 69 isolated hail days north of the Alps and for all 89 clustered and 42 isolated hail days south of the Alps (Table 5.1). We further build composites of the atmospheric circulation prior to a hail event on the first (d-1), second (d-2), and third (d-3) nonhail days. For clustered hail events, the first hail day of the cluster is day 0. For clustered hail events north of the Alps (south of the Alps), d-1, d-2, and d-3 each include 31 (21) days. North of the Alps, the number of days is not 32, the number of independent clusters, because in 2003, two independent clustering periods shared a d-1, d-2, and d-3 day. For isolated hail events, north of the Alps we find 69 days and south of the Alps 42 days for d-1, d-2, and d-3 (see Table 5.1). Because half of the clustered hail days south of the Alps are also clustered north of the Alps, we also created composites comparing hail days that are clustered both south and north of the Alps and compare them to hail days that are only clustered south of the Alps.

5.4.4 Calculating the statistical significance of the differences

We apply two-sample Kolmogorov-Smirnov (KS) tests (Bonamente, 2017, p.219-221) on 500 series of hail days. These were resampled from the original set of hail days to infer the statistical significance of the differences between composites of the atmospheric variables during clustered and isolated hail days. More details on the resampling method are available in the Appendix Sect. C.2. By applying the KS-test to 500 resample series, each difference between isolated and clustered hail day composites has 500 significance test results. To reduce the chance of type I errors in the large number of p-values, we control the probability of rejecting the null hypothesis with the false discovery rate and limit the probability that a rejected null hypothesis should have been accepted to $\alpha_{FDR} = 10$ % (Wilks, 2016). In the difference maps presented in the results

section, we highlight the areas where more than 50 % or 80 % of the 500 tests indicate significant differences. The binary variables, fronts and blocks, are not tested for significance, and differences are only discussed qualitatively.

5.5 Results

5.5.1 Seasonality of isolated and clustered hail days

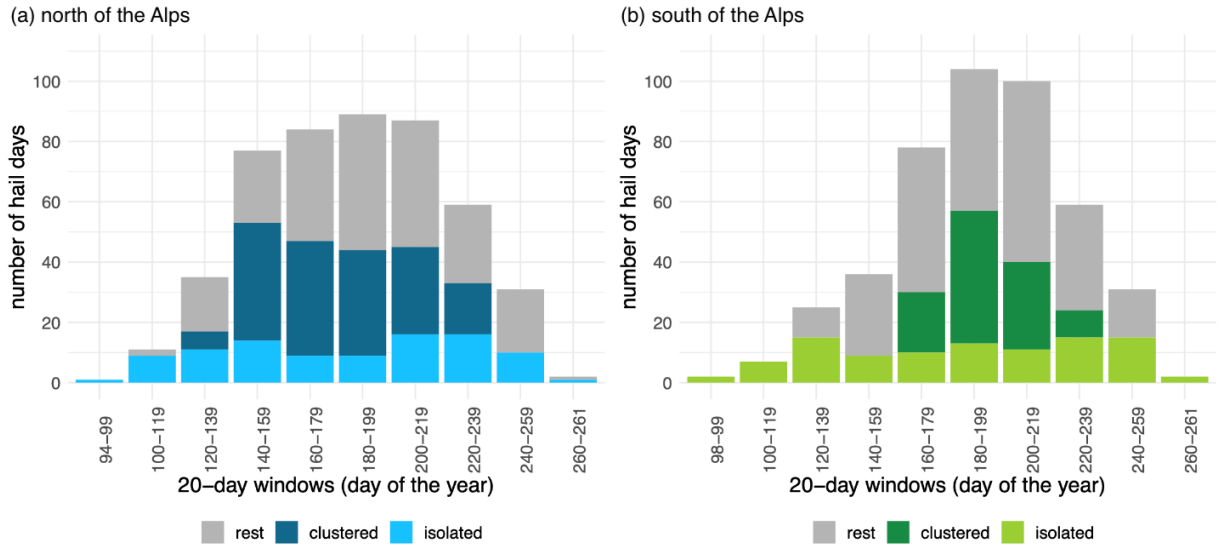

Figure 5.3: The total number of hail days (grey bars), the number of clustered hail days (darker color), and the number of isolated hail days (lighter color) for 20-day windows across the hail season for hail events north of the Alps (a, blue colors) and south of the Alps (b, green colors) between 2002–2019. The numbers on the x-axis indicate the day of the year. In this graph, the isolated hail days are also shown for the period outside of which clustered hail days are defined.

The seasonality of all hail days, the clustered hail days, and the isolated hail days are illustrated by showing the total number of hail days per 20-day window across the hail season in the period from 2002 to 2019 (Fig. 5.3). On both sides of the Alps, isolated hail days occur earlier and later in the year than clustered hail days. Clustered hail days occur earlier in the hail season north of the Alps than south of the Alps. South of the Alps, clustered hail days exhibit a pronounced peak in the middle of the year, whereas isolated hail days are distributed more uniformly across the hail season. There is a notable year-to-year variability (see Fig. 5.2). Both north and south of the Alps, 20-day periods without clustered or isolated hail days occur in at least one of the 18 years.

5.5.2 Weather type classifications

We start by comparing the distribution of central European weather types during clustered and isolated hail days. For all classes of hail days, the two most common weather types are westerly

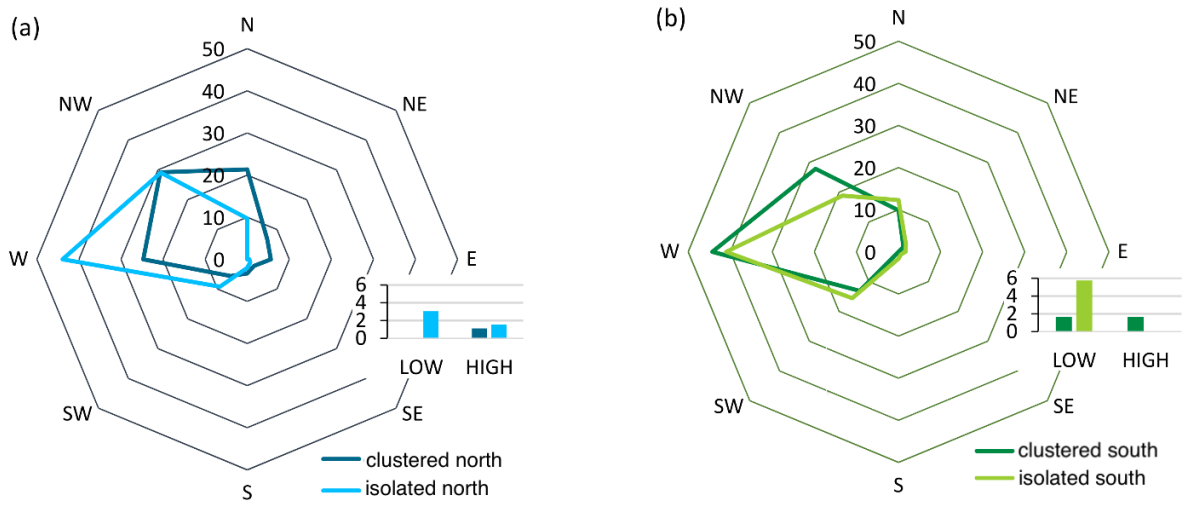

Figure 5.4: Relative frequency [%] of weather types per hail day category for hail events a) north of the Alps and b) south of the Alps. The eight directional weather types are shown in the spider plot, and low- and high-pressure weather type frequencies are indicated in the inset.

(W) and southwesterly (SW) winds in central Europe (Fig. 5.4). North of the Alps, westerly flow is more common during isolated hail days (43 % compared to 27 %), and northerly flow is more common during clustered hail days (22 % compared to 10 %). South of the Alps, the fraction of days with a northwesterly flow is about twice as frequent (8 % higher) for clustered hail days than for isolated ones. However, the overall similarity in fraction of weather types for all classes of hail days suggests that the main wind direction over central Europe is too general an indicator to differentiate between clustered and isolated hail days, and we therefore present more detailed composites of the large-scale flow next.

5.5.3 Large-scale weather situation during clustered and isolated hail days

5.5.3.1 Atmospheric conditions over Europe and the North Atlantic on hail days north of the Alps

The upper-level atmospheric flow is represented by composites of the PV on the 335K isentrope. In the PV composite for the clustered hail days, a trough is located over western Europe with its axis at 10°W (Fig. 5.5a). The trough extends meridionally to southern Spain. A downstream ridge with its axis located at 12°E over Italy tilts anticyclonically. Downstream of the ridge, an anticyclonically tilted trough extends from the Black Sea to the eastern Mediterranean. The anticyclonic tilt of the ridge and the trough point to anticyclonic Rossby wave breaking over central and eastern Europe. During at least 5-10% of clustered hail days, an atmospheric block is located at 66°N between Scandinavia and Iceland.

In the PV composite for isolated hail days, a trough is present over western Europe with the trough axis located at 2°W (Fig. 5.5c). The meridional amplification of this trough is slightly

weaker than the western European trough during the clustered hail days (Fig. 5.5e), but the trough on isolated days is deeper at 50°N. The axis of the downstream ridge is located at 12°E (Fig. 5.5c). The ridge does not exhibit a noticeable tilt. A downstream trough is located over the Black Sea. The meridional amplitude of this trough is smaller than that of its counterpart in the clustered hail day composite (Fig. 5.5e). Zonal winds at 250 hPa are significantly weaker over the Mediterranean and central Europe (Fig. 5.5e). In the isolated hail days composite, a stronger ridge is located upstream of Europe over the Atlantic at 30°W and a trough at 55°W (Fig. 5.5a, c and e). During at least 5-10% of the isolated hail days, atmospheric blocking occurred over North America at 60-70°N and 50-80°W. The jet over the central Atlantic has a southwest–northeast tilt during isolated hail days. Hence, the upper-level flow during clustered hail days is characterized by a longer wavelength of the waves over Europe, by a stronger meridional amplification of the troughs, by wave breaking, and by a weaker zonal flow over Europe than during the isolated hail days. All of these factors indicate a more stationary flow situation over Europe during the clustered hail periods.

The air contains significantly more moisture (3-5 mm) over western and central Europe north of the Alps on clustered hail days (Fig. 5.5b, d and f). On clustered hail days, the winds are on average weaker than 4m s^{-1} at 850hPa and flow from SSW over northern Switzerland (Fig. 5.5b). On isolated hail days, the winds at 850hPa are significantly stronger and southwesterly (Fig. 5.5d and f).

On clustered hail days, maximum daily temperatures are significantly warmer than isolated hail days (+1 to +4 K) over northern Switzerland (Fig. 5.6a). Furthermore, the sea level pressure is significantly higher than on clustered than on isolated days over Central Europe and northern Europe (+2 to >+5 hPa, Fig. 5.6e). The sea-level pressure pattern and the weaker lower tropospheric zonal winds north of the Alps on clustered hail days (Fig. 5.5b) indicate more stationary conditions north of the Alps. No significant differences in the wind shear are found over Switzerland (not shown).

On clustered hail days, cold fronts are less frequent just north of the Alps and over central France than on isolated hail days (Fig. 5.6b and d) and more frequent over northern France and the English Channel. Hence, cold fronts are further away from northern Switzerland during the clustered hail days.

On clustered hail days, CAPE values over central Europe and the Mediterranean are larger than on isolated hail days (Fig. 5.6b and d). The difference over northern Switzerland is significant and substantial at >400 J kg^{-1} (Fig. 5.6f). In summary, persistent warmer and more humid conditions north of the Alps during clustered hail days contribute to significantly higher instability.

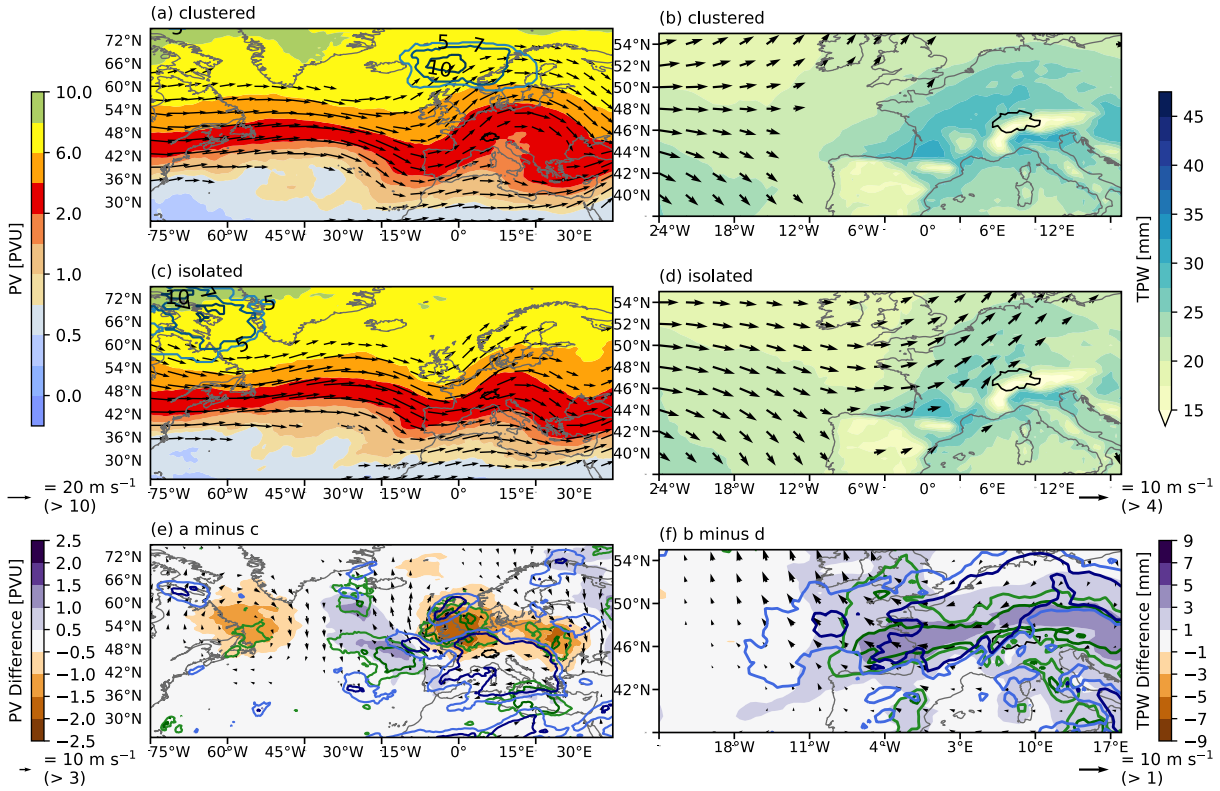

Figure 5.5: For hail events north of the Alps, left column: Potential vorticity at 335 K (color shading; in PVU), wind at 250 hPa (vectors, in m s^{-1}) and atmospheric blocking frequency (dark green contours for 5 %, 7%, and 10%). a) clustered hail days (n = 135 days), b) isolated hail days (n = 69 days), e) difference a minus c of PV at 335 K (color shading) and winds (vectors). Statistically significant differences for more than 50% (lighter contour lines) and more than 80% of the resample composites (darker contour lines) of all 500 ks-test and FDR-corrected p-values are shown in e) in green for PV at 335 K and in blue for wind at 250 hPa. Right column: TPW (filled contours, in mm) and wind at 850 hPa (vectors). b) clustered hail days (n = 135 days), d) isolated hail days (n = 69 days), f) difference b minus d. Using the same framework as in e), significant differences are shown in green for TPW and in blue for wind at 850 hPa. The brackets below the wind vector legends indicate the minimum wind speed that is visualized. The grey contours show coastlines, and the black contour line shows the border of Switzerland.

5.5.3.2 Atmospheric conditions over Europe and the North Atlantic prior to hail events north of the Alps

Composites of the days preceding the hail days illustrate the evolution of the upper-level changes that result in the more meridionally amplified flow over Europe on clustered hail days. These composites each consist of 31 days (see Table 5.1). Three days prior to the clustered hail events north of the Alps, the trough over western Europe at 18°W exists, and a ridge is present over central Europe (Fig. 5.7a). On at least 10% of the days, an atmospheric block is present over Scandinavia north of that ridge. The flow is highly diffluent upstream of the ridge, and the

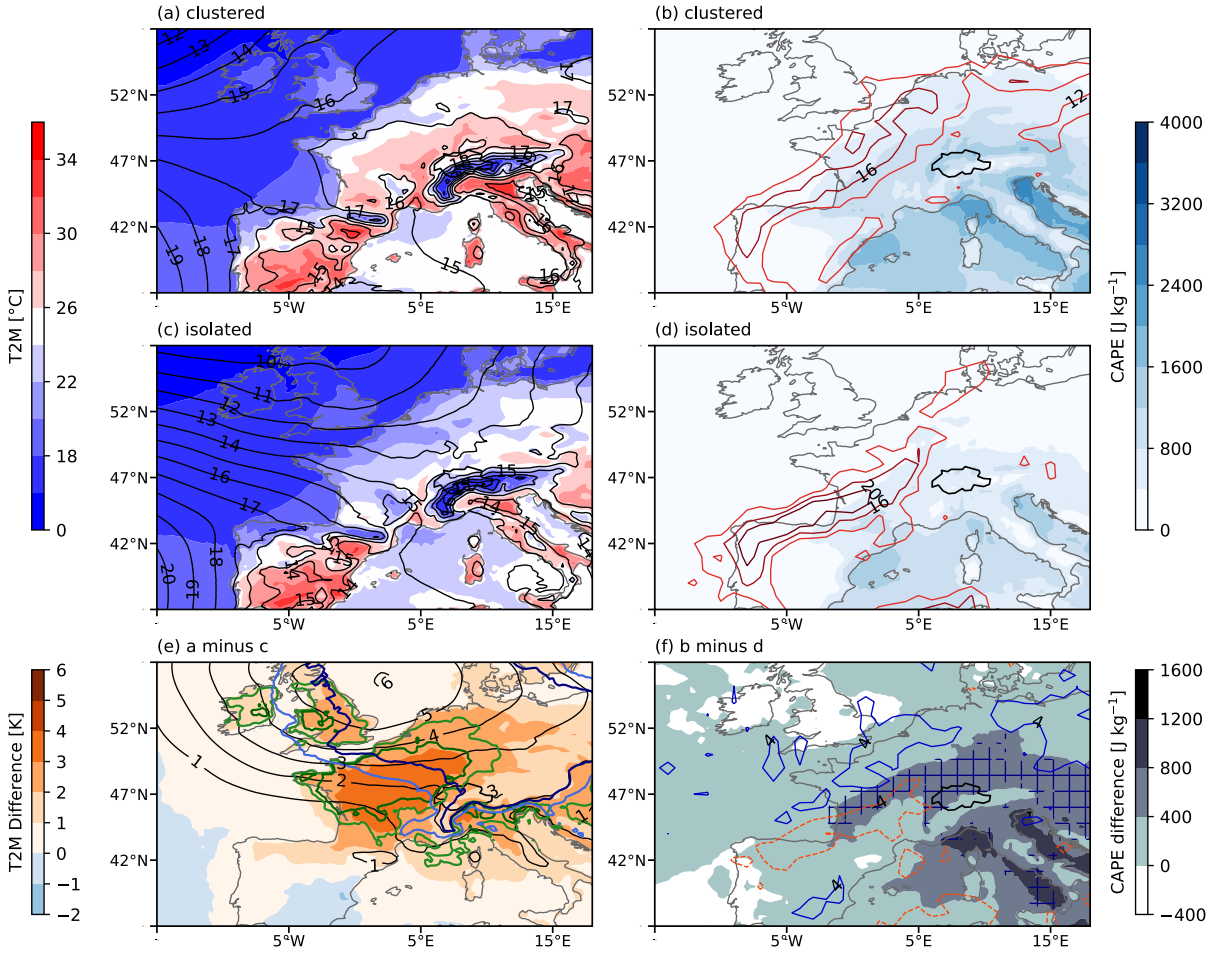

Figure 5.6: For hail events north of the Alps, left column: a) and c) daily maximum T2M (color shading; in °C) and daily mean sea- level pressure (black contour lines, labels indicate by how much the MSLP exceeds 1000 hPa in hPa) a) clustered hail days (n = 135 days), c) isolated hail days (n = 69 days). e) Difference a minus c. Statistically significant differences for more than 50 % of the resample composites (lighter contour lines) and more than 80 % (darker contour lines) of all 500 ks-test and FDR-corrected p-values are shown in green for T2M and in blue for MSLP in e) Right column: CAPE (filled contours, J kg^{-1}) and front frequencies (red, labelled lines) for b) clustered hail days, d) isolated hail days, f) difference b-d, the areas with > 50 % significant differences in CAPE are shown in blue hashes (> 80 % almost never present).

zonal flow over Europe is very weak. This diffluence sustains the meridional amplification of the upstream trough. One day later, the upstream trough over the east Atlantic widens zonally (Fig. 5.7d). The ridge over Central Europe starts tilting anticyclonically at d-2 and so does the downstream trough (Fig. 5.7d and g).

The troughs and ridges over Europe strongly amplify from d-3 to d-1 before isolated hail days (Fig. 5.7b, e and h). Over the North Atlantic a ridge amplifies at 40°W, and atmospheric block-

ing is present over the western and central Atlantic (Fig. 5.7b, e and h). The moisture content of the atmosphere also increases by 2mm from d-2 to d-1 (Fig. 5.8b, e and h).

In summary, the flow is more diffluent and meridionally amplified over Europe three days prior to clustered hail days compared to isolated hail days. Prior to both clustered and isolated hail days, the local atmospheric moisture content increases slightly from d-2 to d-1.

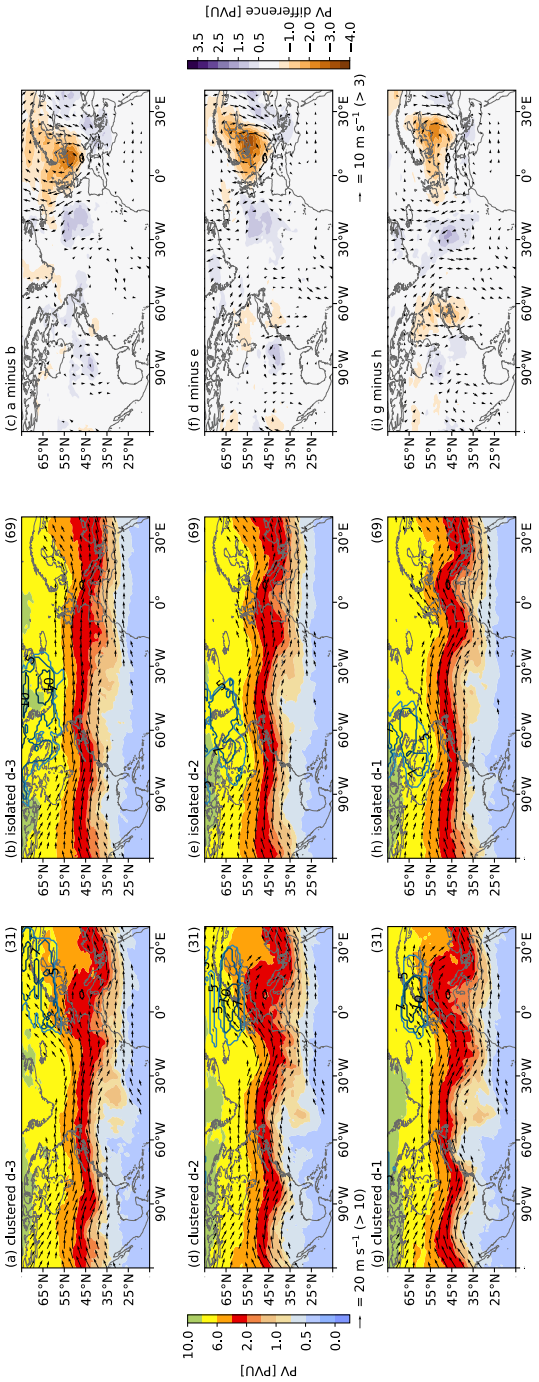

Figure 5.7: PV at 335 K (colored contours), wind at 250 hPa (vectors) and blocking frequency (contour lines) for the three days prior to clustered (left column) and isolated (central column) hail days north of the Alps. The differences are shown in the right column. See caption of Fig. 5.5 for details.

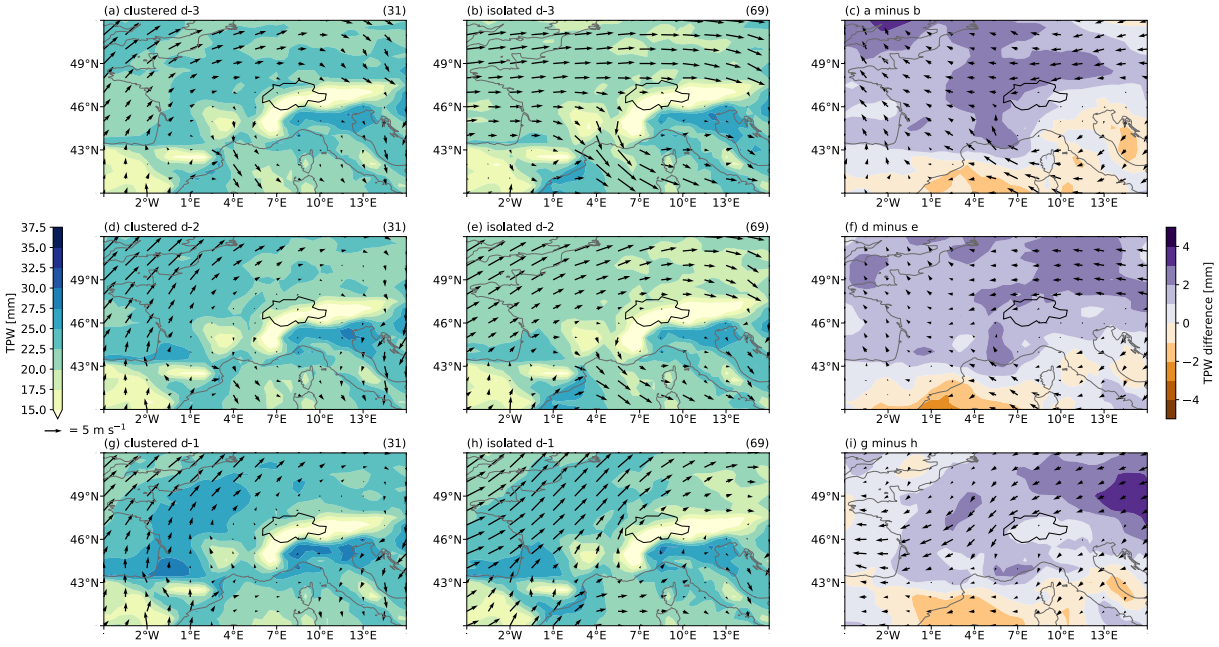

Figure 5.8: TPW (colored contours) and wind at 850 hPa (vectors) for the three days prior to clustered (left column) and (central column) isolated hail days north of the Alps. The differences are shown in the right column. See caption of Fig. 5.5 for details.

5.5.3.3 Atmospheric conditions over Europe and the North Atlantic on hail days south of the Alps

Similar to the north side of the Alps, a trough over western Europe is located further west on clustered days (2°W) (Fig. 5.9a) than on isolated days (5°E) (Fig. 5.9c). The trough in the clustered days composite is tilted anticyclonically in the subtropics. The downstream ridge over Europe is centered at 18°E in both composites, pointing to a longer wavelength and hence slower propagation of the waves on the clustered days. The ridge over central Europe is more amplified in the clustered composite (Fig. 5.9e). The downstream trough over the Black sea and the eastern Mediterranean tilts anticyclonically in the clustered days composites (Fig. 5.9a). No blocks are present over Europe during either clustered or isolated hail days.

In the lower troposphere, the winds on clustered hail days are on average weak (< 4 m s^{-1}) in and around Switzerland (Fig. 5.9b). On isolated hail days, westerly winds are slightly but significantly stronger north of the Alpine ridge (Fig. 5.9d). The moisture content of the atmosphere is higher over most of Europe during clustered hail days (Fig. 5.9b and d) and marginally significantly higher (1-3 mm) over southern Switzerland (Fig. 5.9f).

Daily maximum temperatures during clustered hail events are significantly warmer by 2–3K south of the Alps and by 1–2K over the Mediterranean Sea close to Italy (Fig. 5.10a, c, and e). The differences in mean sea level pressure between clustered and isolated hail days south of the Alps show a weaker low-pressure area during clustered days east of Denmark and south of the Alps (marginally significant, Fig. 5.10e), weaker high-pressure area over the east Atlantic (not significant), and hence a weaker north-south pressure gradient upstream of the Alps. The mean sea level pressure is higher north of the Alps than south of the Alps. Mean sea level pressure is significantly higher on clustered hail days in a band along the southern edge of the Alps (Fig. 5.10e), and there is hence a stronger pressure gradient across the Alps on isolated hail days. On clustered hail days, cold fronts are more often present northwest of the Alps (Fig. 5.10b) and over the Bay of Biscay than on isolated hail days (Fig. 5.10d). In contrast, more cold fronts are located directly over the Alpine ridge on isolated hail days than on clustered hail days.

CAPE values are statistically significantly larger on clustered hail days over parts of Italy (Fig. 5.10f), the northern Adriatic (difference >800 J kg^{-1}), and the Gulf of Genoa. In the study region south of the Alps, CAPE values are insignificantly larger by 350-400 J kg^{-1} on clustered hail days than on isolated days. No significant differences in the wind shear are found over Switzerland (not shown).

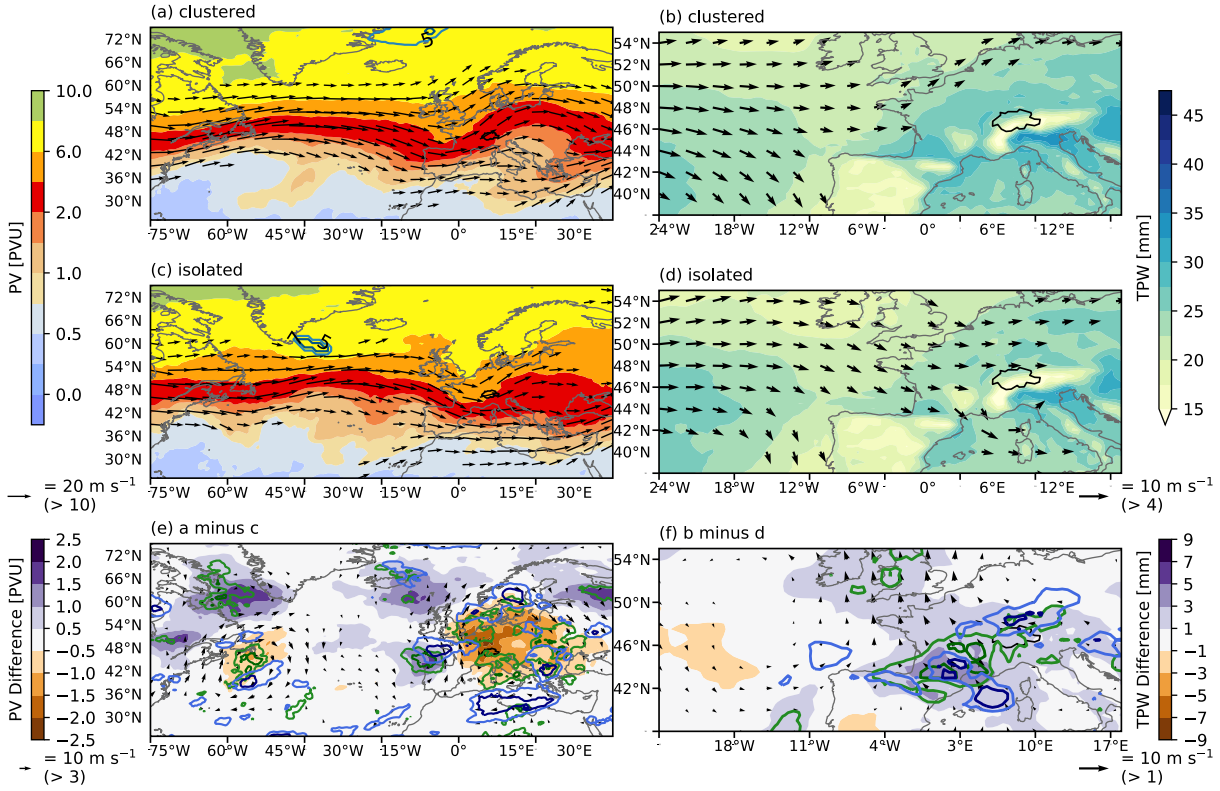

Figure 5.9: During clustered and isolated hail days south of the Alps: left column: PV, winds at 250hPa and atmospheric blocking, right column: TPW and winds at 850hPa. In e), significant differences in 250hPa winds are shown with blue contours and in PV with green contours. In f), the significant differences in 850hPa winds are shown with blue contours and in TPW with green contours. See Fig. 5.5 for more details.

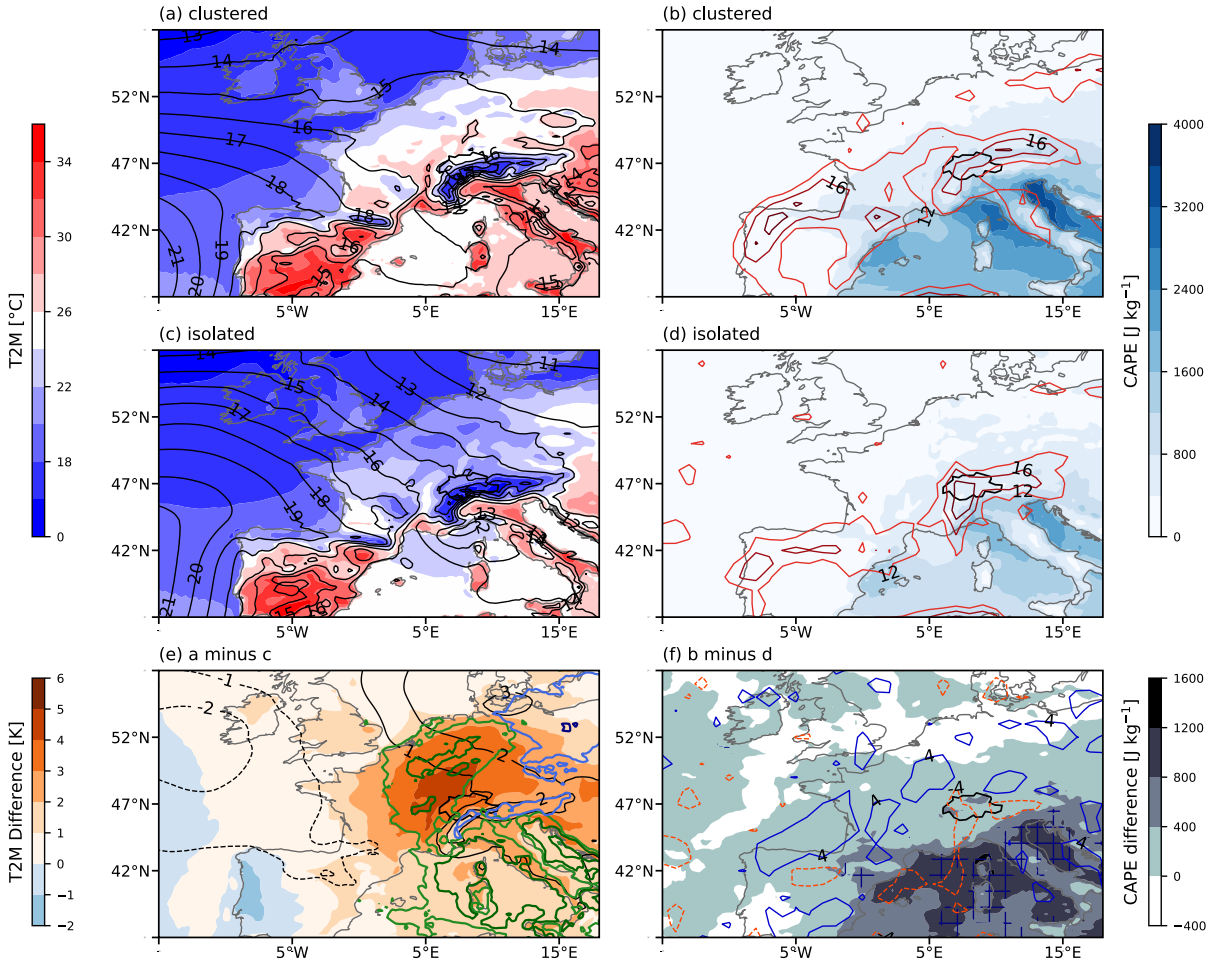

Figure 5.10: During clustered and isolated hail days south of the Alps: left column: T2M and MSLP, right column: Fronts and CAPE. In e) the significances of differences for MSLP (T2M) are shown in blue (green) and in f) the significances of differences for CAPE are shown in blue hashes. See Fig. 5.6 for more details.

5.5.3.4 Atmospheric conditions over Europe and the North Atlantic prior to hail events south of the Alps

The Rossby wave pattern over Europe during clustered hail days exhibits a stronger ridge over central Europe compared to the pattern on isolated hail days. Composites of the days preceding the hail days illustrate the evolution of the flow resulting in this ridge formation.

Three days prior to clustered hail events, a trough is present at 18°W and a ridge upstream is centered at 50°W (Fig. 5.11a). The trough at 18°W is tilted cyclonically. Although atmospheric blocking is present over the northeastern Atlantic and Scandinavia on d-3, the blocked area decreases in the following two days (Fig. 5.11a, d, and g). Over western Europe, the flow is southwesterly, and a small ridge is present at d-3. A downstream trough is present over Greece. Both the ridge over the central North Atlantic and the trough at 18°W amplify over the next two days, and a strong southwesterly flow remains present over western Europe. Over the same period, the ridge over central Europe amplifies, and the downstream trough over Greece breaks anticyclonically. Hence, a typical example of downstream wave propagation is visible in the lagged composites.

The moisture content of the atmosphere over the study region south of the Alps increases by 3 mm between d-3 and d-2 in the composite of the clustered hail days (Fig. 5.12a and d). Over the study region, the air contains 1–3 mm more TPW prior to clustered days than to isolated hail days (Fig. 5.12c).

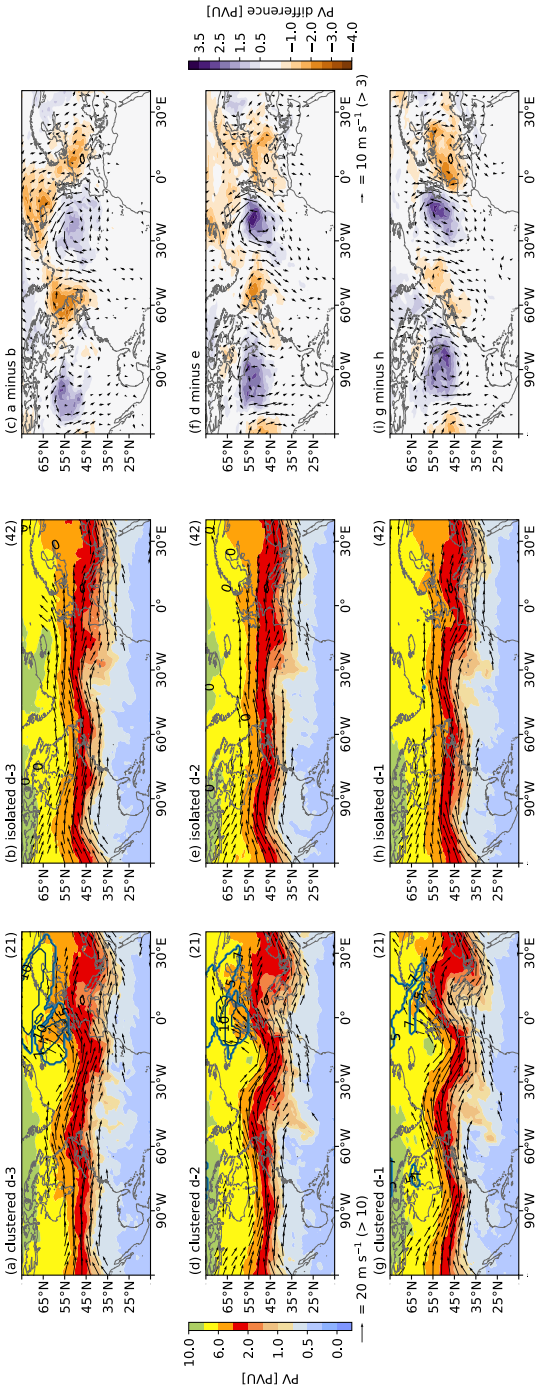

Figure 5.11: Composites of PV at 335K, winds at 250hPa, and atmospheric blocking frequency for the three days prior to clustered (left column) and isolated (central column) hail days south of the Alps. The differences are shown in the right column. See caption of Fig. 5.5 for details.

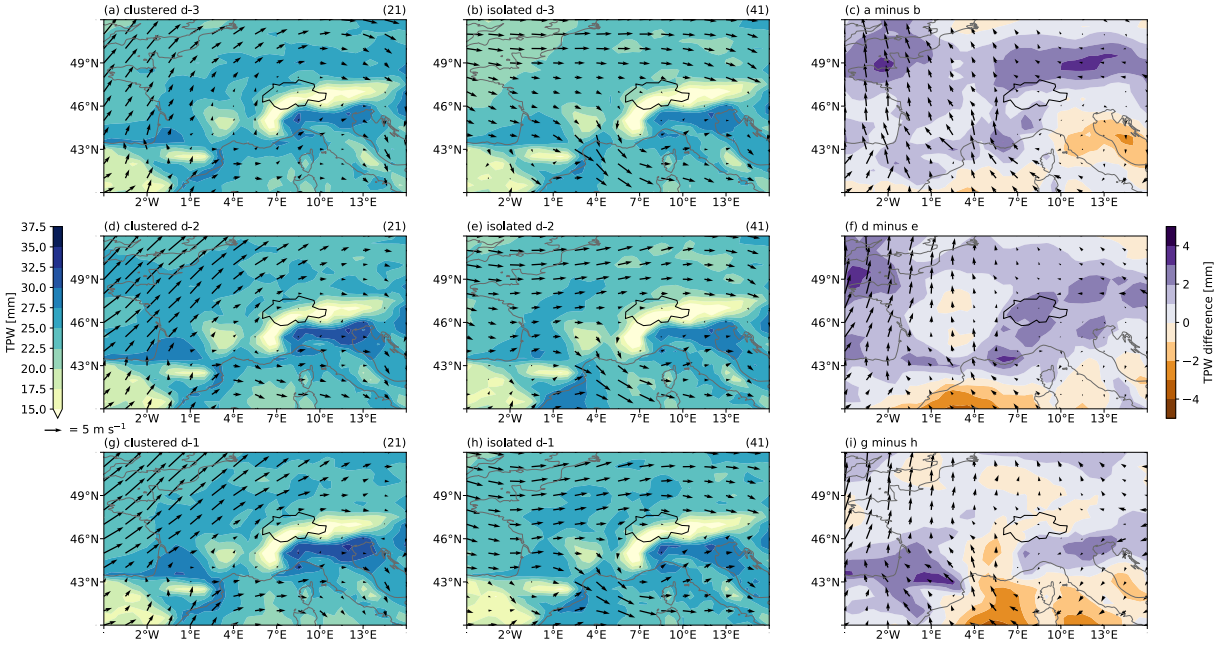

Figure 5.12: Composites of TPW (colored contours) and winds at 850hPa (vectors) for the three days prior to clustered (left column) and isolated (central column) hail days south of the Alps. The differences are shown in the right column. See caption of Fig. 5.5 for details.

5.5.3.5 Concurrent clustered hail days north and south of the Alps versus hail days that are clustered only south of the Alps

During clustered hail days only south of the Alps, the dynamical tropopause (2 PVU contour) is located over southern France. When clustered hail days also occur north of the Alps, the dynamical tropopause is located further north (Fig. 5.13a, c, and e). During clustered hail days on both sides of the Alps, fronts are frequent northwest of the Alps and over the Bay of Biscay. The frontal frequencies on clustered hail days only south of the Alps are very different. The frequency of fronts is highest on top of the Alpine ridge and close to zero northwest of the Alps (Fig. 5.13b, d, and f). The daily maximum CAPE values are higher (> 800 J kg^{-1}) over most of the Mediterranean, the difference culminating at >1200 J kg^{-1} in the gulf of Genoa and west of Corsica and Sardinia (Fig. 5.13f). When clustered hail days occur only south of the Alps, thunderstorms seem to be influenced more by convergence zones with stronger southwesterly low-level winds and nearby fronts. During clustered hail days north and south of the Alps, fronts are located further away to the northwest of the Alps, and the local instability governs thunderstorm activity south of the Alps with high values of CAPE.

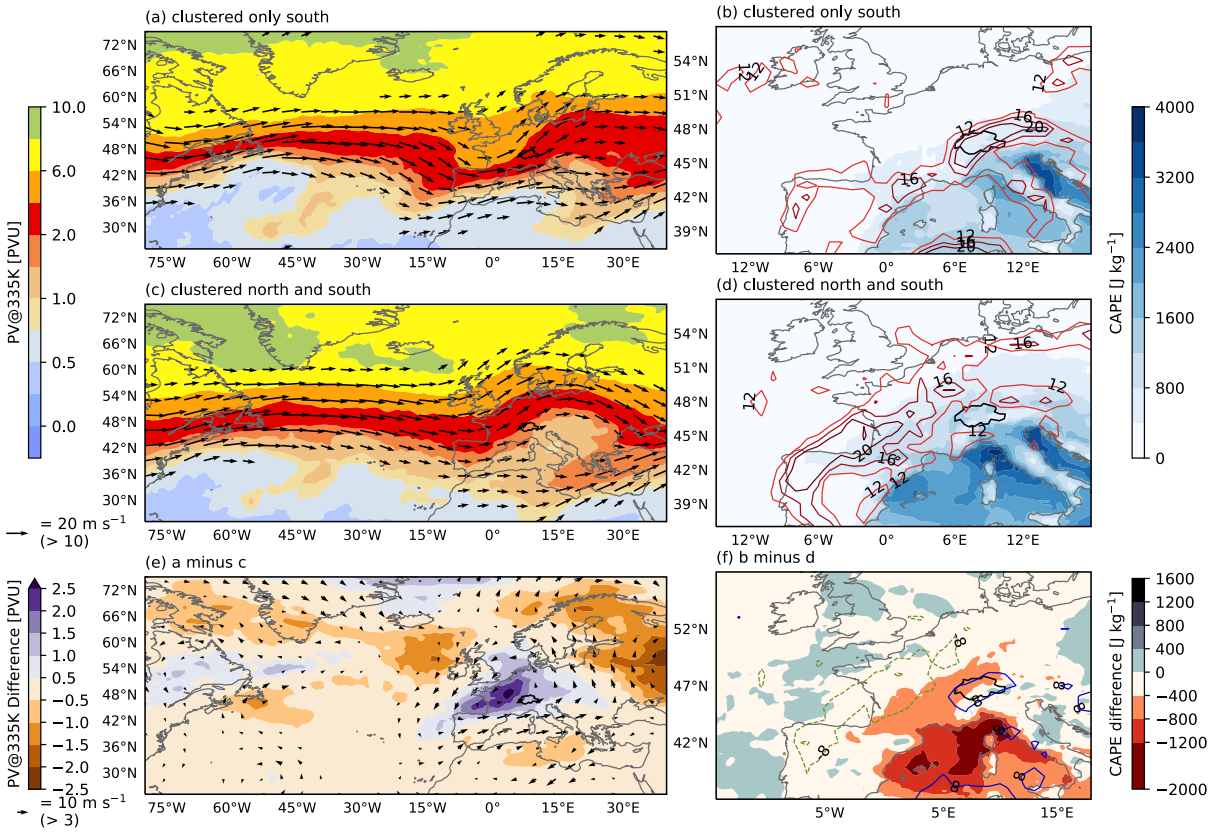

Figure 5.13: PV at 335K, horizontal winds at 250hPa (left column), CAPE and fronts (right column) during clustered hail days occurring only south of the Alps (top row), during clustered hail days occurring both north and south of the Alps (middle row), and their differences (bottom row; significances not shown). See Fig. 5.5 and 5.6 for details.

5.6 Summary and discussion

The large-scale and local-scale atmospheric conditions during and prior to multi-day hail clusters and isolated hail days in northern and southern Switzerland are characterized and compared. Hail days between April and September 2002–2019 are defined for a region within the range of Swiss radar stations north of the Alps and for a similar region south of the Alps (Fig. 5.1). The atmospheric situations are described using a weather type classification and large-scale and local-scale atmospheric conditions.

Clustered hail days occur only during the summer months, whereas isolated hail days also occur earlier and later in the year (Fig. 5.2 and 5.3). Differences between isolated and clustered hail days in the prevailing central European weather types are small (Fig. 5.4).

5.6.1 Atmospheric conditions prior to and during hail events north of the Alps

Several characteristics of the large-scale flow over the North Atlantic and Europe point to more stationary flow conditions on clustered hail days than on isolated hail days (Fig. 5.5). The flow is more amplified meridionally, and Rossby waves break downstream of Switzerland, resulting in weak upper-level zonal winds over central and eastern Europe and weaker surface zonal winds. In addition, blocking anticyclones over Scandinavia can contribute to more persistent flow over central Europe (Mohr et al., 2020). There are fewer fronts over western Europe on clustered hail days than on isolated hail days, and the fronts are located further from northern Switzerland (Fig. 5.6). This difference in front locations points to thermotopographic winds being more relevant for convection initiation during clustered hail days (see e.g., Trefalt et al., 2018; Schemm et al., 2016), whereas prefrontal convergence and prefrontal orographic flow could be more relevant for the initiation of hailstorms on isolated hail days (Schemm et al., 2016; Nisi et al., 2020). On clustered hail days, the air is more humid and warmer by $>$ 2–3 K in Central Europe north of the Alps, and CAPE is on average 400-800 J kg^{-1} higher than on isolated hail days over a large area north of the Alps. Hence, instability is substantially higher on clustered hail days, and the stationary flow allows these conditions to persist. The slow-moving large-scale flow signal suggests that clustered hail days might be more predictable than isolated hail days (e.g., Trapp, 2014; Dalcher and Kalnay, 1987).

A high atmospheric moisture content across western and central Europe may be needed for sustained convection over several days north of the Alps. Because northern Switzerland is about 600 km away from oceanic moisture sources, evapotranspiration over land is an important moisture source in summer (Sodemann and Zubler, 2010).

The higher temperatures, higher CAPE values, and the location of blocks during clustered hail days compared to isolated hail days agree well with the differences in local conditions between a month with many hail days and a month with few hail days in northern Switzerland described

in Madonna et al. (2018).

Upper-level Rossby waves over the Atlantic are more amplified meridionally prior to clustered days than isolated days (Fig. 5.7). In addition, blocking over Scandinavia on days prior to clustered hail days may contribute to a diffluent flow over Europe that amplifies the troughs that reach Europe from upstream (e.g., Shutts, 1983). The local moisture content increases by 3 mm prior to both clustered and isolated hail days.

5.6.2 Atmospheric conditions prior to and during hail events south of the Alps

Differences in the large-scale flow conditions between clustered and isolated hail days south of the Alps are similar to the differences north of the Alps with the distinction that atmospheric blocks do not occur over Europe. The large-scale flow is more stationary during clustered hail days than isolated hail days. Locally, the atmosphere is slightly warmer and contains more humidity on clustered days; however, the difference in local CAPE between isolated and clustered days is small and not significant (Fig. 5.9 and 5.10). The pressure gradient across the Alps is stronger on isolated hail days, and slightly more fronts are present south of the Alps on isolated hail days. Hence, the favorable conditions for hail are more transient on isolated hail days due to a less stationary large-scale flow, and the stronger cross-Alpine pressure gradients may indicate that Foehn winds support short-lived prefrontal convergence zones. The trough just west of Switzerland and the fronts located directly over the south side of the Alps on isolated hail days typically produce a low-level convergence with thermal winds over this area. These conditions are known to last only several hours and to develop severe hailstorms in the southern Prealps (Luca Nisi, personal communication).

Prior to clustered hail days south of the Alps, the Rossby waves over western Europe have a larger meridional amplitude than prior to isolated hail days (Fig. 5.11). A trajectory analysis could assess whether this strong amplification results in the transport of moist air masses from the subtropics towards the Alps.

South of the Alps, 58 % of isolated hail days are outside of the seasonal window within which clustered hail days occur. This may be related to the different convective environments of hailstorms in the middle of the convective season and in the shoulder seasons. In midsummer, the supply of humid and unstable air from the Mediterranean towards the Alps ahead of the trough over the eastern Atlantic may create conditions favorable for hail day clustering. The unstable air masses do not require additional synoptic forcing for the formation of hailstorms. Given the slow propagation of the trough eastward, additional moisture and warm air can be advected towards the Alps on subsequent days. Furthermore, strong radiative heating over the Alps, the resulting thermo-topographic flows, and the evapotranspiration of moisture may support hail day clustering. In contrast, at the beginning and end of the hail season, the air masses are not as unstable to begin with and therefore need stronger triggers to produce hailstorms.

Even though the flow over the Atlantic on isolated hail days resembles the flow during the negative North Atlantic Oscillation (NAO) phase, we do not find any significant correlation between the NAO index and the isolated hail days (not shown). This is in agreement with Piper and Kunz (2017), who comment that convective activity in southern Switzerland can occur regardless of large-scale forcing thanks to the complex orographic mechanisms. More fronts occur over the Alpine ridge during isolated hail days and northwest of Switzerland during clustered hail days.

Different atmospheric conditions are associated with clustered hail days on both sides of the Alps than with clustering only south of the Alps (cf. Fig. 5.13). When clustered hail days occur only south of the Alps, stronger winds and bulk wind shears combined with lower CAPE values suggest a stronger dynamic forcing of the thunderstorms. On clustered hail days affecting all of Switzerland, CAPE values are > 1200 J kg^{-1} larger over the Ligurian sea, and the winds and bulk wind shear are weaker. Furthermore, fronts almost never occur directly over the Alpine ridge and are mostly located over the Bay of Biscay or 400 km northwest of the Alps. In this situation, both prefrontal convergence zones and orographic heating may well be the more likely drivers of convective activity.

5.7 Conclusions and outlook

Multi-day hail clusters are a regular phenomenon in Switzerland both north and south of the Alps. We observe on average 10 clustered days per year in the north and 6 clustered days in the south. We compared the large- and local-scale conditions prior to and during multi-day hail clusters and isolated hail days between 2002 and 2019, within the range of Swiss radar stations. Multi-day hail clusters occur only between mid-May and end of August on the north side of the Alps and between mid-June and mid-August on the south side of the Alps, whereas isolated hail days occur during the entire convective season.

For the regions both north and south of the Alps, the large-scale atmospheric flow over the east Atlantic and Europe prior to and during clustered hail days is more amplified meridionally and characterized by a trough located in the east Atlantic. The meridional amplification is enhanced by atmospheric blocks located over Scandinavia prior to clustered hail days. On the north side of the Alps, furthermore, warmer and more humid local conditions with significantly higher CAPE values are found during clustered hail days. Fronts northwest of Switzerland are located farther away than on isolated hail days. Our findings suggest that on the north side of the Alps, thermotopographic winds are more relevant for convection initiation during clustered hail days, whereas prefrontal convergence and prefrontal orographic flow may be more relevant for the initiation of hailstorms on isolated hail days.

The local conditions south of the Alps are warmer and more humid on clustered hail days; however, differences in local CAPE between clustered and isolated hail days are not significant. We

observe a stronger pressure gradient across the Alps on isolated hail days, which may indicate that Foehn winds support short-lived prefrontal convergence zones south of the Alps. On isolated hail days south of the Alps, local conditions supporting convection are dispersed faster by the large-scale flow.

During hail days that are clustered both north and south of the Alps, fronts are frequent over the Bay of Biscay and 400 km northwest of Switzerland. In contrast, on hail days that are clustered solely south of the Alps, fronts are almost exclusively located over the Alpine ridge. Furthermore, low-level winds are stronger south of the Alpine ridge, and CAPE values are lower north of the Alps and over the Gulf of Genoa. This suggests that when hail days cluster only south of the Alps, dynamic processes are responsible for maintaining convective conditions over several days.

Future research could compare the characteristics of hailstorms between clustered and isolated hail days, such as their duration, speed, and direction of movement, and the hour of the day at which they are most likely to occur in which regions of Switzerland. These characteristics could provide more insight into the likely trigger mechanisms during and isolated hail days. The results of this study also pose a question: Do average higher daily maximum temperatures, the weaker zonal flow, and the meridionally amplified atmospheric waves on clustered hail days mean that climate warming may increase the frequency of multi-day hail clusters?"

5.8 Acknowledgements

Many thanks go to Sebastian Schemm and Michael Sprenger for kindly providing the ERA-Interim front data set. Furthermore, many thanks go to Andrey Martynov for maintaining the servers and data sets with which the analyses were conducted. Finally, we thank Simon Milligan for editing this manuscript for its language.

Chapter 6

Summary, concluding remarks and outlook

6.1 Summary

This three-part doctoral thesis contributed to a better understanding and nowcasting of hail in Switzerland and improved surface hail observations. In the first part, crowdsourced hail reports were validated against the radar reflectivity and compared with radar-based hail algorithms. In the second part, several machine learning models were assembled to nowcast hail in individual thunderstorms in a quasi-operational setting. In the third part, synoptic- and local-scale atmospheric flow characteristics during and prior to multi-day hail clusters and isolated hail days were characterized. Each part contains its own summary and conclusion section, the key points are summarized in this last chapter again.

Chapter 2 presents the MeteoSwiss crowdsourced hail reports collected between May 2015 and October 2018. In May 2015, a hail size reporting function was introduced to the MeteoSwiss app. This function allows users to report the presence of hail and the observed hail diameter through well-known objects. The categories are "no hail", "smaller than a coffee bean" ($<$ 5–8 mm), "coffee bean" (5–8 mm), "1 Swiss franc coin" (23 mm), "5 Swiss franc coin" (32 mm), "golf ball" (43 mm) and "tennis ball" (68 mm). By October 2018, $>$ 50'000 reports had been collected. The most frequently reported hail category is "coffee bean" (5–8 mm) and it is reported most in the late afternoon. The spatial distribution of the reports reflects primarily the population density of Switzerland. A comparison to radar reflectivity data reveals that quality control and filtering using plausibility criteria is crucial for the further use of these reports. The most important filter requires reports to be close to radar reflectivity areas of at least 35 dBZ. They remove about half of the reports. Except for the largest size category, enough false reports are filtered out for them to not substantially influence statistical analyses. A positive correlation was found between the reported size and POH and MESHS. MESHS values tend to be on average 1.5 cm greater than the reported size.

Chapter 3 gives an update on the crowdsourced hail reports presented in chapter 2. The additional crowdsourced reports confirm the findings in chapter 2. As of the end of September 2020, 119'549 crowdsourced hail reports had been submitted and, excluding the category "no hail", 41'191 filtered reports remained for the analyses. Furthermore, the longer time series of reports revealed the effect of varying the visibility of the crowdsourcing function in the app and hence the importance of graphical user guidance for app-based crowdsourcing data collection.

In chapter 4 XGBoost (extreme gradient boosted tree) models are created to nowcast the presence and size of hail for the lead-times t+5', t+10', t+15', t+25', t+35' and t+45' in a quasi-operational setting. Thunderstorm environmental parameters extracted from radar, satellite, COSMO-CH model, lightning, topography, and other metadata serve as predictor variables. Twelve statistics of variables (called features) were extracted within 23 km circles along thunderstorm paths of the 2018 convective season. The target variables are the maximum POH and maximum MESHS within 23 km circles of the future cell positions. Results show that binary XGBoost models (hail yes/no) provide a better performing nowcast of the presence of hail than the Lagrangian persistence for all lead-times greater than t+5'. For t+5' both prediction skills are equal. Models predicting the occurrence of POH $\geq$ 10 % have higher Critical Success Index (0.75 for t+5' and 0.58 for t+45') than MESHS (0.5 for t+5' and 0.35 for t+45'), likely because MESHS values occur less frequently than POH. Sensitivity analyses suggest that 500–1000 top features are needed to reach the same nowcasting performance as the performance of models using all features. These top features include variables from all data sources. The most frequently used data sources are radar, model, and satellite variables. The model interpretation results show that feature values describing an intense/low storm activity at t0 increase/decrease the probability of POH $\geq$ 10 % and MESHS $\geq$ 2 cm for all lead-times.

Chapter 5 characterizes the synoptic- and local-scale atmospheric flow conditions before and during multi-day hail clusters and isolated hail days that occurred between April and September 2002–2019. Clustered and isolated hail days are defined for two regions within the Swiss radar coverage area, north of the Alps and south of the Alps respectively. The large- and local-scale atmospheric flow characteristics are captured using ERA-5 reanalysis variables, objectively identified cold fronts, atmospheric blocks, and a weather type classification. In both regions, composites of atmospheric variables indicate a more stationary and meridionally amplified atmospheric flow over Europe during multi-day hail clusters. On clustered hail days north of the Alps, blocks are frequent over the North Sea and surface fronts are located farther away from Switzerland than on isolated hail days. Furthermore, CAPE, the daily maximum temperature, and the total precipitable water are significantly larger than on isolated hail days. On the alpine south side, differences in CAPE, temperature, and moisture are smaller. On isolated hail days, the mean sea level pressure is significantly deeper south of the Alps, pointing to Foehn like winds possibly intensifying convergence zones south of the Alps. Compared to clustered hail days, convective conditions are less persistent. Three days prior to clustered hail days north and south of the Alps, Rossby waves over the eastern Atlantic are more amplified meridionally than prior to isolated hail days. Before clustered hail days, atmospheric blocks are present over Scandinavia

in more than 10 % of cases. This is not the case before isolated hail days.

6.2 Concluding remarks

The results of the first part of this project have revealed that the crowdsourced hail reports close the surface hail observation gap in Switzerland to a great degree. Their strength lies in their unprecedented number and spatial coverage, while also giving plausible estimates of the approximate hailstone diameters. Thus, crowdsourced hail observations are of great value for hail research. They also serve as a bridge between the general population and the world of research. Thanks to crowdsourced reports, the probability of different hail diameters occurring in Switzerland was determined and radar-based algorithms were compared against an observational data set with several thousand data points. The comparison of crowdsourced reports with hail algorithms has increased the understanding of POH and MESHS. It gives a more transparent perspective on the predictions of maximum POH and MESHS developed in the nowcasting chapter (chapter 4). The crowd-sourcing function has had the great advantage of being installed on a national weather service app that already had on average $>$ 500'000 daily users. Particularly in urban areas and during daytime, it is very likely that app users capture most hail events. For these populated areas, crowd-sourced reports could be used to evaluate radar-based algorithms and predictions of hail categorically. Nisi et al. (2016) attempted it with car insurance damage claims and Noti (2016) with crowdsourced hail reports. Relatively strict criteria were applied to filter the crowdsourced reports for the comparison with POH and MESHS. Future projects with other applications could potentially apply more relaxed criteria.

From the crowdsourced reports and the automatic hail sensors, MeteoSwiss developed a new radar-based hail algorithm called LEHA ("Largest Expected Hail size on a reference Area of a given size"; NCCS, 2021). This algorithm was created in the context of the national project "Hail Climatology Switzerland" (see www.hagelklima.ch). LEHA has the purpose of estimating the largest expected hail size on any area smaller than 1 km^2. It is derived from MESHS and rests on the idea that, while the maximum expected severe hail size may truly be the given MESHS value within a square kilometer, the likelihood of observing that diameter within a much smaller area within that square kilometer is small. As an example, if MESHS estimates the maximum diameter of hail to be 6 cm within 1 km^2, then LEHA estimates the likely maximum size of hail inside a 100 m^2 area to be 3.8 cm. This algorithm is expected to be applied in the context of damage to buildings, cars, and crops.

To my knowledge, the nowcasting project was the first attempt to predict the hail occurrence and hailstone diameter for individual thunderstorms and for nowcasting lead-times in Europe with a machine learning model. This project had the advantage that machine learning methods use knowledge about the atmosphere that is still waiting to be revealed within measurements and observations. Local and meso- scale processes that are specific to locations, influencing thunderstorms and the probability of hail, have likely been captured in the data. These hail

nowcasting models are not yet used operationally. This project serves as a base to do so and for future similar projects. The large amount of data assimilation and preprocessing that precedes the actual prediction implies that a good and fast data assimilation system in an operational setting is crutial. Ideally, the data collection and preprocessing would take at most a few minutes.

The nowcasting project and the analysis contrasting multi-day hail clusters to isolated hail days have shown the importance of analyzing environmental characteristics at different spatial and temporal scales. The differences in large- and local-scale situations during and up to three days in advance suggest that another level of predictability could be added to hail prediction models. The locations of fronts within central Europe, the amplitude of Rossby waves over the Euro-Atlantic sector and knowing whether hail occurred during the previous days could be interesting predictive factors. This new level of predictability would likely benefit forecasting models more than nowcasting models. However, the presented nowcasting models did not explore using past data going beyond 45 minutes before the most recent observations. Whether the multi-day evolution of atmospheric conditions influence are an added value to predicting the duration and size of hail in existing thunderstorms still needs investigating.

This doctoral thesis has led to a better understanding of crowdsourced hail reports and the ability to apply them in hail research. The comparison with the two operational radar-based hail algorithms has made their strengths and weaknesses more transparent. Hail occurrence and size nowcasting models were developed and may still be implemented in the MeteoSwiss operational nowcasting system in the future. The results of the nowcasting chapter prove the great value of using different data sources and machine learning methods, and serve as an example for future similar projects. The third part of this thesis made a step towards understanding the atmospheric conditions before and during multi-day hail clusters. This information is likely to be relevant for future forecasts. Knowing whether hail is likely to occur several days in a row gives insurances, event managers, wine growers, farmers and other concerned people a better chance to prepare and prevent hail-caused damage.

6.3 Outlook

The work presented in this thesis leads to several follow-up research ideas, some of which are mentioned in the individual chapters. The most important ones are summarized again below.

- It would be interesting to conduct an inquiry with the MeteoSwiss app users to receive feedback on the hail crowdsourcing function and to better understand their reporting behavior. To help with interpreting the reports, it would be helpful to gain a greater understanding on how well a typical user is capable of estimating the size of an object and mentally comparing it with another. Could it be, that comparing the hail diameter with an object increases the reporting error, because the mind may miss estimate the objects that the hail diameter is compared to? How many users wish to give a more accurate indication on hail size? Users

could be given the option of estimating the hail diameter in centimeters or millimeters as well.

- Given the success of crowdsourcing, the same technique could be used to capture other phenomena. Examples within atmospheric science are fog, snow, wind gusts, or also particular optical weather phenomena such as rainbows, halos, or anticrepuscular rays.

- Once a larger number of crowdsourced hail reports has been collected, they could replace POH and MESHS as a target variable to train machine learning models for hail nowcasting. Convolutional neural networks could model hail using fields of environmental variables (instead of statistics within a certain diameter) and predict the probability that a report of a certain size will be sent. The population bias could be accounted for by incorporating the population density as a predictor variable. The population density would likely be a particularly relevant predictor. To attempt a prediction that is independent of the population bias, the population density could artificially be set to the countrywide maximum in all grid-boxes. A population bias would remain, because populated areas are in valley bottoms and flat areas. The great advantage of using crowdsourced reports as a target variable would be the complete independence from other data sources. Furthermore, predicting hail in populated areas and on flat grounds is most relevant, since these are the locations where damage caused by hail are likely most relevant.

- The list of variables used to nowcast hail with machine learning could be extended to include dual-polarization variables such as the differential reflectivity (Z_{DR}) and specific differential phase (K_{DP}) and with the hydrometeor classification introduced by Besic et al. (2016). Z_{DR} and K_{DP} columns would give the models information on updraft strengths and widths and the hydrometeor classification detects small and large hail aloft.

- The SHAP method presented in the nowcasting chapter could be exploited much more. Seasonal and diurnal cycles of feature contributions to predictions could be explored and spatial maps of SHAP values could determine regional differences in feature importance. These analyses could provide a better understanding on hail predictability and help understand how the model uses different features.

- Future thunderstorm and hail prediction projects should move towards allowing seamless predictions, joining the nowcasting and the forecasting time scales. An example of an ongoing larger project aiming at developing a seamless prediction system is the SINFONY project by the German Weather Service (DWD; DWD, 2021; Blahak et al., 2018).

- The characteristics description of multi-day hail clusters and isolated hail days could extend to thunderstorm properties.

- The third part of this thesis discovered that on multi-day hail clusters Rossby waves are meridionally amplified, surface temperatures and the moisture content of the atmosphere are higher, and the CAPE more intense than on isolated hail days. These findings pose

the question, whether with anthropogenic climate change, the frequency of multi-day hail clusters will increase.

Bibliography

Adams-Selin, R. D. and Ziegler, C. L.: Forecasting hail using a one-dimensional hail growth model within WRF, Monthly Weather Review, 144, 4919–4939, https://doi.org/10.1175/MWR-D-16-0027.1, 2016.

Allen, J. T. and Tippett, M. K.: The Characteristics of United States Hail Reports: 1955-2014, Electronic J. Severe Storms Meteor, 10, 1–31, URL https://ejssm.org/ojs/index.php/ejssm/article/viewArticle/149, 2015.

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., and Avellano, A.: The Data Assimilation Research Testbed: A Community Facility, Bulletin of the American Meteorological Society, 90, 1283–1296, https://doi.org/10.1175/2009BAMS2618.1, 2009.

Anthony, M. and Holden, S. B.: Cross-validation for binary classification by real-valued functions: theoretical analysis, in: Proceedings of the eleventh annual conference on Computational learning theory, pp. 218–229, 1998.

Bachmann, K., Keil, C., Craig, G. C., Weissmann, M., and Welzbacher, C. A.: Predictability of Deep Convection in Idealized and Operational Forecasts: Effects of Radar Data Assimilation, Orography, and Synoptic Weather Regime, Monthly Weather Review, 148, 63–81, https://doi.org/10.1175/MWR-D-19-0045.1, 2020.

Barras, H., Hering, A., Martynov, A., Noti, P. A., Germann, U., and Martius, O.: Experiences with >50,000 crowdsourced hail reports in Switzerland, Bulletin of the American Meteorological Society, 100, 1429–1440, https://doi.org/10.1175/BAMS-D-18-0090.1, 2019.

Barras, H., Martius, O., Nisi, L., Schroeer, K., Hering, A., and Germann, U.: Multi-day hail clusters and isolated hail days in Switzerland – large-scale flow conditions and precursors, Weather Clim. Dynam. Discuss. [preprint], 2021, 1–32, https://doi.org/10.5194/wcd-2021-25, 2021.

Barrett, A. I., Gray, S. L., Kirshbaum, D. J., Roberts, N. M., Schultz, D. M., and Fairman Jr, J. G.: Synoptic versus orographic control on stationary convective banding, Quarterly Journal of the Royal Meteorological Society, 141, 1101–1113, https://doi.org/https://doi.org/10.1002/qj.2409, 2015.

Bell, J. R. and Molthan, A. L.: Evaluation of approaches to identifying hail damage to crop vegetation using satellite imagery, Journal of Operational Meteorology, 4, 142–159, https://doi.org/10.15191/nwajom.2016.0411, 2016.

Bell, J. R., Gebremichael, E., Molthan, A. L., Schultz, L. A., Meyer, F. J., Hain, C. R., Shrestha, S., and Payne, K. C.: Complementing Optical Remote Sensing with Synthetic Aperture Radar Observations of Hail Damage Swaths to Agricultural Crops in the Central United States, Journal of Applied Meteorology and Climatology, 59, 665–685, https://doi.org/10.1175/JAMC-D-19-0124.1, 2020.

Besic, N., Figueras i Ventura, J., Grazioli, J., Gabella, M., Germann, U., and Berne, A.: Hydrometeor classification through statistical clustering of polarimetric radar measurements: a semi-supervised approach, Atmospheric Measurement Techniques, 9, 4425–4445, https://doi.org/10.5194/amt-9-4425-2016, 2016.

Besic, N., Gehring, J., Praz, C., Figueras i Ventura, J., Grazioli, J., Gabella, M., Germann, U., and Berne, A.: Unraveling hydrometeor mixtures in polarimetric radar measurements, Atmospheric Measurement Techniques, 11, 4847–4866, https://doi.org/10.5194/amt-11-4847-2018, 2018.

Bider, M.: Statistische Untersuchungen über die Hagelhäufigkeit in der Schweiz und ihre Beziehungen zur Großwetterlage, Archiv für Meteorologie, Geophysik und Bioklimatologie, Serie B, 6, 66–90, 1954.

Blahak, U., Wapler, K., Paulat, M., Potthast, R., Seifert, A., Bach, L., Bauernschubert, E., Feger, R., Feige, K., Hoff, M., et al.: Development of a new seamless prediction system for very short range convective-scale forecasting at Deutscher Wetterdienst, in: EGU General Assembly Conference Abstracts, p. 9642, 2018.

Bonamente, M.: Hypothesis Testing and Statistics, in: Statistics and Analysis of Scientific Data. Graduate Texts in Physics, pp. 117–146, Springer, https://doi.org/10.1007/978-1-4939-6572-4_7, 2017.

Bonelli, P. and Marcacci, P.: Thunderstorm nowcasting by means of lightning and radar data: algorithms and applications in northern Italy, Natural Hazards and Earth System Sciences, 8, 1187–1198, https://doi.org/10.5194/nhess-8-1187-2008, 2008.

Bouttier, F. and Marchal, H.: Probabilistic thunderstorm forecasting by blending multiple ensembles, Tellus A: Dynamic Meteorology and Oceanography, 72, 1–19, https://doi.org/10.1080/16000870.2019.1696142, 2020.

Breiman, L.: Random forests, Machine learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Brimelow, J. C. and Taylor, N.: Verification of the MESH product over the Canadian prairies using a high-quality surface hail report dataset sourced from social media, in: 38th Conference

on Radar Meteorology, AMS, URL https://ams.confex.com/ams/38RADAR/webprogram/Paper321272.html, 2017.

Brimelow, J. C., Reuter, G. W., and Poolman, E. R.: Modeling maximum hail size in Alberta thunderstorms, Weather and forecasting, 17, 1048–1062, 2002.

Brooks, H. E.: Proximity soundings for severe convection for Europe and the United States from reanalysis data, Atmospheric Research, 93, 546–553, https://doi.org/10.1016/j.atmosres.2008.10.005, 2009.

Browning, K. A., Collier, C. G., Larke, P. R., Menmuir, P., Monk, G. A., and Owens, R. G.: On the Forecasting of Frontal Rain Using a Weather Radar Network, Monthly Weather Review, 110, 534–552, https://doi.org/10.1175/1520-0493(1982)110⟨0534:OTFOFR⟩2.0.CO;2, 1982.

Burke, A., Snook, N., Gagne, D. J., McCorkle, S., and McGovern, A.: Calibration of machine learning–based probabilistic hail predictions for operational forecasting, Weather and Forecasting, 35, 149–168, https://doi.org/10.1175/WAF-D-19-0105.1, 2020.

Cassola, F., Ferrari, F., and Mazzino, A.: Numerical simulations of Mediterranean heavy precipitation events with the WRF model: A verification exercise using different approaches, Atmospheric Research, 164–165, 210–225, https://doi.org/https://doi.org/10.1016/j.atmosres.2015.05.010, 2015.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pp. 785–794, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/2939672.2939785, 2016.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y.: xgboost: Extreme Gradient Boosting, URL https://CRAN.R-project.org/package=xgboost, r package version 1.2.0.1, 2020.

Chen, X.-W. and Lin, X.: Big Data Deep Learning: Challenges and Perspectives, IEEE Access, 2, 514–525, https://doi.org/10.1109/ACCESS.2014.2325029, 2014.

Cifelli, R., Doesken, N., Kennedy, P., Carey, L. D., Rutledge, S. A., Gimmestad, C., and Depue, T.: The community collaborative rain, hail, and snow network: Informal education for scientists and citizens, Bulletin of the American Meteorological Society, 86, 1069–1077, https://doi.org/10.1175/BAMS-86-8-1069, 2005.

Crook, N. A.: Sensitivity of Moist Convection Forced by Boundary Layer Processes to Low-Level Thermodynamic Fields, Monthly Weather Review, 124, 1767–1785, https://doi.org/10.1175/1520-0493(1996)124⟨1767:SOMCFB⟩2.0.CO;2, 1996.

Czernecki, B., Taszarek, M., Marosz, M., Półrolniczak, M., Kolendowicz, L., Wyszogrodzki, A., and Szturc, J.: Application of machine learning to large hail prediction - The importance of

radar reflectivity, lightning occurrence and convective parameters derived from ERA5, Atmospheric Research, 227, 249–262, https://doi.org/10.1016/j.atmosres.2019.05.010, 2019.

Dalcher, A. and Kalnay, E.: Error growth and predictability in operational ECMWF forecasts, Tellus A, 39A, 474–491, https://doi.org/https://doi.org/10.1111/j.1600-0870.1987.tb00322.x, 1987.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Dixon, M. and Wiener, G.: TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting—A Radar-based Methodology, Journal of Atmospheric and Oceanic Technology, 10, 785–797, https://doi.org/10.1175/1520-0426(1993)010⟨0785:TTITAA⟩2.0.CO;2, 1993.

Donaldson, R. J.: Radar reflectivity profiles in thunderstorms, Journal of Atmospheric Sciences, 18, 292–305, https://doi.org/10.1175/1520-0469(1961)018⟨0292:RRPIT⟩2.0.CO;2, 1961.

Donner, S.: Hagel ist unberechenbar – ein Hagelmessnetz soll das nun ändern, Higgs, URL https://www.higgs.ch/hagel-ist-unberechenbar-ein-hagelmessnetz-soll-das-nun-aendern/33379/, last accessed: 2021-04-23, 2020.

Dotzek, N., Groenemeijer, P., Feuerstein, B., and Holzer, A. M.: Overview of ESSL's severe convective storms research using the European Severe Weather Database ESWD, Atmospheric research, 93, 575–586, https://doi.org/10.1016/j.atmosres.2008.10.020, 2009.

DWD: Development of DWD's Seamless Integrated Forecasting System, URL https://www.dwd.de/EN/research/researchprogramme/sinfony_iafe/sinfony_en_node.html;jsessionid=6D71EC4168C95B839ACBFC30AFF414B1.live21074, last accessed: 2021-04-30, 2021.

Ebert, E. E.: Ability of a Poor Man's Ensemble to Predict the Probability and Distribution of Precipitation, Monthly Weather Review, 129, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129⟨2461:AOAPMS⟩2.0.CO;2, 2001.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 15, 51–64, https://doi.org/10.1002/met.25, 2008.

Ebert, E. E., Wilson, L. J., Brown, B. G., Nurmi, P., Brooks, H. E., Bally, J., and Jaeneke, M.: Verification of Nowcasts from the WWRP Sydney 2000 Forecast Demonstration Project, Weather and Forecasting, 19, 73 – 96, https://doi.org/10.1175/1520-0434(2004)019⟨0073:VONFTW⟩2.0.CO;2, 2004.

Eilts, M. D., Johnson, J., Mitchell, E. D., Sanger, S., Stumpf, G., Witt, A., Thomas, K. W., Hondl, K. D., Rhue, D., and Jain, M.: Severe weather warning decision support system, in: Preprints, 18th Conf. on Severe Local Storms, pp. 536–540, Amer. Meteor. Soc San Francisco, CA, 1996.

Elmore, K. L., Flamig, Z., Lakshmanan, V., Kaney, B., Farmer, V., Reeves, H. D., and Rothfusz, L. P.: mPING: Crowd-sourcing weather reports for research, Bulletin of the American Meteorological Society, 95, 1335–1342, https://doi.org/10.1175/BAMS-D-13-00014.1, 2014.

Federal Statistical Office of Switzerland: Statistics of buildings and housing (StatBL). Subset: Total Population in the year 2017 (STATPOP2017)., www.bfs.admin.ch/bfsstatic/dam/assets/6027943/master, last accessed: 4 October 2018, 2017.

Federer, B., Waldvogel, A., Schmid, W., Schiesser, H., Hampel, F., Schweingruber, M., Stahel, W., Bader, J., Mezeix, J., Doras, N., et al.: Main results of Grossversuch IV, Journal of Applied Meteorology and Climatology, 25, 917–957, 1986.

Ferro, C. A. and Stephenson, D. B.: Extremal dependence indices: Improved Verification measures for deterministic forecasts of rare binary events, Weather and Forecasting, 26, 699–713, https://doi.org/10.1175/WAF-D-10-05030.1, 2011.

Flora, M. L., Potvin, C. K., Skinner, P. S., Handler, S., and McGovern, A.: Using machine learning to calibrate storm-scale probabilistic guidance of severe weather hazards in the warn-on-forecast system, arXiv, pp. 1–25, https://doi.org/10.1175/mwr-d-20-0194.1, 2020.

FOEN: Umgang mit Naturgefahren in der Schweiz - Bericht des Bundesrats in Erfüllung des Postulats 12.4271 Darbellay vom 14.12.2012. Technical Report, Swiss Federal Office for the Environment (FOEN), https://www.bafu.admin.ch/dam/bafu/de/dokumente/naturgefahren/dossiers/umgang_mit_naturgefahreninderschweiz.pdf.download.pdf/umgang_mit_naturgefahreninderschweiz.pdf, last accessed: 5 May 2021, 2016.

Foote, G. B.: A Study of Hail Growth Utilizing Observed Storm Conditions, Journal of Applied Meteorology and Climatology, 23, 84–101, https://doi.org/10.1175/1520-0450(1984)023⟨0084:ASOHGU⟩2.0.CO;2, 1984.

Foote, G. B., Krauss, T. W., and Makitov, V.: Hail metrics using conventional radar, in: Proc., 16th Conference on Planned and Inadvertent Weather Modification, URL https://ams.confex.com/ams/Annual2005/techprogram/paper_86773.htm, 2005a.

Foote, G. B., Krauss, T. W., and Makitov, V.: Hail metrics using conventional radar, in: Proc., 16th Conference on Planned and Inadvertent Weather Modification, URL https://ams.confex.com/ams/Annual2005/techprogram/paper_86773.htm, 2005b.

Foresti, L., Sideris, I. V., Nerini, D., Beusch, L. E., and Germann, U. R.: Using a 10-year radar archive for nowcasting precipitation growth and decay: A probabilistic machine learning approach, Weather and Forecasting, 34, 1547–1569, https://doi.org/10.1175/WAF-D-18-0206.1, 2019.

Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, 29, 1189–1232, URL http://www.jstor.org/stable/2699986, 2001.

Gagne, D. J., Mcgovern, A., Brotzge, J., Coniglio, M., Jr, J. C., and Xue, M.: Day-Ahead Hail Prediction Integrating Machine Learning with Storm-Scale Numerical Weather Models, Twenty-Seventh Conference on Innovative Applications of Artificial Intelligence, In AAAI, pp. 3954–3960, 2015.

Gagne, D. J., McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., and Xue, M.: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles, Weather and Forecasting, 32, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1, 2017.

Gagne, D. J., McGovern, G. R., Schwartz, C., Snook, N., Sobash, R., and Gallo, B.: Evaluation of Hail Size Forecasting Models during the 2016 Hazardous Weather Testbed Spring Experiment, in: 98th American Meteorological Society Annual Meeting, AMS, 2018.

Gagne, D. J., Haupt, S. E., Nychka, D. W., and Thompson, G.: Interpretable deep learning for spatial analysis of severe hailstorms, Monthly Weather Review, 147, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1, 2019.

Gallo, B. T., Clark, A. J., Jirak, I., Kain, J. S., Weiss, S. J., Coniglio, M., Knopfmeier, K., Correia, J., Melick, C. J., Karstens, C. D., Iyer, E., Dean, A. R., Xue, M., Kong, F., Jung, Y., Shen, F., Thomas, K. W., Brewster, K., Stratman, D., Carbin, G. W., Line, W., Adams-Selin, R., and Willington, S.: Breaking new ground in severe weather prediction: The 2015 NOAA/hazardous weather testbed spring forecasting experiment, Weather and Forecasting, 32, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1, 2017.

Gallo, K., Smith, T., Jungbluth, K., and Schumacher, P.: Hail Swaths Observed from Satellite Data and Their Relation to Radar and Surface-Based Observations: A Case Study from Iowa in 2009, Weather and Forecasting, 27, 796–802, https://doi.org/10.1175/WAF-D-11-00118.1, 2012.

Gensini, V. A., Gold, D., Allen, J. T., and Barrett, B. S.: Extended U.S. Tornado Outbreak During Late May 2019: A Forecast of Opportunity, Geophysical Research Letters, 46, 10 150–10 158, https://doi.org/10.1029/2019GL084470, 2019.

Germann, U., Boscacci, M., Gabella, M., and Sartori, M.: Radar design for prediction in the Swiss Alps, Meteorological Technology International, 42–45, URL www.ukimediaevents.com/publication/574f8129/44, 2015.

Germann, U., Figueras i Ventura, J., Gabella, M., Hering, A., Sideris, I., and Calpini, B.: Triggering innovation: The latest MeteoSwiss Alpine weather radar network, Meteorological Technology International, 62–65, URL www.ukimediaevents.com/publication/2d183b22/64, 2016.

Golding, B. W.: Nimrod: a system for generating automated very short range forecasts, Meteorological Applications, 5, 1–16, https://doi.org/https://doi.org/10.1017/S1350482798000577, 1998.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative Adversarial Networks, URL https://arxiv.org/abs/1406.2661, 2014.

Goyette, S.: Development of a model-based high-resolution extreme surface wind climatology for Switzerland, Natural Hazards: Journal of the International Society for the Prevention and Mitigation of Natural Hazards, 44, 329–339, https://doi.org/10.1007/s11069-007-9130-5, 2008.

Groenemeijer, P., Púčik, T., Holzer, A. M., Antonescu, B., Riemann-Campe, K., Schultz, D. M., Kühne, T., Feuerstein, B., Brooks, H. E., Doswell III, C. A., et al.: Severe convective storms in Europe: Ten years of research and education at the European Severe Storms Laboratory, Bulletin of the American Meteorological Society, 98, 2641–2651, https://doi.org/10.1175/BAMS-D-16-0067.1, 2017.

Hamann, U., Zeder, J., Beusch, L., Clementi, L., Foresti, L., Hering, A., Nerini, D., Nisi, L., Sassi, M., and Germann, U.: Nowcasting of thunderstorm severity with Machine Learning in the Alpine Region, URL https://repositorio.aemet.es/bitstream/20.500.11765/10617/1/NTSP5_Hamann_3ENC2019.pdf, 2019.

Heim, C., Panosetti, D., Schlemmer, L., Leuenberger, D., and Schär, C.: The Influence of the Resolution of Orography on the Simulation of Orographic Moist Convection, Monthly Weather Review, 148, 2391–2410, https://doi.org/10.1175/MWR-D-19-0247.1, 2020.

Hering, A., Morel, C., Galli, G., Sénési, S., Ambrosetti, P., and Boscacci, M.: Nowcasting thunderstorms in the Alpine region using a radar based adaptive thresholding scheme, in: Proceedings of ERAD, vol. 1, pp. 206–211, URL https://www.copernicus.org/erad/2004/online/ERAD04_P_206.pdf, 2004.

Hering, A., Nisi, L., Della Bruna, G., Gaia, M., Nerini, D., Ambrosetti, P., Hamann, U., Trefalt, S., and Germann, U.: Fully automated thunderstorm warnings and operational nowcasting at MeteoSwiss, in: European Conference on Severe Storms 2015, URL https://www.researchgate.net/profile/Luca-Nisi/publication/309817864_Fully_automated_thunderstorm_warnings_and_operational_nowcasting_at_MeteoSwiss/links/58246a0608ae61258e3cf68c/Fully-automated-thunderstorm-warnings-and-operational-nowcasting-at-MeteoSwiss.pdf, 2015.

Hering, A. M., Germann, U., Boscacci, M., and Sénési, S.: Operational nowcasting of thunderstorms in the Alps during MAP D-PHASE, in: Proceedings of the 5th European Conference on Radar Meteorology (ERAD 2008), June, 2008.

Herman, G. R. and Schumacher, R. S.: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests, Monthly Weather Review, 146, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1, 2018a.

Herman, G. R. and Schumacher, R. S.: "Dendrology" in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation, Monthly Weather Review, 146, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1, 2018b.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Hill, A. J., Herman, G. R., and Schumacher, R. S.: Forecasting Severe Weather with Random Forests, Monthly Weather Review, 148, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1, 2020.

Hohl, R., Schiesser, H.-H., and Knepper, I.: The use of weather radars to estimate hail damage to automobiles: an exploratory study in Switzerland, Atmospheric research, 61, 215–238, https://doi.org/10.1016/S0169-8095(01)00134-X, 2002.

Huntrieser, H., Schiesser, H. H., Schmid, W., and Waldvogel, A.: Comparison of traditional and newly developed thunderstorm indices for Switzerland, Weather and Forecasting, 12, 108–123, https://doi.org/10.1175/1520-0434(1997)012⟨0108:COTAND⟩2.0.CO;2, 1997.

James, P. M., Reichert, B. K., and Heizenreder, D.: NowCastMIX: Automatic Integrated Warnings for Severe Convection on Nowcasting Time Scales at the German Weather Service, Weather and Forecasting, 33, 1413–1433, https://doi.org/10.1175/WAF-D-18-0038.1, 2018.

Joe, P., Burgess, D., Potts, R., Keenan, T., Stumpf, G., and Treloar, A.: The S2K severe weather detection algorithms and their performance, Weather and Forecasting, 19, 43–63, https://doi.org/10.1175/1520-0434(2004)019⟨0043:TSSWDA⟩2.0.CO;2, 2004.

Kärnä, T. and Baptista, A. M.: Evaluation of a long-term hindcast simulation for the Columbia River estuary, Ocean Modelling, 99, 1–14, https://doi.org/10.1016/j.ocemod.2015.12.007, 2016.

Keenan, T., Joe, P., Wilson, J., Collier, C., Golding, B., Burgess, D., May, P., Pierce, C., Bally, J., Crook, A., Seed, A., Sills, D., Berry, L., Potts, R., Bell, I., Fox, N., Ebert, E., Eilts, M., O'Loughlin, K., Webb, R., Carbone, R., Browning, K., Roberts, R., and Mueller, C.: The Sydney 2000 World Weather Research Programme Forecast Demonstration Project: Overview and Current Status: Overview and Current Status, Bulletin of the American Meteorological Society, 84, 1041–1054, https://doi.org/10.1175/BAMS-84-8-1041, 2003.

Kober, K. and Tafferner, A.: Tracking and Nowcasting of Convective Cells Using Remote Sensing Data from Radar and Satellite, Meteorologische Zeitschrift, 1, 75–84, URL https://elib.dlr.de/56285/, 2009.

Kopp, J., Rivoire, P., Ali, S. M., Barton, Y., and Martius, O.: A novel method to identify sub-seasonal clustering episodes of extreme precipitation events and their contributions to large accumulation periods, Hydrology and Earth System Sciences Discussions, 2021, 1–27, https://doi.org/10.5194/hess-2021-67, 2021.

Koza, J. R., Bennett, F. H., Andre, D., and Keane, M. A.: Automated design of both the topology and sizing of analog electrical circuits using genetic programming, in: Artificial Intelligence in Design'96, edited by Gero, J. and Sudweeks, F., pp. 151–170, Springer, Dordrecht, https://doi.org/10.1007/978-94-009-0279-4_9, 1996.

Krzywinski, M. and Altman, N.: Visualizing samples with box plots, Nat Methods, 11, 119–120, https://doi.org/10.1038/nmeth.2813, 2014.

Kuhn, M. and Johnson, K.: Regression trees and rule-based models, in: Applied predictive modeling, pp. 173–220, Springer, 2013.

Kunz, M., Blahak, U., Handwerker, J., Schmidberger, M., Punge, H. J., Mohr, S., Fluck, E., and Bedka, K. M.: The severe hailstorm in southwest Germany on 28 July 2013: Characteristics, impacts and meteorological conditions, Quarterly Journal of the Royal Meteorological Society, 144, 231–250, https://doi.org/10.1002/qj.3197, 2018.

Kunz, M., Wandel, J., Fluck, E., Baumstark, S., Mohr, S., and Schemm, S.: Ambient conditions prevailing during hail events in central Europe, Natural Hazards and Earth System Sciences, 20, 1867–1887, https://doi.org/10.5194/nhess-20-1867-2020, 2020.

Lagerquist, R., McGovern, A., and Smith, T.: Machine learning for real-time prediction of damaging straight-line convective wind, Weather and Forecasting, 32, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1, 2017.

Lagerquist, R., McGovern, A., Homeyer, C. R., II, D. J. G., and Smith, T.: Deep Learning on Three-Dimensional Multiscale Data for Next-Hour Tornado Prediction, Monthly Weather Review, 148, 2837–2861, https://doi.org/10.1175/MWR-D-19-0372.1, 2020.

Lean, H. W., Roberts, N. M., Clark, P. A., and Morcrette, C.: The Surprising Role of Orography in the Initiation of an Isolated Thunderstorm in Southern England, Monthly Weather Review, 137, 3026–3046, https://doi.org/10.1175/2009MWR2743.1, 2009.

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D.: Handwritten digit recognition with a back-propagation network, in: Advances in neural information processing systems, pp. 396–404, URL https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf, 1990.

Li, P. W. and Lai, E. S. T.: Applications of radar-based nowcasting techniques for mesoscale weather forecasting in Hong Kong, Meteorological Applications, 11, 253–264, https://doi.org/https://doi.org/10.1017/S1350482704001331, 2004.

Liang, Q., Feng, Y., Deng, W., Hu, S., Huang, Y., Zeng, Q., and Chen, Z.: A composite approach of radar echo extrapolation based on TREC vectors in combination with model-predicted winds, Advances in Atmospheric Sciences, 27, 1119–1130, https://doi.org/10.1007/s00376-009-9093-4, 2010.

Liu, Y. and Just, A.: SHAPforxgboost: SHAP Plots for 'XGBoost', URL https://CRAN.R-project.org/package=SHAPforxgboost, r package version 0.0.4, 2020.

Löffler-Mang, M., Schön, D., and Landry, M.: Characteristics of a new automatic hail recorder, Atmospheric research, 100, 439–446, https://doi.org/10.1016/j.atmosres.2010.10.026, 2011.

Lorenz, E.: The butterfly effect, World Scientific Series on Nonlinear Science Series A, 39, 91–94, 2000.

Lucas, B. D. and Kanade, T.: An iterative image registration technique with an application to stereo vision, in: Proceedings DARPA Image Image Understanding Workshop, pp. 121–130, Vancouver, British Columbia, URL https://www.ri.cmu.edu/pub_files/pub3/lucas_bruce_d_1981_2/lucas_bruce_d_1981_2.pdf, 1981.

Lundberg, S. M. and Lee, S. I.: A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems, 2017, 4766–4775, 2017.

Madonna, E., Ginsbourger, D., and Martius, O.: A Poisson regression approach to model monthly hail occurrence in Northern Switzerland using large-scale environmental variables, Atmospheric Research, 203, 261–274, https://doi.org/10.1016/j.atmosres.2017.11.024, 2018.

Madsen, H. O., Krenk, S., and Lind, N. C.: Methods of structural safety, Dover Publications, Inc., 2006.

Mandapaka, P., Germann, U., and Panziera, L.: Diurnal cycle of precipitation over complex Alpine orography: inferences from high-resolution radar observations, Quarterly Journal of the Royal Meteorological Society, 139, 1025–1046, https://doi.org/10.1002/qj.2013, 2013.

Manzato, A.: Hail in Northeast Italy: A neural network ensemble forecast using sounding-derived indices, Weather and Forecasting, 28, 3–28, https://doi.org/10.1175/WAF-D-12-00034.1, 2013.

Marzban, C. and Witt, A.: A Bayesian neural network for severe-hail size prediction, Weather and Forecasting, 16, 600–610, https://doi.org/10.1175/1520-0434(2001)016⟨0600:ABNNFS⟩2.0.CO;2, 2001.

Matsumura, Y.: MlBayesOpt: Hyper Parameter Tuning for Machine Learning, Using Bayesian Optimization, URL https://CRAN.R-project.org/package=MlBayesOpt, r package version 0.3.4, 2019.

Maydl, P. and Schulter, D.: Risk management and robustness as part of sustainability assessment, in: Life-Cycle and Sustainability of Civil Infrastructure Systems, edited by Strauss, A., Frangopol, D. M., and Bergmeister, K., pp. 1644–1649, Taylor and Francis Group, 2013.

McGill, R., Tukey, J. W., and Larsen, W. A.: Variations of box plots, The American Statistician, 32, 12–16, https://doi.org/10.1080/00031305.1978.10479236., 1978.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T., and Williams, J. K.: Using artificial intelligence to improve real-time decision-making for high-impact weather, Bulletin of the American Meteorological Society, 98, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1, 2017.

McGovern, A., Karstens, C. D., Smith, T., and Lagerquist, R.: Quasi-Operational Testing of Real-Time Storm-Longevity Prediction via Machine Learning, Weather and Forecasting, 34, 1437 – 1451, https://doi.org/10.1175/WAF-D-18-0141.1, 2019a.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T.: Making the black box more transparent: Understanding the physical implications of machine learning, Bulletin of the American Meteorological Society, 100, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1, 2019b.

MeteoSwiss: Coalition-2, URL https://www.meteoswiss.admin.ch/home/research-and-cooperation/projects.subpage.html/en/data/projects/2020/coalition-2.html, last accessed: 2021-03-31, 2020a.

MeteoSwiss: COSMO forecasting system, URL https://www.meteoswiss.admin.ch/home/measurement-and-forecasting-systems/warning-and-forecasting-systems/cosmo-forecasting-system.html, last accessed: 2021-03-31, 2020b.

MeteoSwiss: Coalition-4, URL https://www.meteoswiss.admin.ch/home/research-and-cooperation/projects.subpage.html/en/data/projects/2020/coalition-4.html, last accessed: 2021-03-31, 2021.

Miyoshi, T., Kunii, M., Ruiz, J., Lien, G.-Y., Satoh, S., Ushio, T., Bessho, K., Seko, H., Tomita, H., and Ishikawa, Y.: "Big Data Assimilation" Revolutionizing Severe Weather Prediction, Bulletin of the American Meteorological Society, 97, 1347–1354, https://doi.org/10.1175/BAMS-D-15-00144.1, 2016.

Mobiliar Lab for Natural Risks: Sensoren ermöglichen neue Hagelanalysen - ein Fallbeispiel, 4th Newsletter, URL https://www.mobiliarlab.unibe.ch/ueber_uns/news/alle_newsletters/index_ger.html, last accessed: 2021-04-14, 2018.

Mobiliar Lab for Natural Risks: The Swiss Hail Network, URL https://www.mobiliarlab.unibe.ch/research/applied_research_on_hail_and_wind_gusts/the_swiss_hail_network/index_eng.html, last accessed: 2021-04-21, 2021.

Mockus, J., Tiesis, V., and Zilinskas, A.: The application of Bayesian methods for seeking the extremum, Towards global optimization, 2, 2, URL https://www.researchgate.net/publication/248818761_The_application_of_Bayesian_methods_for_seeking_the_extremum, 1978.

Mohr, S., Wandel, J., Lenggenhager, S., and Martius, O.: Relationship between atmospheric blocking and warm-season thunderstorms over western and central Europe, Quarterly Journal of the Royal Meteorological Society, 145, 3040–3056, https://doi.org/10.1002/qj.3603, 2019.

Mohr, S., Wilhelm, J., Wandel, J., Kunz, M., Portmann, R., Punge, H. J., Schmidberger, M., Quinting, J. F., and Grams, C. M.: The role of large-scale dynamics in an exceptional sequence of severe thunderstorms in Europe May–June 2018, Weather and Climate Dynamics, 1, 325–348, https://doi.org/10.5194/wcd-1-325-2020, 2020.

Molnar, C.: Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, https://christophm.github.io/interpretable-ml-book/, last accessed: 30 April 2021, 2021.

Morel, S.: Verification of radar-based hail detection algorithms with insurance loss data in Switzerland, Master's thesis, University of Bern, 2014.

Mueller, C. K., Saxen, T., Roberts, R., and Wilson, J.: Evaluation of the NCAR thunderstorm Auto-nowcast system, in: Preprints, 20th Conf. on Severe Local Storms, Orlando, FL, Amer. Meteor. Soc., J40–J45, 2000.

NCCS: Hail climate Switzerland - National hail hazard maps, Brochure, https://www.nccs.admin.ch/dam/nccs/en/dokumente/website/hagel/nccs_broschuere_hagelklima_schweiz.pdf.download.pdf/NCCS_Brochure_Hail_Climate_Switzerland.pdf, last accessed: 7 May 2021, 2021.

Nelson, S. P.: The Influence of Storm Flow Structure on Hail Growth, Journal of Atmospheric Sciences, 40, 1965–1983, https://doi.org/10.1175/1520-0469(1983)040⟨1965:TIOSFS⟩2.0.CO;2, 1983.

Niculescu-Mizil, A. and Caruana, R.: Predicting Good Probabilities with Supervised Learning, in: Proceedings of the 22nd International Conference on Machine Learning, ICML '05, pp. 625—-632, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/1102351.1102430, 2005.

Nisi, L.: Spatial and temporal distribution of hailstorms in the Alpine region. A long term, high resolution, radar-based analysis, Ph.D. thesis, University of Bern, 2019.

Nisi, L., Ambrosetti, P., and Clementi, L.: Nowcasting severe convection in the Alpine region: The COALITION approach, Quarterly Journal of the Royal Meteorological Society, 140, 1684–1699, https://doi.org/10.1002/qj.2249, 2014.

Nisi, L., Martius, O., Hering, A., Kunz, M., and Germann, U.: Spatial and temporal distribution of hailstorms in the Alpine region: A long-term, high resolution, radar-based analy-

sis, Quarterly Journal of the Royal Meteorological Society, 142, 1590–1604, https://doi.org/10.1002/qj.2771, 2016.

Nisi, L., Hering, A., Germann, U., and Martius, O.: A 15-year hail streak climatology for the Alpine region, Quarterly Journal of the Royal Meteorological Society, 144, 1429–1449, https://doi.org/10.1002/qj.3286, 2018.

Nisi, L., Hering, A., Germann, U., Schroeer, K., Barras, H., Kunz, M., and Martius, O.: Hailstorms in the Alpine region: Diurnal cycle, 4D-characteristics, and the nowcasting potential of lightning properties, Quarterly Journal of the Royal Meteorological Society, 146, 4170–4194, https://doi.org/10.1002/qj.3897, 2020.

Noti, P. A.: Verification of radar-based hail detection algorithms with insurance loss data in Switzerland, Master's thesis, University of Bern, 2016.

Panziera, L., Gabella, M., Zanini, S., Hering, A., Germann, U., and Berne, A.: A radar-based regional extreme rainfall analysis to derive the thresholds for a novel automatic alert system in Switzerland, Hydrology and earth system sciences, 20, 2317–2332, 2016.

Pennelly, C., Reuter, G., and Flesch, T.: Verification of the WRF model for simulating heavy precipitation in Alberta, Atmospheric Research, 135–136, 172–192, https://doi.org/https://doi.org/10.1016/j.atmosres.2013.09.004, 2014.

Pielke, R.: Mesoscale Meteorological Modeling, 3rd ed., Elsevier, 2013.

Pierce, C. E., Hardaker, P. J., Collier, C. G., and Haggett, C. M.: GANDOLF: a system for generating automated nowcasts of convective precipitation, Meteorological Applications, 7, 341–360, https://doi.org/https://doi.org/10.1017/S135048270000164X, 2000.

Pinto, J. G., Gómara, I., Masato, G., Dacre, H. F., Woollings, T., and Caballero, R.: Large-scale dynamics associated with clustering of extratropical cyclones affecting Western Europe, Journal of Geophysical Research: Atmospheres, 119, 13,704–13,719, https://doi.org/https://doi.org/10.1002/2014JD022305, 2014.

Piper, D. and Kunz, M.: Spatiotemporal variability of lightning activity in Europe and the relation to the North Atlantic Oscillation teleconnection pattern, Natural Hazards and Earth System Sciences, 17, 1319–1336, https://doi.org/10.5194/nhess-17-1319-2017, 2017.

Piper, D., Kunz, M., Ehmele, F., Mohr, S., Mühr, B., Kron, A., and Daniell, J.: Exceptional sequence of severe thunderstorms and related flash floods in May and June 2016 in Germany - Part 1: Meteorological background, Natural Hazards and Earth System Sciences, 16, 2835–2850, https://doi.org/10.5194/nhess-16-2835-2016, 2016.

Piper, D. A., Kunz, M., Allen, J. T., and Mohr, S.: Investigation of the temporal variability of thunderstorms in central and western Europe and the relation to large-scale flow and teleconnection patterns, Quarterly Journal of the Royal Meteorological Society, 145, 3644–3666, https://doi.org/10.1002/qj.3647, 2019.

Púčik, T., Castellano, C., Groenemeijer, P., Kühne, T., Rädler, A. T., Antonescu, B., and Faust, E.: Large Hail Incidence and Its Economic and Societal Impacts across Europe, Monthly Weather Review, 147, 3901–3916, https://doi.org/10.1175/MWR-D-19-0204.1, 01 Nov. 2019.

Púčik, T., Groenemeijer, P., Rỳva, D., and Kolář, M.: Proximity soundings of severe and nonsevere thunderstorms in central Europe, Monthly Weather Review, 143, 4805–4821, https://doi.org/10.1175/MWR-D-15-0104.1, 2015.

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., and Foresti, L.: Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0), Geoscientific Model Development, 12, 4185–4219, https://doi.org/10.5194/gmd-12-4185-2019, 2019.

Pullman, M., Gurung, I., Maskey, M., Ramachandran, R., and Christopher, S. A.: Applying Deep Learning to Hail Detection: A Case Study, IEEE Transactions on Geoscience and Remote Sensing, 57, 10 218–10 225, https://doi.org/10.1109/TGRS.2019.2931944, 2019.

Punge, H. J. and Kunz, M.: Hail observations and hailstorm characteristics in Europe: A review, Atmospheric Research, 176-177, 159–184, https://doi.org/10.1016/j.atmosres.2016.02.012, 2016.

Quinto, B.: Introduction to Machine Learning, pp. 1–27, Apress, Berkeley, CA, https://doi.org/10.1007/978-1-4842-5669-5_1, 2020.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, in: Adaptive Computation and Machine Learning, edited by Dietterich, T., the MIT Press, URL http://www.gaussianprocess.org/gpml/chapters/RW.pdf, 2006.

Reges, H. W., Doesken, N., Turner, J., Newman, N., Bergantino, A., and Schwalbe, Z.: CoCoRaHS: The evolution and accomplishments of a volunteer rain gauge network, Bulletin of the American Meteorological Society, 97, 1831–1846, https://doi.org/10.1175/BAMS-D-14-00213.1, 2016.

Roebber, P. J.: Visualizing multiple measures of forecast quality, Weather and Forecasting, 24, 601–608, https://doi.org/10.1175/2008WAF2222159.1, 2009.

Rohrer, M., Brönnimann, S., Martius, O., Raible, C. C., Wild, M., and Compo, G. P.: Representation of extratropical cyclones, blocking anticyclones, and alpine circulation types in multiple reanalyses and model simulations, Journal of Climate, 31, 3009–3031, https://doi.org/10.1175/JCLI-D-17-0350.1, 2018.

Rotach, M. W., Ambrosetti, P., Ament, F., Appenzeller, C., Arpagaus, M., Bauer, H.-S., Behrendt, A., Bouttier, F., Buzzi, A., Corazza, M., Davolio, S., Denhard, M., Dorninger, M., Fontannaz, L., Frick, J., Fundel, F., Germann, U., Gorgas, T., Hegg, C., Hering, A., Keil, C., Liniger, M. A., Marsigli, C., McTaggart-Cowan, R., Montaini, A., Mylne, K., Ranzi, R.,

Richard, E., Rossa, A., Santos-Muñoz, D., Schär, C., Seity, Y., Staudinger, M., Stoll, M., Volkert, H., Walser, A., Wang, Y., Werhahn, J., Wulfmeyer, V., and Zappa, M.: MAP D-PHASE: Real-Time Demonstration of Weather Forecast Quality in the Alpine Region, Bulletin of the American Meteorological Society, 90, 1321 – 1336, https://doi.org/10.1175/2009BAMS2776.1, 2009.

Samuel, A. L.: Some Studies in Machine Learning Using the Game of Checkers, IBM Journal of Research and Development, 3, 210–229, https://doi.org/10.1147/rd.33.0210, 1959.

Schemm, S., Rudeva, I., and Simmonds, I.: Extratropical fronts in the lower troposphere–global perspectives obtained from two automated methods, Quarterly Journal of the Royal Meteorological Society, 141, 1686–1698, https://doi.org/https://doi.org/10.1002/qj.2471, 2015.

Schemm, S., Nisi, L., Martinov, A., Leuenberger, D., and Martius, O.: On the link between cold fronts and hail in Switzerland, Atmospheric Science Letters, 17, 315–325, https://doi.org/https://doi.org/10.1002/asl.660, 2016.

Schiesser, H.: Hailfall: the relationship between radar measurements and crop damage, Atmospheric Research, 25, 559–582, https://doi.org/10.1016/0169-8095(90)90038-E, 1990.

Schmetz, J., Pili, P., Tjemkes, S., Just, D., Kerkmann, J., Rota, S., and Ratier, A.: An introduction to Meteosat second generation (MSG), Bulletin of the American Meteorological Society, 83, 977–992, https://doi.org/10.1175/1520-0477(2002)083⟨0977:AITMSG⟩2.3.CO;2, 2002.

Schmid, W., Schiesser, H., and Waldvogel, A.: The kinetic energy of hailfalls. Part IV: Patterns of hailpad and radar data, Journal of Applied Meteorology and Climatology, 31, 1165–1178, https://doi.org/10.1175/1520-0450(1992)031⟨1165:TKEOHP⟩2.0.CO;2, 1992.

Schroder, Z. and Elsner, J. B.: Quantifying relationships between environmental factors and power dissipation on the most prolific days in the largest tornado "outbreaks", International Journal of Climatology, 40, 3150–3160, https://doi.org/10.1002/joc.6388, 2020.

Schroeer, K., Trefalt, S., Hering, A., Germann, U., and Schwierz, C.: Hagelgefährdung in der Schweiz (in German), Mobiliar Lab Herbstveranstaltung 2019, URL https://www.mobiliarlab.unibe.ch/unibe/portal/fak_naturwis/g_dept_kzen/d_c_oeschger/abt_mobilab/content/e325728/e527468/e527651/e1031643/pane1031701/e1031706/Hagelgefhrdung-in-der-Schweiz_KatharinaSchrer_ger.pdf, 2019.

Schwierz, C., Croci-Maspoli, M., and Davies, H. C.: Perspicacious indicators of atmospheric blocking, Geophysical Research Letters, 31, L06 125, https://doi.org/https://doi.org/10.1029/2003GL019341, 2004.

Shafer, C., Doswell III, C. A.: Identifying and Ranking Multi-Day Severe Weather Outbreaks, in: 26th Conference on Severe Local Storms, AMS, URL https://ams.confex.com/ams/26SLS/webprogram/Paper211637.html, 2012.

Shapley, L. S.: Stochastic Games, Proceedings of the National Academy of Sciences, 39, 1095–1100, https://doi.org/10.1073/pnas.39.10.1095, 1953.

Shavlik, J. W., Dietterich, T., and Dietterich, T. G.: Readings in machine learning, Morgan Kaufmann Publishers, Inc., 1990.

Shi and Tomasi: Good features to track, in: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 593–600, https://doi.org/10.1109/CVPR.1994.323794, 1994.

Shutts, G. J.: The propagation of eddies in diffluent jetstreams: Eddy vorticity forcing of 'blocking' flow fields, Quarterly Journal of the Royal Meteorological Society, 109, 737–761, https://doi.org/https://doi.org/10.1002/qj.49710946204, 1983.

Smith, P. L. and Waldvogel, A.: On determinations of maximum hailstone sizes from hailpad observations, Journal of Applied Meteorology, 28, 71–76, 1989.

Snoek, J., Larochelle, H., and Adams, R. P.: Practical Bayesian Optimization of Machine Learning Algorithms, in: Advances in Neural Information Processing Systems, edited by Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., vol. 25, pp. 2951–2959, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf, 2012.

Sobash, R. A., Schwartz, C. S., Romine, G. S., Fossell, K. R., and Weisman, M. L.: Severe weather prediction using storm surrogates from an ensemble forecasting system, Weather and Forecasting, 31, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1, 2016.

Sodemann, H. and Zubler, E.: Seasonal and inter-annual variability of the moisture sources for Alpine precipitation during 1995–2002, International Journal of Climatology, 30, 947–961, https://doi.org/https://doi.org/10.1002/joc.1932, 2010.

Sprenger, M., Schemm, S., Oechslin, R., and Jenkner, J.: Nowcasting Foehn Wind Events Using the AdaBoost Machine Learning Algorithm, Weather and Forecasting, 32, 1079–1099, https://doi.org/10.1175/WAF-D-16-0208.1, 2017.

Stucki, M. and Egli, T.: Elementarschutzregister Hagel, Untersuchungen zur Hagelgefahr und zum Widerstand der Gebäudehülle, synthesis report in German, Präventionsstiftung der kantonalen Gebäudeversicherungen, URL http://www.fluelerpolymer.ch/documents/Synthesebericht_Hagel_D.pdf, 2007.

Stucki, P., Dierer, S., Welker, C., Gómez-Navarro, J. J., Raible, C. C., Martius, O., and Brönnimann, S.: Evaluation of downscaled wind speeds and parameterised gusts for recent and historical windstorms in Switzerland, Tellus A: Dynamic Meteorology and Oceanography, 68, 31 820, https://doi.org/10.3402/tellusa.v68.31820, 2016.

Taszarek, M., Brooks, H. E., and Czernecki, B.: Sounding-derived parameters associated with convective hazards in Europe, Monthly Weather Review, 145, 1511–1528, https://doi.org/10.1175/MWR-D-16-0384.1, 2017.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical Research: Atmospheres, 106, 7183–7192, https://doi.org/https://doi.org/10.1029/2000JD900719, 2001.

Thompson, G., Rasmussen, R. M., and Manning, K.: Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part I: Description and Sensitivity Analysis, Monthly Weather Review, 132, 519–542, https://doi.org/10.1175/1520-0493(2004)132⟨0519:EFOWPU⟩2.0.CO;2, 2004.

Thompson, G., Field, P. R., Rasmussen, R. M., and Hall, W. D.: Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization, Monthly Weather Review, 136, 5095–5115, https://doi.org/10.1175/2008MWR2387.1, 2008.

Trapp, R. J.: On the significance of multiple consecutive days of tornado activity, Monthly Weather Review, 142, 1452–1459, https://doi.org/10.1175/MWR-D-13-00347.1, 2014.

Trefalt, S.: Hail and Severe Wind Gusts in the Convective Season in Switzerland, Ph.D. thesis, University of Bern, 2017.

Trefalt, S., Martynov, A., Barras, H., Besic, N., Hering, A. M., Lenggenhager, S., Noti, P., Röthlisberger, M., Schemm, S., Germann, U., and Martius, O.: A severe hail storm in complex topography in Switzerland - Observations and processes, Atmospheric Research, 209, 76–94, https://doi.org/10.1016/j.atmosres.2018.03.007, 2018.

Treloar, A.: Vertically integrated radar reflectivity as an indicator of hail size in the greater Sydney region of Australia, in: Preprints, 19th Conf. on Severe Local Storms, Minneapolis, MN, Amer. Meteor. Soc, pp. 48–51, 1998.

VKF: Naturgefahren und Prävention innerhalb der VKF. Technical Report., 2013.

von Matt, C.: ZDR-column detection in Switzerland – Verification, Sensitivity Analysis and Associations with MAXECHO, POH and MESHS, Master's thesis, University of Bern, 2020.

Waldvogel, A., Federer, B., and Grimm, P.: Criteria for the detection of hail cells, Journal of Applied Meteorology and Climatology, 18, 1521–1525, https://doi.org/10.1175/1520-0450(1979)018⟨1521:CFTDOH⟩2.0.CO;2, 1979.

Wapler, K., Harnisch, F., Pardowitz, T., and Senf, F.: Characterisation and predictability of a strong and a weak forcing severe convective event–a multi-data approach, Meteorologische Zeitschrift, 24, 393–410, https://doi.org/10.1127/metz/2015/0625, 2015.

Weusthoff, T.: Weather Type Classification at MeteoSwiss - Introduction of new automatic classification schemes, Arbeitsberichte der MeteoSchweiz, p. 46, 2011.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100 of *International Geophysics Series*, Elsevier Inc, third edn., https://doi.org/10.1016/B978-0-12-385022-5.00001-4, 2011.

Wilks, D. S.: " the Stippling Shows Statistically Significant Grid Points ", Bulletin of the American Meteorological Society, 97, 2263–2274, https://doi.org/10.1175/BAMS-D-15-00267.1, 2016.

Willemse, S.: A statistical analysis and climatological interpretation of hailstorms in Switzerland, Ph.D. thesis, ETH Zurich, 1995.

Wilson, J. W., Crook, N. A., Mueller, C. K., Sun, J., and Dixon, M.: Nowcasting Thunderstorms: A Status Report, Bulletin of the American Meteorological Society, 79, 2079 – 2100, https://doi.org/10.1175/1520-0477(1998)079⟨2079:NTASR⟩2.0.CO;2, 1998.

Wilson, J. W., Feng, Y., Chen, M., and Roberts, R. D.: Nowcasting Challenges during the Beijing Olympics: Successes, Failures, and Implications for Future Nowcasting Systems, Weather and Forecasting, 25, 1691–1714, https://doi.org/10.1175/2010WAF2222417.1, 2010.

Witt, A., Eilts, M. D., Stumpf, G. J., Johnson, J., Mitchell, E. D. W., and Thomas, K. W.: An enhanced hail detection algorithm for the WSR-88D, Weather and Forecasting, 13, 286–303, https://doi.org/10.1175/1520-0434(1998)013⟨0286:AEHDAF⟩2.0.CO;2, 1998.

Wong, W. K. and Lai, E. S. T.: RAPIDS–Operational blending of nowcast and NWP QPF, in: The 2nd International Symposium on Quantitative Precipitation Forecasting and Hydrology, 2006.

Yan, Y.: rBayesianOptimization: Bayesian Optimization of Hyperparameters, URL https://CRAN.R-project.org/package=rBayesianOptimization, r package version 1.1.0, 2016.

Yao, H., Li, X., Pang, H., Sheng, L., and Wang, W.: Application of random forest algorithm in hail forecasting over Shandong Peninsula, Atmospheric Research, 244, 105 093, https://doi.org/10.1016/j.atmosres.2020.105093, 2020.

Zhou, K., Zheng, Y., Li, B., Dong, W., and Zhang, X.: Forecasting Different Types of Convective Weather: A Deep Learning Approach, Journal of Meteorological Research, 33, 797–809, https://doi.org/10.1007/s13351-019-8162-6, 2019.

Appendix A

Comparison of MeteoSwiss crowdsourced hail reports with other hail observational datasets

Elaborating on Chapter 2.7.2, the following comparisons compare MeteoSwiss crowdsourced hail reports to European Severe Weather Database (ESWD, Dotzek et al., 2009) reports and to measurements from automatic hail sensors. This comparison had been done as a response to reviews during the publication on the MeteoSwiss crowdsourced hail reports. We had checked the availability of ESWD, EWOB, mPING and skywarn reports for Switzerland for the time period May 2015 — July 2018. While we received ESWD reports and managed to access the mPING reports, we did not receive any reply to our request for data from EWOB or skywarn. In the entire mPING dataset within the coordinates (5°E, 45°N) and (11°E, 48°N; any time) we found 4 reports done on March 9, 2015 from the MeteoSwiss building in Kloten, Switzerland. These four mPING reports were not made during a hail event and could therefore not be further used.

Independent of the comparison with other observational data sets, there are already signs that point to the quality of the reports. The first sign is visible as soon as the hail reports appear on the MeteoSwiss app animation. The reports appear to reflect typical hail swaths, being located at the centres of heavy precipitation fields and rarely appearing in locations where clouds do not appear. We cannot prove the validity of each single report. However, in the big picture, a large number of reports appear at locations and times that, given the radar-based precipitation fields, are likely true. Furthermore, the results of the comparison with POH and MESHS (e.g. Fig. 2.6), are in fact both an indication of quality of the radar-based hail algorithms and of the reports. If there was no positive correlation between the algorithms and the reports, we would indeed need to assess if it was due to a possibly bad accuracy of the reports or if we would have to redefine the algorithms. With the statistical tests (and notches) we show that the differences in median are statistically significant, which suggests that the likelihood of the positive correlation existing due to chance is very small.

A.1 MeteoSwiss crowdsourced hail reports versus ESWD

We conducted a comparison with ESWD reports which had a quality control of at least QC0+ or higher, to be sure that they were checked for validity. Within the time period May 2015 to August 2018, we found 181 reports within the coordinates (5.7°E, 45.5°N) and (10.6°E, 47.9°N), most of which are located outside of Switzerland (see Fig. A.1) and 108 out of the 181 have an information on the hail size. Reports that do not have information on hail size often have information on hail cover thickness. Some reports indicate the hail size ">2cm". If no hail size is given but the report says "large hail" and if the description indicates that cars were damaged, we also attribute a hail size of ">2cm" to the report.

To match ESWD reports with MeteoSwiss hail reports, we apply the neighbourhood matching method B described in section 2.8.1. When a time uncertainty is given by ESWD, then we alter the temporal search radius to it. To avoid comparing reports with a great uncertainty in location, we considered only the MeteoSwiss crowdsourced hail reports which do not have a manually adapted location. Out of the 108 ESWD reports, we successfully matched 25 with 110 MeteoSwiss crowdsourced hail reports (Fig. A.1).

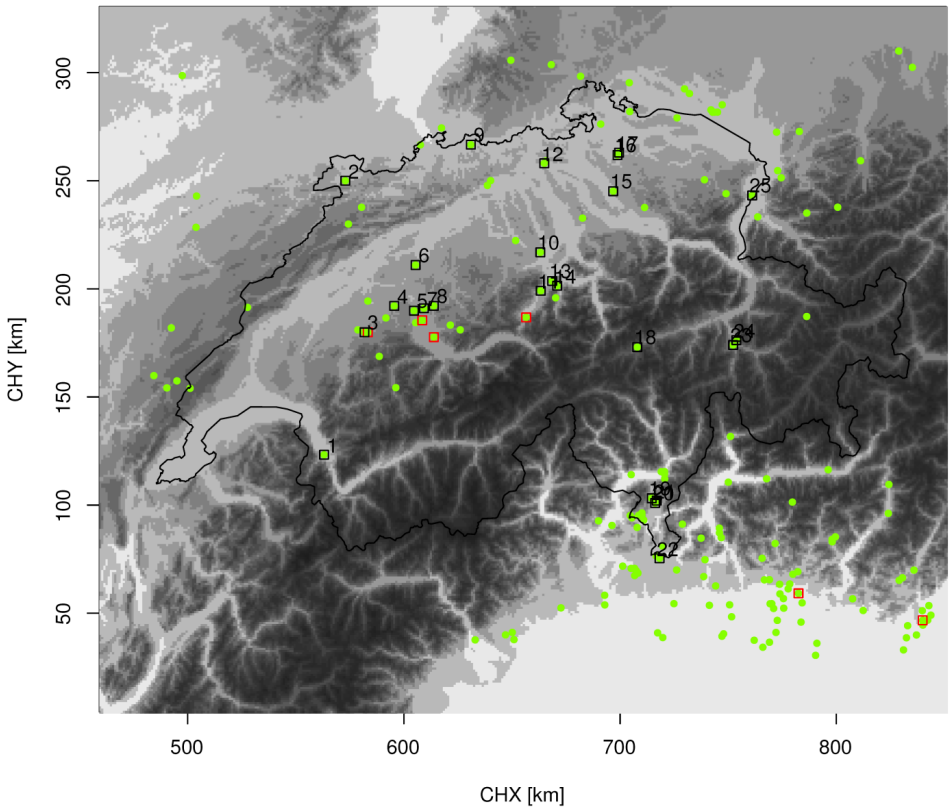

Figure A.1: Map of the research area with all ESWD reports between May 2015 and July 2018 (green dots), the locations of matches with MeteoSwiss crowdsourced hail reports (red squares) and the locations of matches where the ESWD reports has a defined hail size (black squares with numbers at the top right indexing the situations ordered by longitude). The squares have the size 4 x 4 km^2. The grey shading shows the topography.

Figures A.2 and A.3 show for each of the 25 situations POH, MESHS and the crowd-sourced reports. POH and MESHS are included to add some context on the hail activity as estimated by radar. Some ESWD reports are so close in time and space that they are matched to the same MeteoSwiss reports (situations 13 and 14, 16 and 17, 19, 20 and 21, and 23 and 24). The MeteoSwiss reports either agree with the ESWD reports (the majority of the reports) or indicate a smaller hail stone size. For only two ESWD reports (2, 15) the reported size in the MeteoSwiss data is larger than the ESWD report. For all matched situations, within the time uncertainty and within 2 km of each ESWD report, we find at least one MeteoSwiss report that indicates the same range of size as ESWD, except for situations 16, 18, 21 and 22 (Fig. A.3). For locations with several reports (MeteoSwiss or ESWD) we see variation in the hail size estimates. This points to either observational uncertainty or small-scale hail size variability or both. Within the 25 situations, it happens 5 times that a user reports the size "coffee bean" (5mm) and a few minutes later, a few meters away, the size "1 franc coin" (23mm; on Fig. A.2, the MeteoSwiss reports which are connected with a thin black line (see situations 13, 14, 16, 17, 20 and 21) or

which are overlapping (5 and 25)). In these cases, the user apparently observed an increase in hail size within a few minutes. Before April 2018, the users were not specifically instructed to report the maximum hail size. Therefore, the reason for the underestimation in size could be that the MeteoSwiss reports do not indicate the maximum hail size but the average.

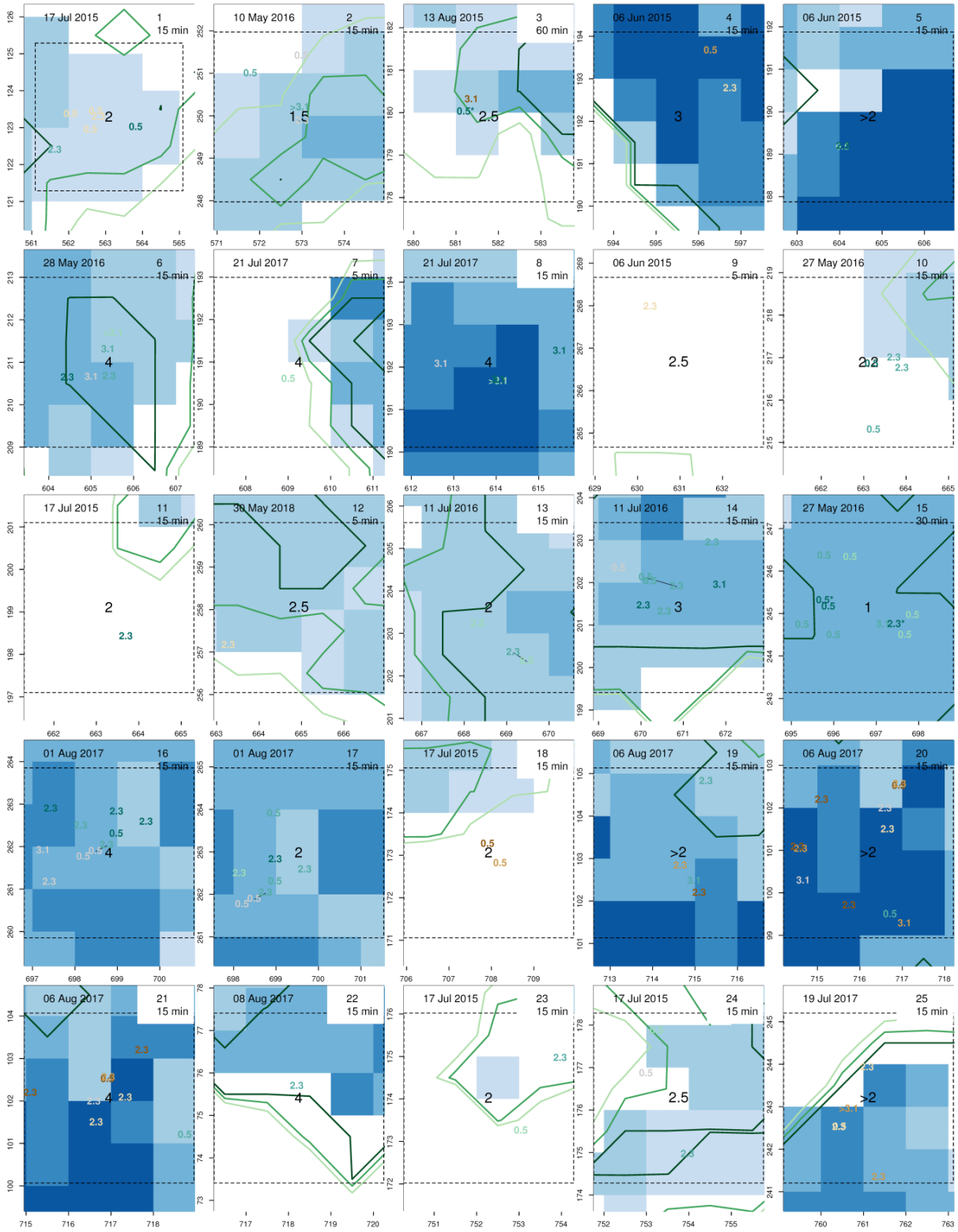

Figure A.2: All 25 situations matching ESWD reports with MeteoSwiss hail reports (numbered squares as in Fig. A.1), the index number, the date and the temporal uncertainty by ESWD is shown at the top of each square. The green contours (POH) and blue shadings (MESHS) have the same legend as in Fig. 2.2. Shown are the maximum POH and MESHS within the matching period (see legend in Fig. 2.2). The squares are centered on the ESWD report and have the size 4 x 4 km^2. The colored numbers show the MeteoSwiss reported size category (in cm). The numbers are colored following the legend in Fig. A.3. Numbers with asterisks indicate a temporal difference between ESWD and MeteoSwiss reports >15 min (situations 3, 15). Reports that were done by the same ID are linked with a thin black line (e.g. situation 14).

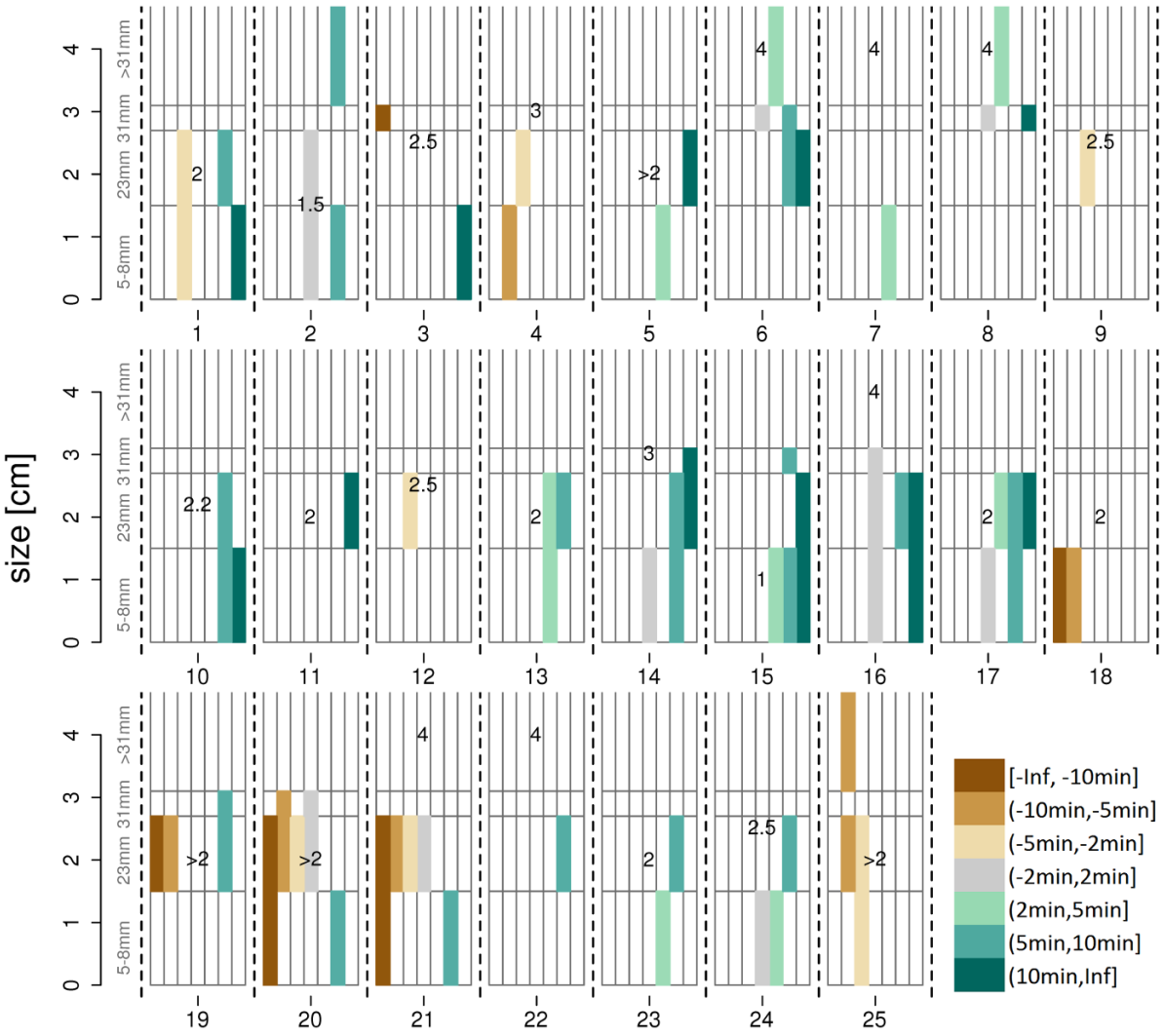

Figure A.3: Same situations as in Fig. A.1 and A.2 but showing the ESWD reported hail size as numbers and the MeteoSwiss crowdsourced hail report categories as size ranges (see Table 2.1; grey grid), colored and ordered according to the time difference to the ESWD report as shown in the legend. Brown (green) colors indicate that the MeteoSwiss reports were done earlier (later) than the ESWD reported time.

A.2 MeteoSwiss crowdsourced hail reports versus automatic hail sensor measurements

The pilot network that existed until 2018 captured five hail and graupel events with between 20 and 50 hail stone impacts and two events with more than 400 impacts. Of the five mentioned events, two did not have any MeteoSwiss hail reports within 2 km. The maximum diameter of 10–12 mm measured for the other three cases were confirmed by MeteoSwiss hail reports that were found within 2 km or the hail sensors. One case had one report of 5–8 mm, another case had nine such reports and one report indicating "no hail" and the third case had 16 5–8 mm reports and one "no hail" report within 2 km of the hail sensor location. These cases were therefore mostly graupel or small hail events.

The two events with more than 400 impacts happened on May 27, 2016 in Aadorf and on July 21, 2017 in Konolfingen. The sensor in Aadorf measured 425 impacts between 18:06 and 18:21 UTC with a mean diameter of 17 mm and a maximum diameter of 27.5 mm. Ninety percent of all hail stones had diameters between 8 mm (5th percentile) and 22.2 mm (95th percentile). During the hail fall at the sensor, three MeteoSwiss hail reports of the size 23 mm were submitted less than 2 km away. Another report of the size 31 mm was submitted within the same time frame at a distance of 2.5 km from the sensor. The sensor in Konolfingen captured 780 impacts between 14:44 and 15:05 UTC. The mean diameter was 9.9 mm, the maximum diameter was 22.6 mm and 90% of all impacts had sizes between 5.2 m (5th percentile) and 13.6 mm (95th percentile). Within the measured time period, 10 (7) MeteoSwiss hail reports were done within 2 km (1 km) of the hail sensor. All reports had the sizes 23 mm or 31 mm, so mostly larger than the size measured by the hail sensor.

With the new hail sensor network that started to be deployed in 2018, we captured one hail event on 6 August 2018 [1] . Fig. A.4 shows a map zooming into the location where automatic hail sensors measured hail on that day. Three out of 15 sensors measured at least 20 impacts of a size between 1 and 2 cm (dark blue dots) and three other sensors measured 1–3 impacts of smaller sizes. The vicinity of several sensor that did not measure hail close to the locations where hail was measured shows nicely how locally hail occurs. Next to the two sensors that measured 20 and 128 impacts, three MeteoSwiss hail reports were submitted. One report indicated a size 23 mm or larger, in accordance with the hail sensor, and two users reported a size smaller than 23 mm. In all areas where the hail sensors measured larger than 1 cm hail, the maximum probability of hail (POH) was between 91–100 %.

These three cases suggest that the MeteoSwiss hail reported diameters tend to be equal or larger, rather than smaller than the diameters measured by the automatic hail sensors. Such a tendency is fortunate/positive, since (Smith and Waldvogel, 1989) found that the maximum size of hail stones within 10 meters of a hail pad tends to be approximately 1 cm larger than the size registered by the hail pad. However, we based this statement on only three hail events, which makes it inconclusive and we therefore did not mention these results in the main publication.

[1] This text was written in January 2019

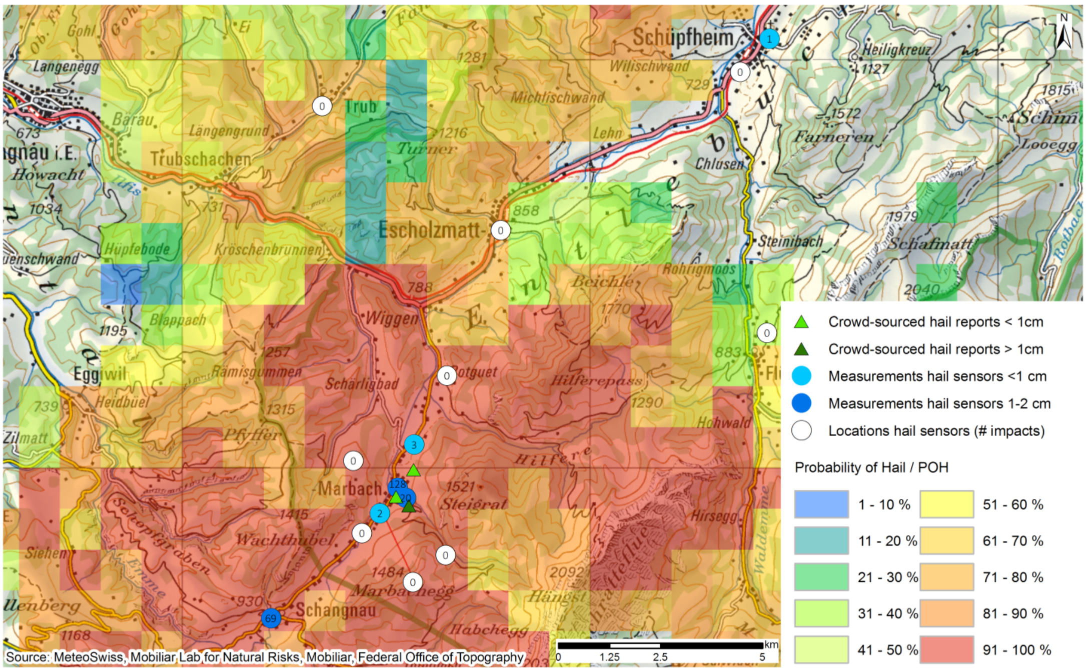

Figure A.4: Map zooming on the locations where automatic hail sensors measured hail on August 6 2018. The circles indicate the locations of hail sensors with the number of impacts. The triangles indicate the MeteoSwiss crowdsourced hail reports done during the event. The colored tiles show the daily maximum POH on that day as explained in the legend. The background map is a 1:200'000 map from the Swiss Federal Office of Topography (one grid-box has the size of 10 km) (Mobiliar Lab for Natural Risks, 2018).

Appendix B

Appendix to chapter 4

B.1 Bayesian Optimisation

The purpose of Bayesian Optimisation (Mockus et al., 1978) in this project is tuning the hyper-parameters by intelligently choosing which hyper-parameters to tune. We assume that the uncertainty of the loss associated with any value of a hyper-parameter can be modelled as a sample of a Gaussian Process. The Gaussian Process is a generalization of the Gaussian probability distribution (Rasmussen and Williams, 2006) and gives a range of likely scenarios of functions that can pass through a range of optionally predefined points in a space of dimensions. In this project, these points are the loss values associated with 15 initial hyper-parameter configurations, with which the models were trained in 15 initial tuning iterations (diamonds in Fig. B.1). The uncertainty of the function, described with confidence intervals, is reduced close to the predefined points and increases with the distance to these points. The configuration of the hyper-parameters to test next is determined by another function, the acquisition function. This function determines which point is most interesting to evaluate, given the information by the previous steps' confidence intervals. Several options to define the acquisition function are explained for example in Snoek et al. (2012) and many example visualizations exist online. The method chosen in this project, the Gaussian Process Upper Confidence Bound, balances exploration and exploitation. Exploration searches for new solutions in unexplored areas of the phase space and exploitation chooses a solution that is expected to be high performing in a promising area with lower uncertainty. The highest value of the acquisition function determines the next point to evaluate. Once the next configuration is evaluated, the Gaussian Process and acquisition functions are fitted again to the new set of known loss points. In this project, 30 subsequent iterations were applied to determine an optimal set up of XGBoost hyper-parameters. Figure B.1 shows the example loss values for tuning a binary XGBoost model predicting POH. The smallest loss during that tuning was found on the 44th iteration, although several other iterations showed a very similar loss value.

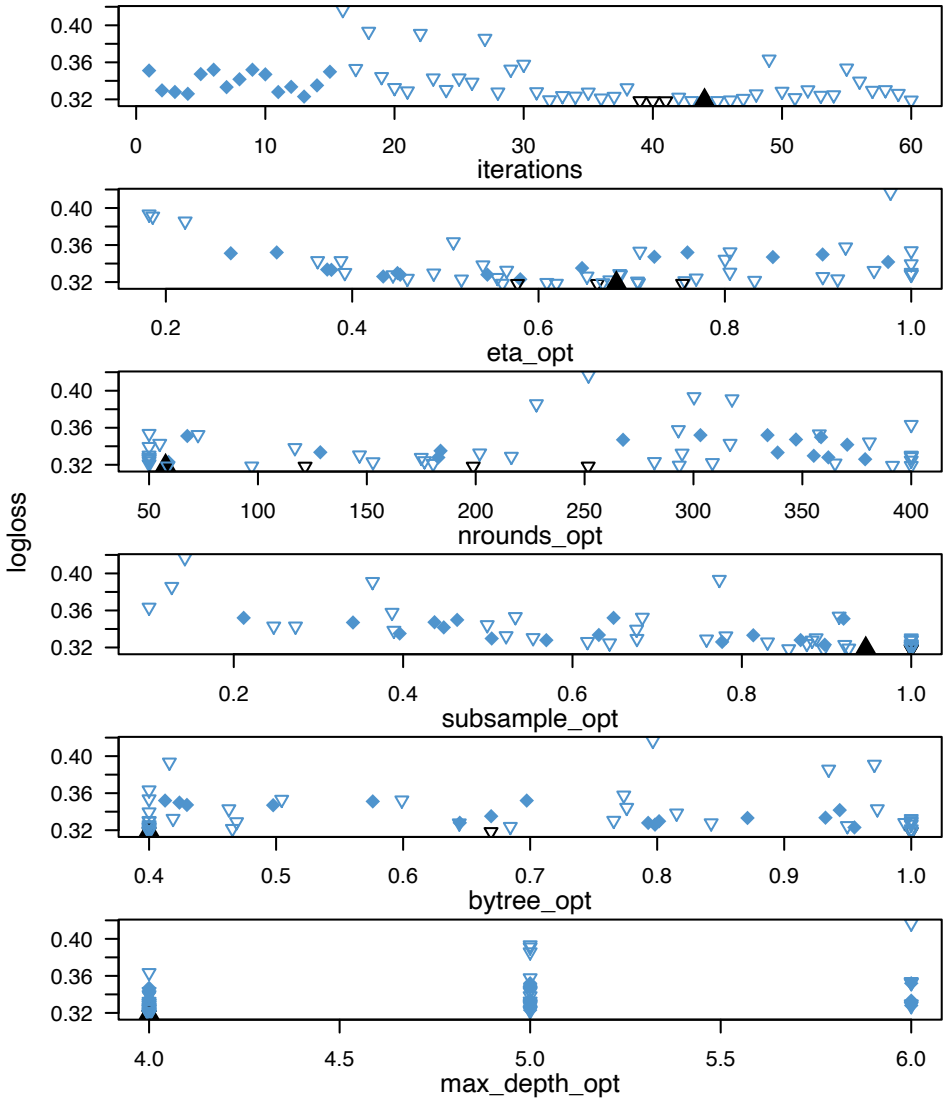

Figure B.1: Logloss for 60 iterations (top) and the tuned hyper-parameters (eta, nrounds, subsample, bytree and max_depth) here as an example for the binary POH model using 1000 features predicting t+5'. The filled diamonds indicate the first 15 iterations, the large filled black triangle is the iteration with the smallest logloss, the other black triangles have a logloss that is almost identical to the black triangle and the remaining blue triangles indicate the logloss for the remaining iterations.

B.2 Verification scores for binary variables

Based on the contingency table (Table B.1), counting the number of matches and mismatches of binary predictions compared to the target, several scores are calculated. In the following equations, the square brackets next to the equation indicate the possible range and in bold the best value of the score.

Table B.1: Contingency table for observed vs. predicted binary hail events

predicted \observed	event (hail)	non-event (no hail)
event (hail)	**hits (a)**	**false alarms (b)**
non-event (no hail)	**misses (c)**	**correct rejections (d)**

The probability of detection (POD) which is also known as the hit rate (H; Wilks, 2011) measures the fraction of correctly predicted events out of all observed events.

$$POD = H = \frac{a}{a+c}; [0, \mathbf{1}] \tag{B.1}$$

The false alarm rate (F) is also known as the probability of false detection (POFD) and determines the fraction of false alarms out of all observed non-events.

$$F = \frac{b}{b+d}; [\mathbf{0}, 1] \tag{B.2}$$

The symmetric extremal dependence index (Ferro and Stephenson, 2011) uses H and F and is a binary performance measure that is suitable for the assessment of strongly skewed binary classifications, such as rare events. SEDI ranges between -1 and 1, with 1 being the best score and 0 meaning that the prediction is as good as predicting with a randomly remixed reference vector.

$$SEDI = \frac{logF - logH - log(1-F) + log(1-H)}{logF + logH + log(1-F) + log(1-H)}; [-1, \mathbf{1}] \tag{B.3}$$

The standard error of the SEDI is estimated as in Ferro and Stephenson (2011) using the following equation. The graphs in the main chapter show the approximate 95 % confidence intervals, which ranges SEDI ± 2*SEDI.se (Ferro and Stephenson, 2011).

$$SEDI.se = \frac{2 * |\frac{(1-H)(1-F)+HF}{(1-H)(1-F)} * \log(F(1-H)) + \frac{2H}{1-H} * \log(H(1-F))|}{H * (\log(F(1-H))) + \log(H(1-F)^2} * \sqrt{\frac{H(1-H)}{p*n}} \tag{B.4}$$

The false alarm ratio (FAR), which should not be confused with the false alarm rate, determines what fraction of predicted events actually did not occur (were false alarms).

$$FAR = \frac{b}{a+b}; [\mathbf{0}, 1] \tag{B.5}$$

The Success Ratio (SR), which is shown on the x-axis of the performance diagram, calculates

which fraction of predicted events were correctly observed.

$$SR = \frac{a}{a+b} = 1 - FAR; [0, \mathbf{1}] \tag{B.6}$$

Another variable shown in the performance diagram is the frequency bias, which is also known as the bias scores. It compares the frequency of observed events to the frequency of predicted events. The perfect score is 1. A value below one indicates that the model has a tendency to underforecast and a value above 1 indicates overforecasting.

$$frequency_bias = \frac{a+b}{a+c} = 1 - FAR; [0, \infty] \tag{B.7}$$

Finally, the critical success index (CSI, also called threat score) measures the fractions of hits compared to the sum of all observed or predicted events (Wilks, 2011).

$$CSI = \frac{a}{a+b+c}; [0, \mathbf{1}] \tag{B.8}$$

B.3 Verification scores for continuous variables

Scores for continuous variables compare the reference vector (O for observed) to the predicted vector (F for forecast). The mean error (ME, also known as the additive bias) measures the average forecast error. A perfect score is zero.

$$ME = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i) \tag{B.9}$$

The mean absolute error (MAE) measures the average magnitude of the forecast error.

$$ME = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i| \tag{B.10}$$

The root mean squared error (RMSE) is similar to the MAE, except that strongly deviating forecasts are penalized more strongly.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2} \tag{B.11}$$

The centered RMSE (cRMSE) compares the difference between debiased observed and debiased predicted values.

$$cRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(y - \overline{y}) - (x - \overline{x})]^2} \tag{B.12}$$

The centered RMSE, the standard deviations (σ_F, σ_O) and the correlation coefficient (CC) are related through equation B.14 and shown in the Taylor diagram (Taylor, 2001).

$$\sigma_F = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - \bar{F})^2}; \sigma_O = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - \bar{O})^2} \tag{B.13}$$

$$CC = \frac{1}{\sigma_F \sigma_O}\frac{1}{N}\sum_{i=1}^{N}(F_i - \bar{F})(O_i - \bar{O}) \tag{B.14}$$

B.4 Evaluation and interpretation of models predicting POH

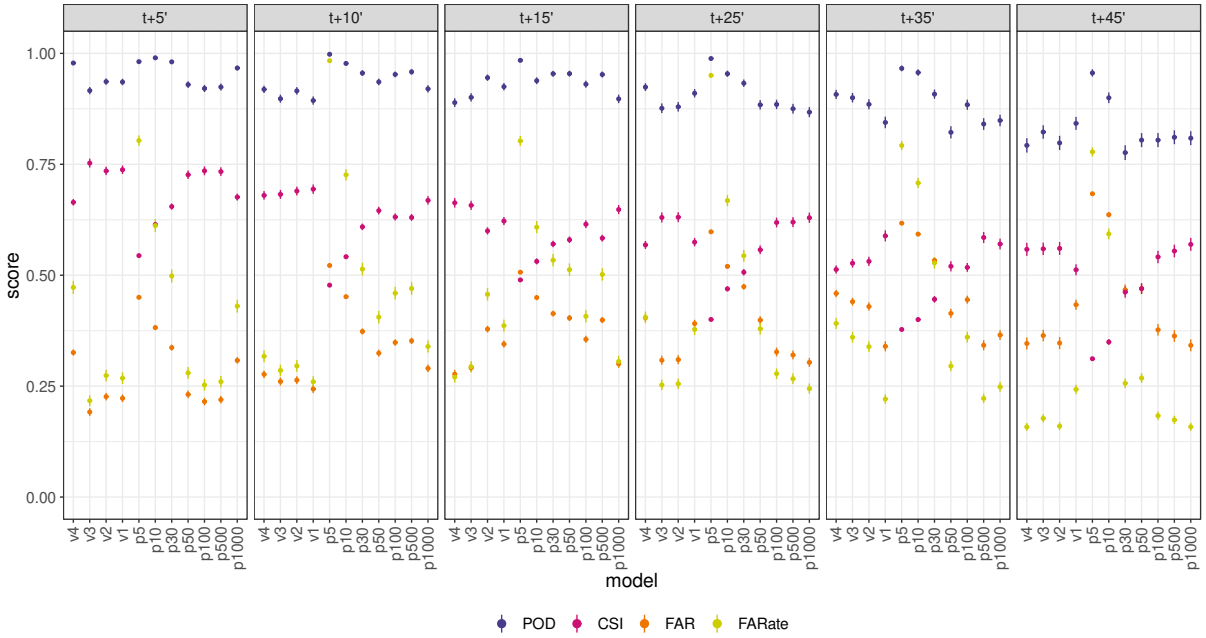

Figure B.2: POD (dark purple), CSI (purple), FAR (1-Success Ratio, orange) and FARate (yellow) and their 95 % confidence intervals (here almost invisible) per each binary XGBoost model and lead-time predicting POH.

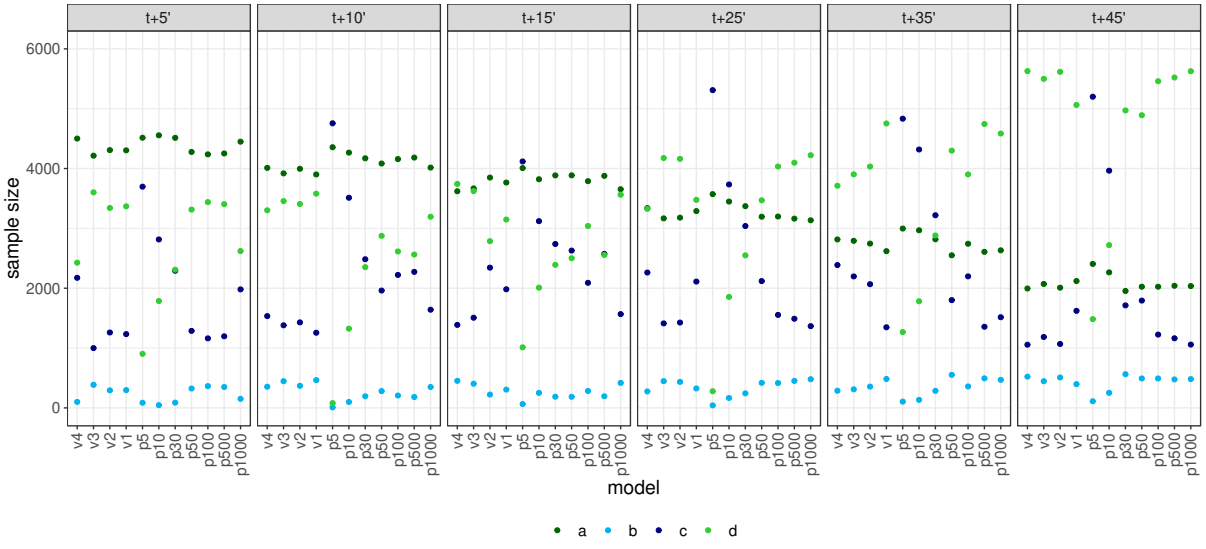

Figure B.3: Number of hits (a), false alarms (b), misses (c) and correct rejections (d) per each binary XGBoost model and lead-time predicting POH.

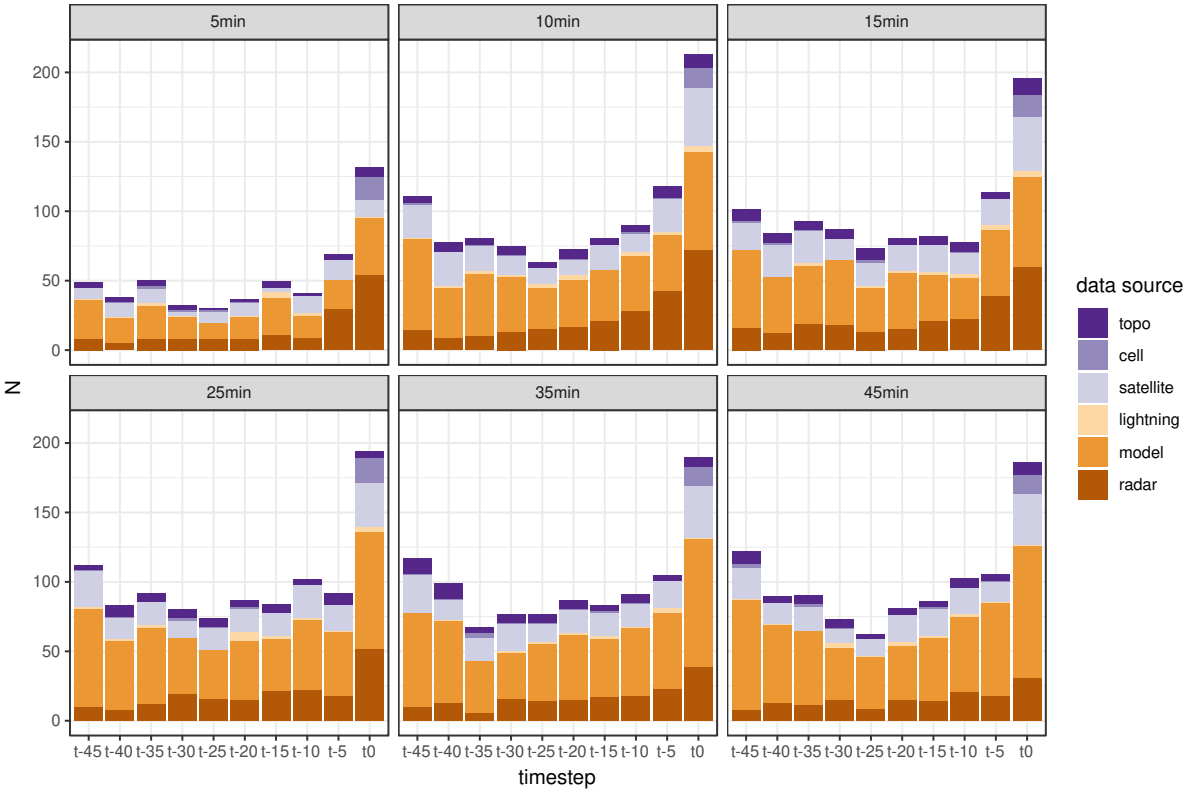

Figure B.4: Number of variables (y-axes) per data source (colors), lead-time (panels) and time step (x-axes) for the binary XGBoost model predicting MESHS with 1000 features (p1000). See Table 4.1) for a list of variables per data source.

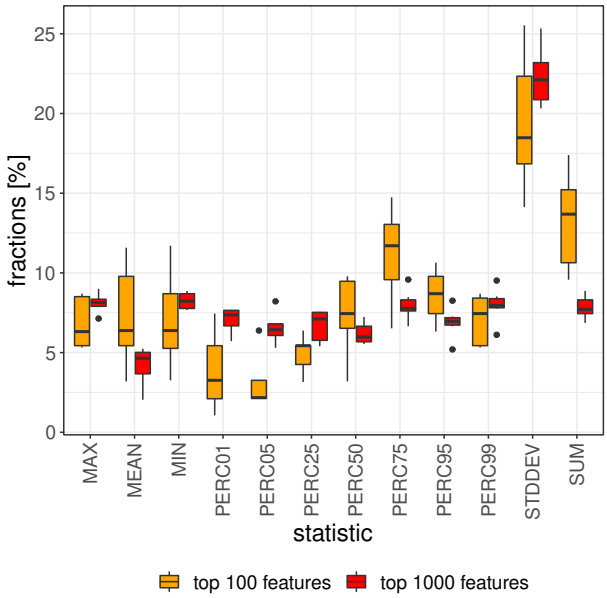

Figure B.5: Fraction per statistic within the top 100 (orange) and top 1000 (red) features used in the p1000 binary XGBoost models predicting POH. The boxplots show the range of fractions of all lead-times.

B.5 Evaluation and interpretation of models predicting MESHS

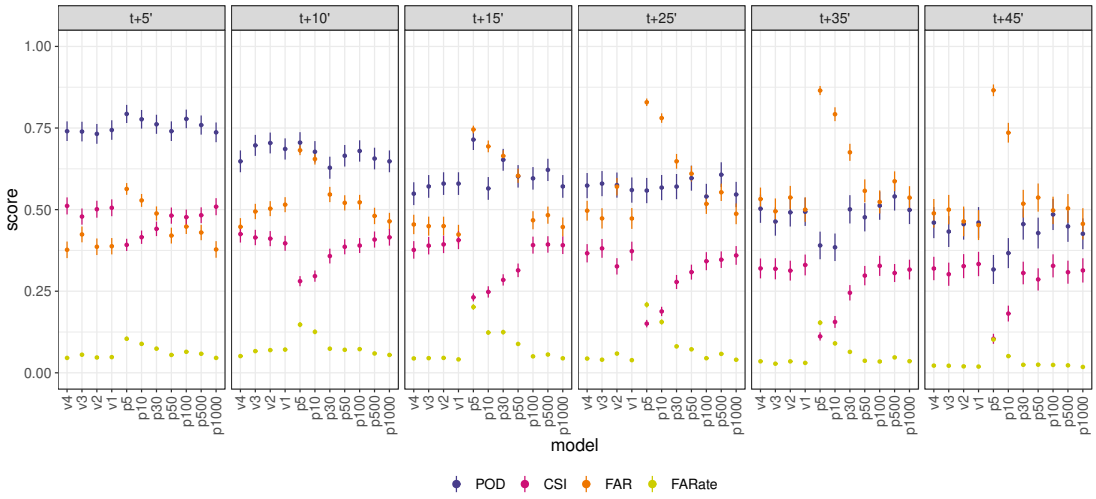

Figure B.6: POD (dark purple), CSI (purple), FAR (1-Success Ratio, orange) and FARate (yellow) and their 95 % confidence intervals (here almost invisible) per each binary XGBoost model and lead-time predicting MESHS.

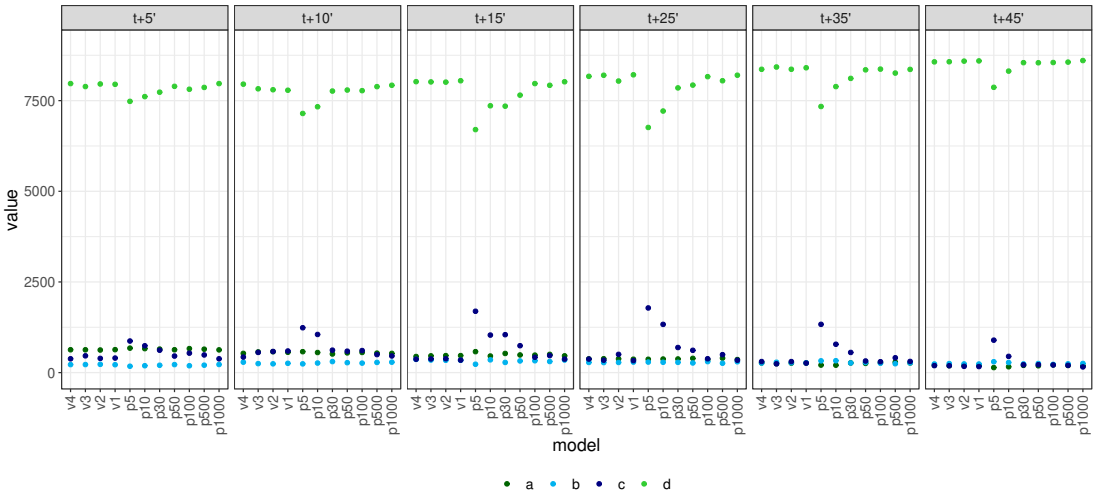

Figure B.7: Number of hits (a), false alarms (b), misses (c) and correct rejections (d) per each binary XGBoost model and lead-time predicting MESHS.

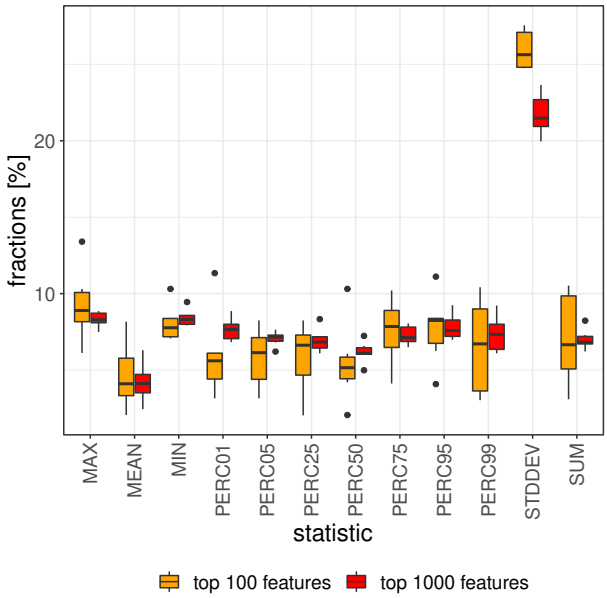

Figure B.8: Fraction per statistic within the top 100 (orange) and top 1000 (red) features used in the p1000 binary XGBoost models predicting MESHS. The boxplots show the range of fractions of all lead-times.

B.6 Additional SHAP summary plots

B.6.1 Binary XGBoost models predicting POH

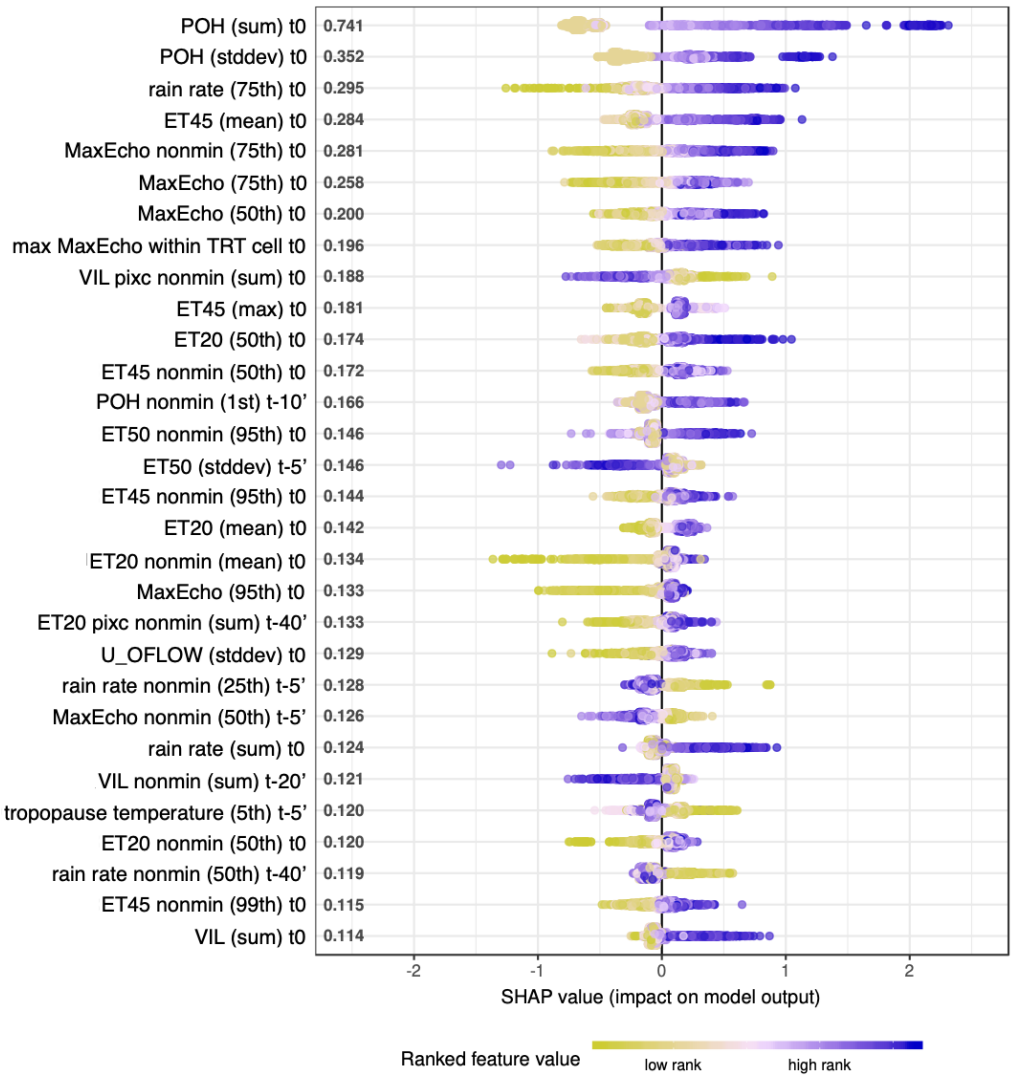

*Figure B.9: SHAP summary plot for the top 30 features of the binary XGBoost model predicting POH at a lead-time of **10 min** using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). See caption of Fig. 4.9 for more details.*

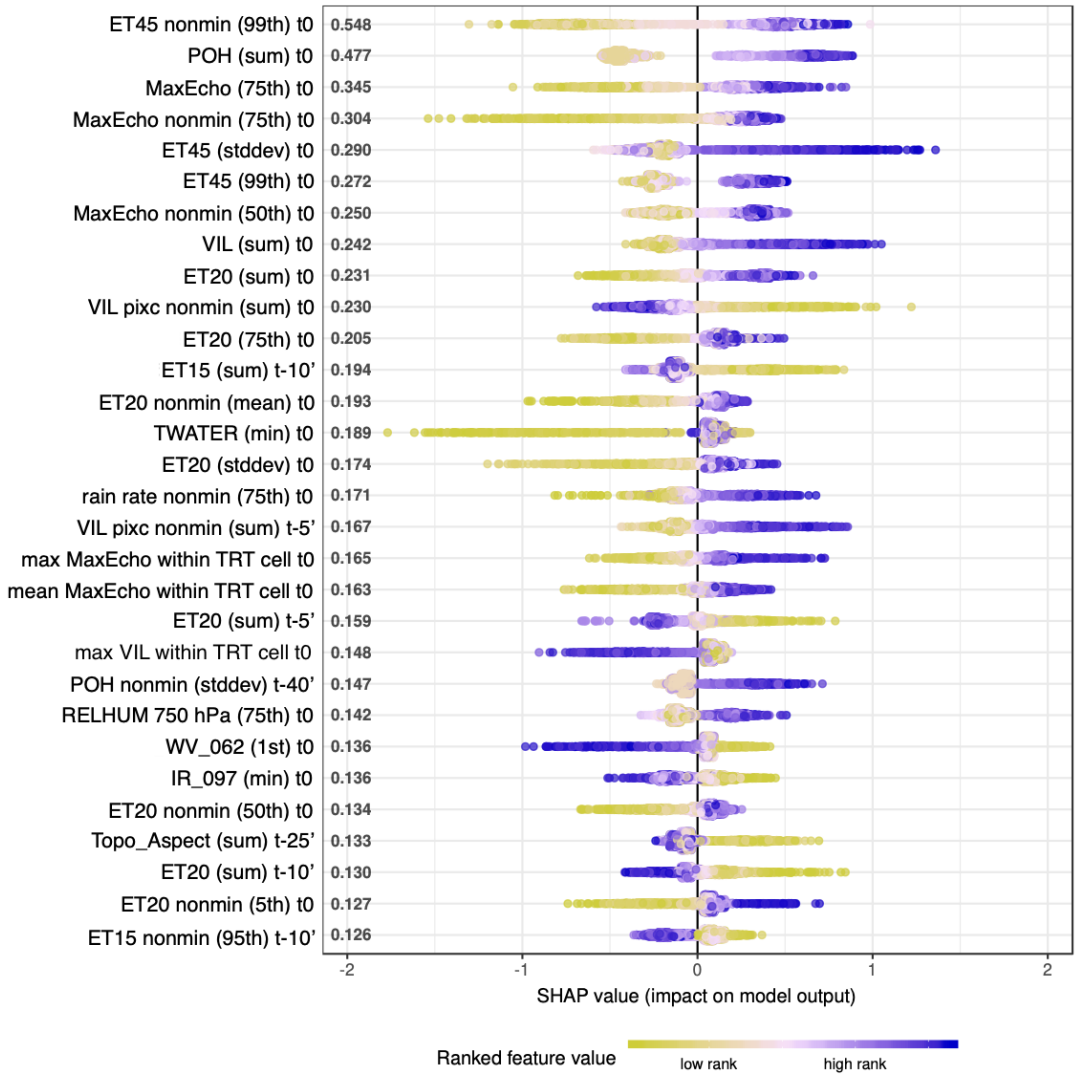

*Figure B.10: SHAP summary plot for the top 30 features of the binary XGBoost model predicting POH at a lead-time of **15 min** using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). See caption of Fig. 4.9 for more details.*

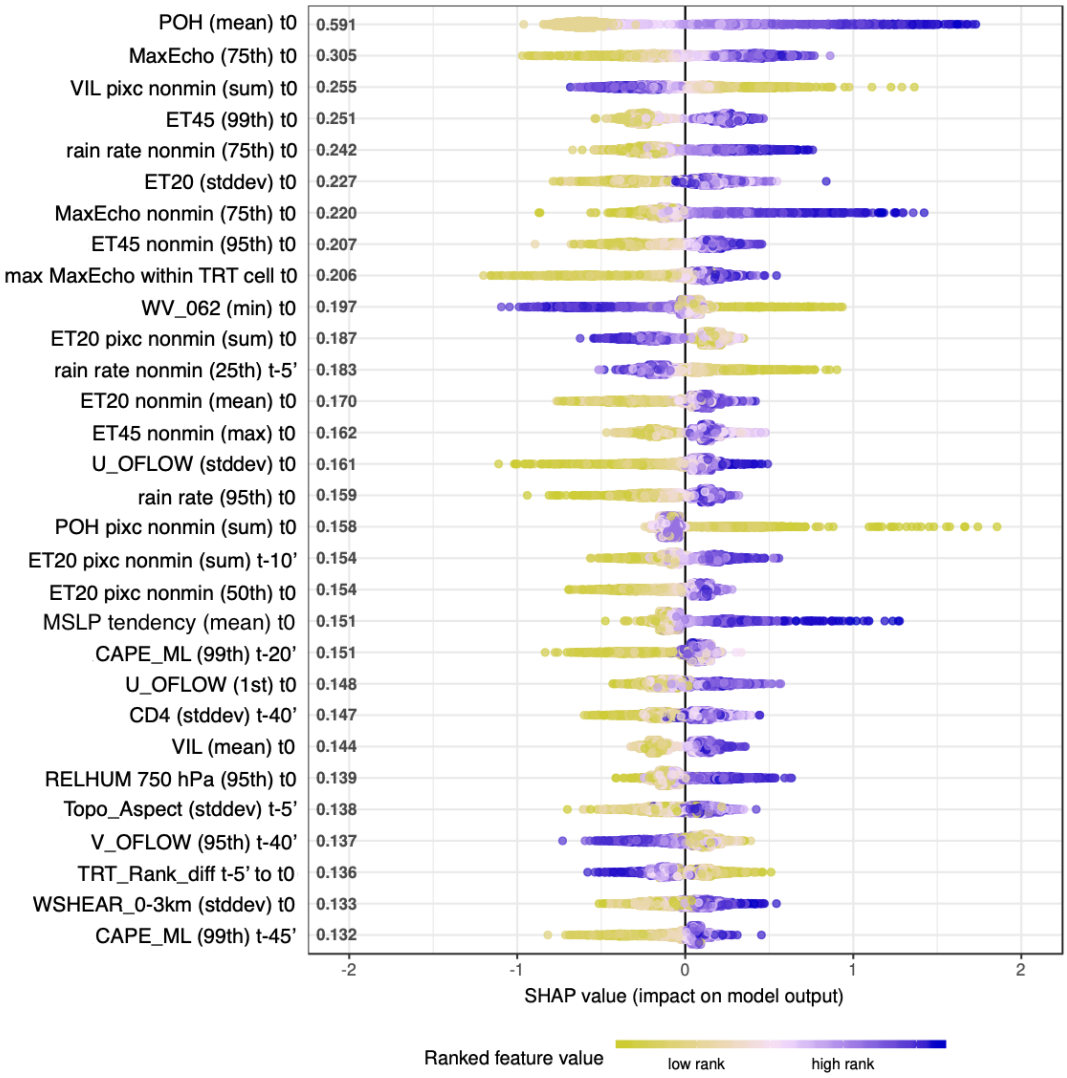

*Figure B.11: SHAP summary plot for the top 30 features of the binary XGBoost model predicting POH at a lead-time of **25 min** using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). See caption of Fig. 4.9 for more details.*

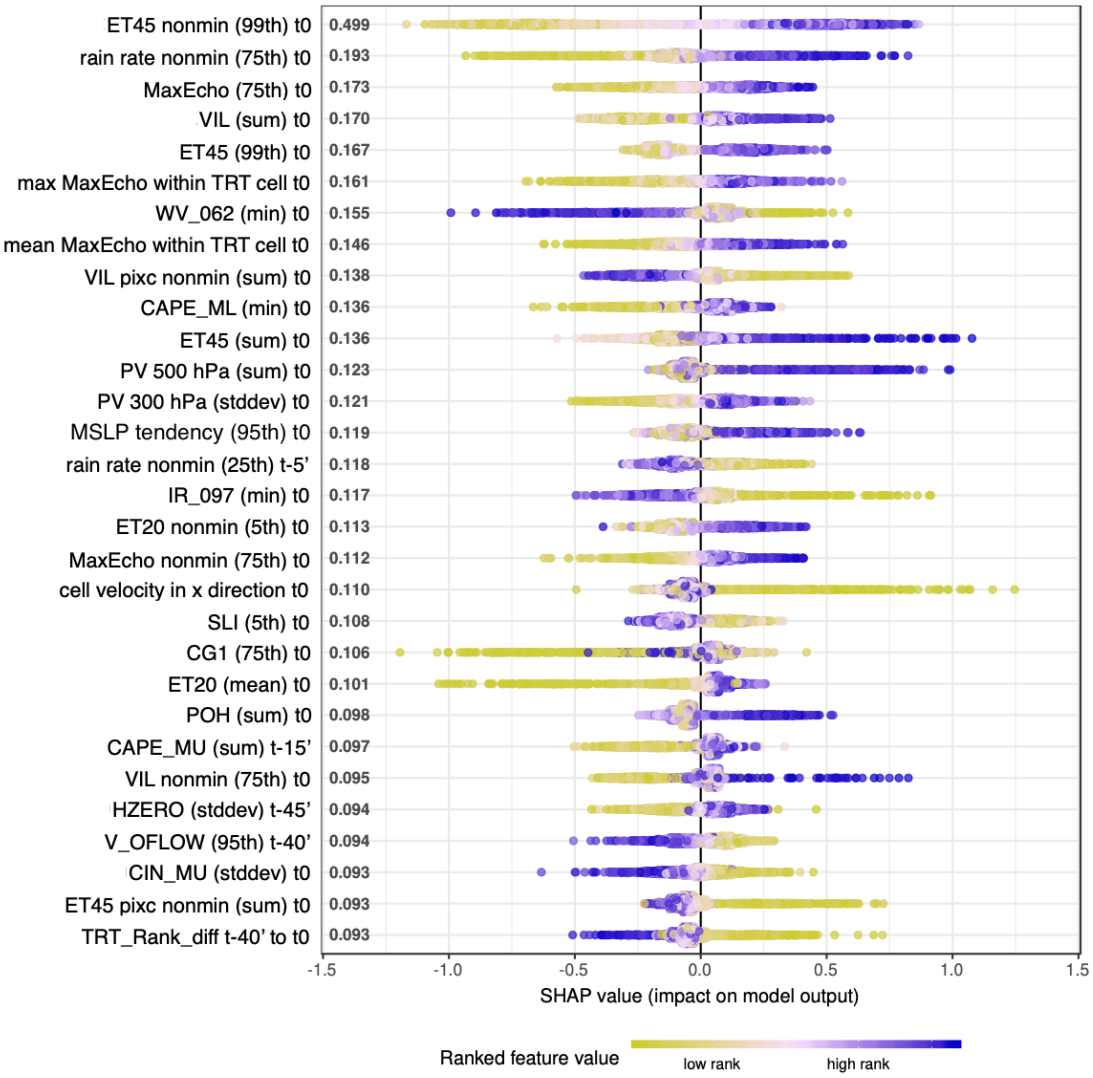

*Figure B.12: SHAP summary plot for the top 30 features of the binary XGBoost model predicting POH at a lead-time of **35 min** using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). See caption of Fig. 4.9 for more details.*

B.6.2 Binary XGBoost models predicting MESHS

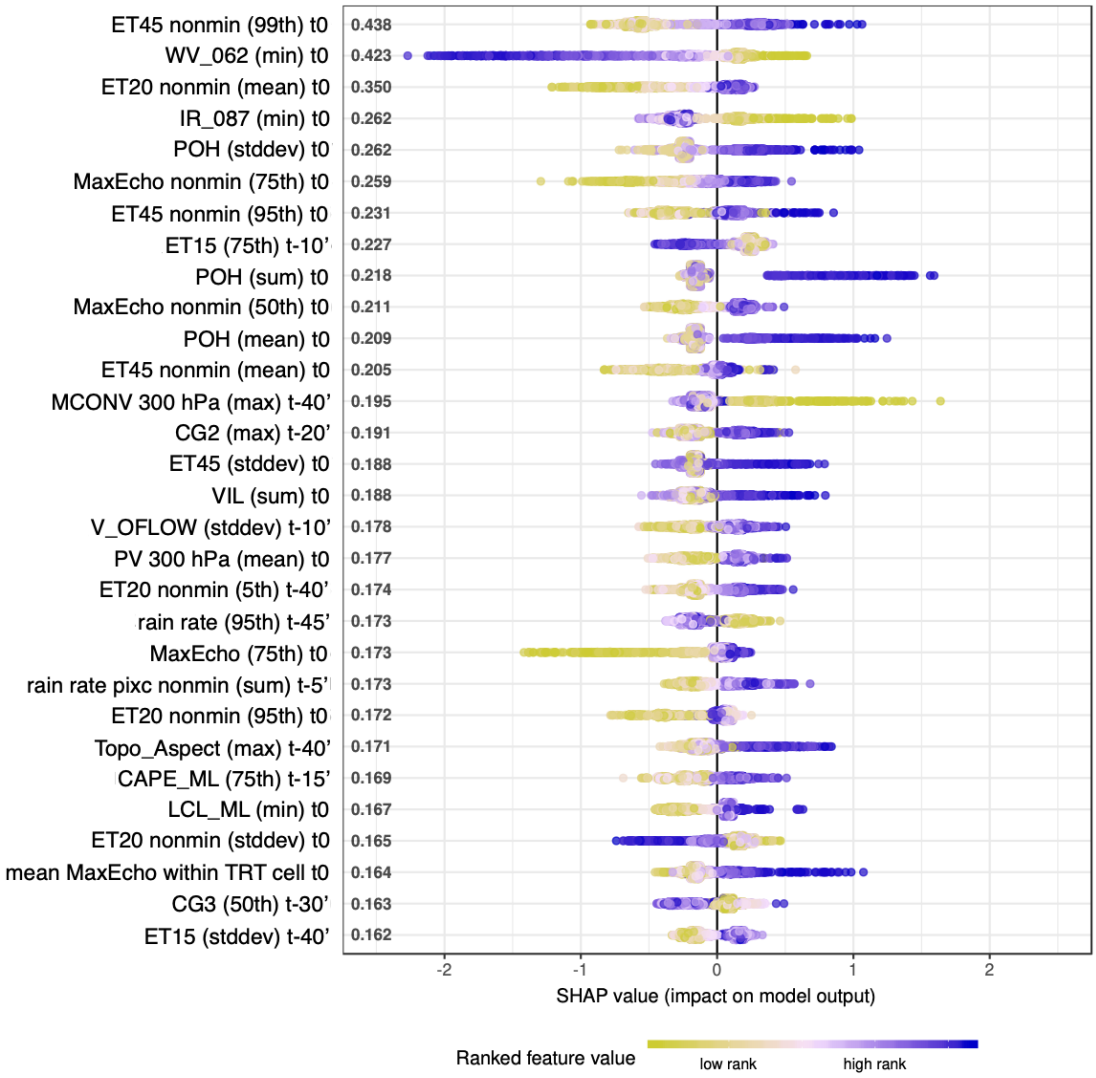

*Figure B.13: SHAP summary plot for the top 30 features of the binary model predicting MESHS at a lead-time of **15 min** using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). See caption of Fig. 4.15 for more details.*

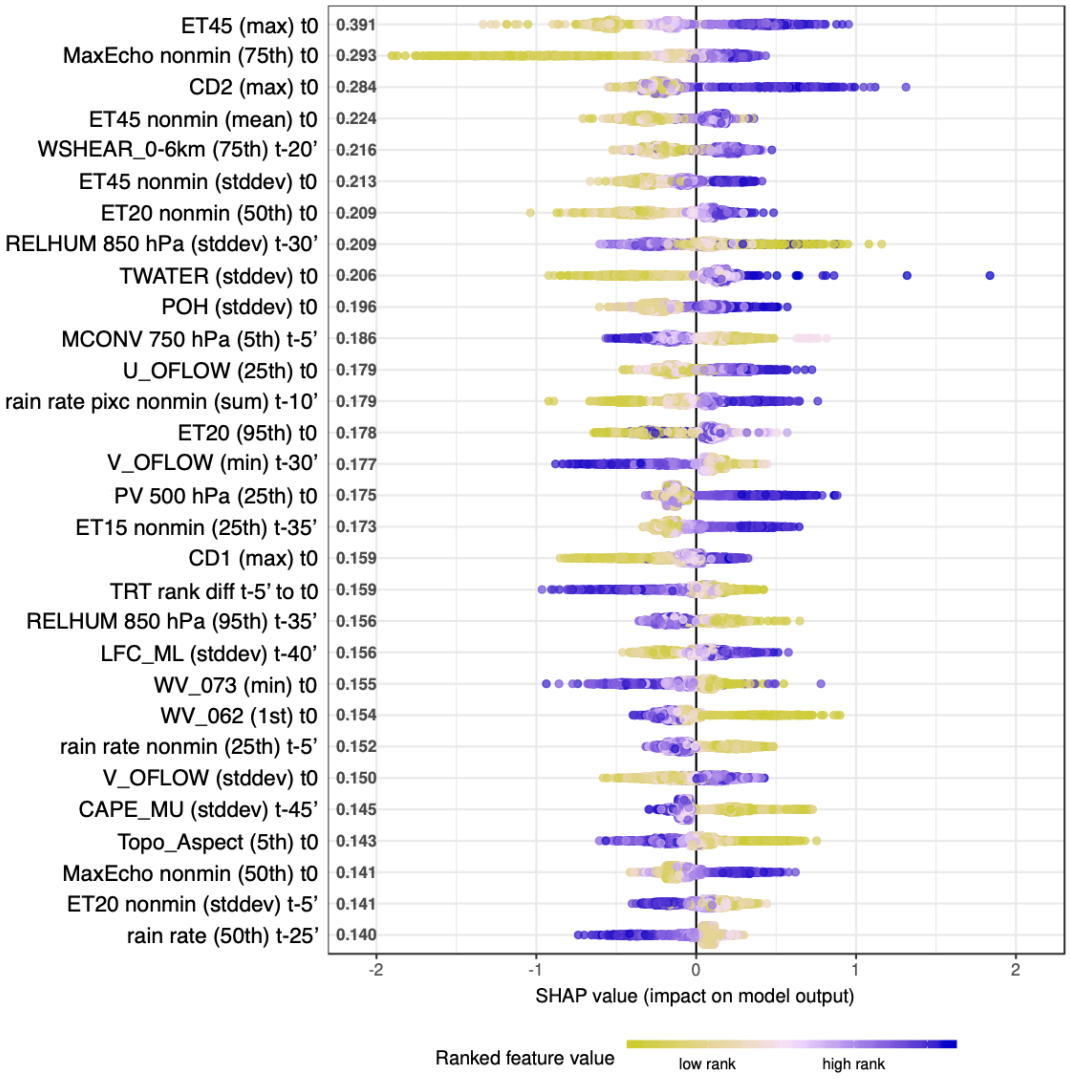

*Figure B.14: SHAP summary plot for the top 30 features of the binary model predicting MESHS at a lead-time of **25 min** using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). See caption of Fig. 4.15 for more details.*

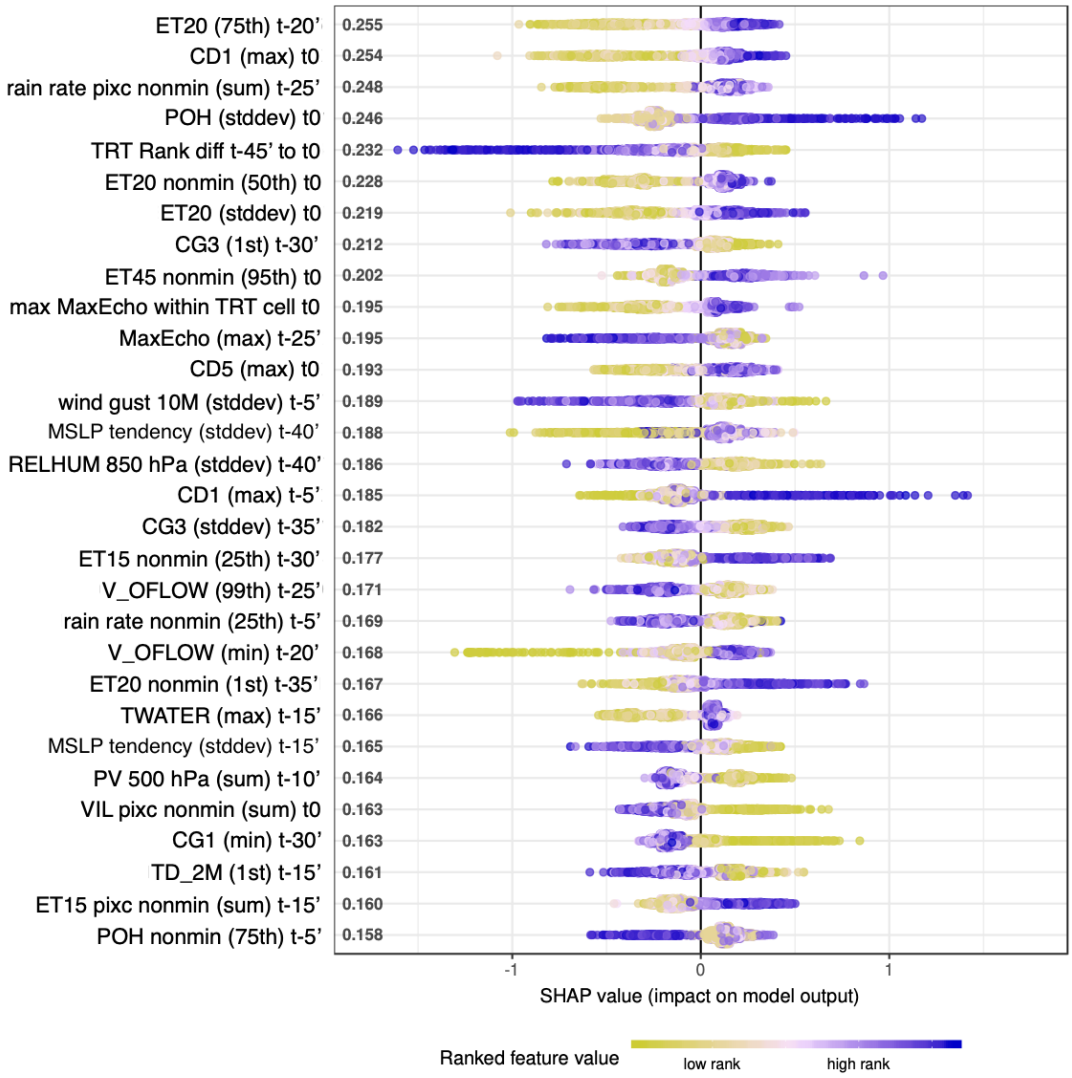

*Figure B.15: SHAP summary plot for the top 30 features of the binary model predicting MESHS at a lead-time of **35 min** using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). See caption of Fig. 4.15 for more details.*

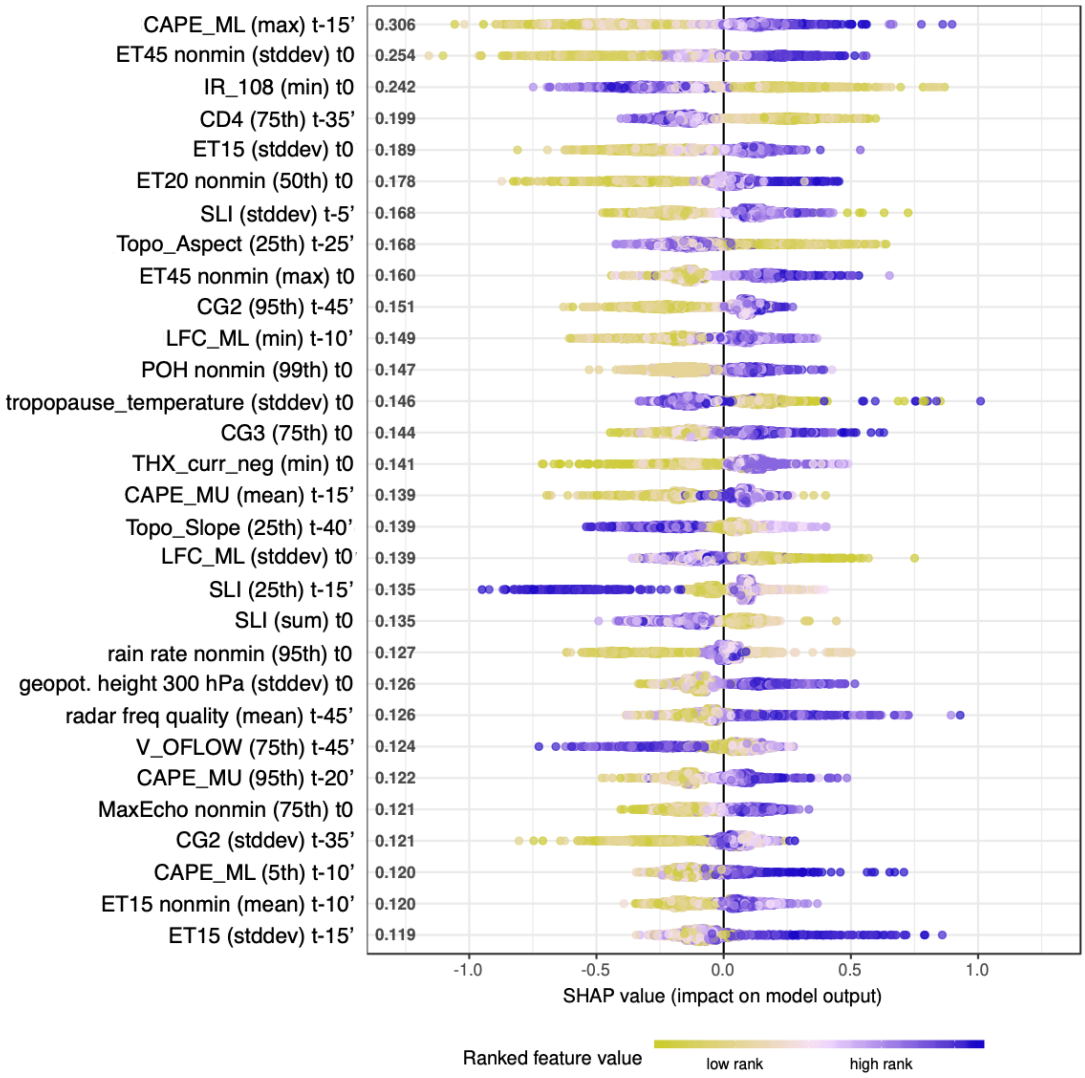

*Figure B.16: SHAP summary plot for the top 30 features of the binary model predicting MESHS at a lead-time of **45 min** using 1000 features. Lime green (dark blue) colors indicate low (high) feature values (colored by rank for visibility). See caption of Fig. 4.15 for more details.*

Appendix C

Appendix to chapter 5

C.1 Selecting the area threshold to define hail days

The method used to choose the area threshold is shown in Table C.1. The number of days on which grid- boxes with POH values $\geq 80\%$ cover at least 200 km^2 are counted for the region north of the Alps (top rows) and the region south of the Alps (bottom rows). Different percentiles (p) of all nonzero POH values $\geq 80\%$ areas are calculated for each region (area in km^2). The days are divided into whether their area value is greater than the percentile (area $\geq$ p) or not (area $<$ p) and whether the days observed at least five car insurance loss reports (≥ 5 losses) or not (< 5 losses). In an ideal case, the columns "≥ 5 losses"- "area $\geq$ p" and "< 5 losses"- "area $<$ p" would have very large numbers of days, and the other two columns would have very low numbers. North of the Alps, 285 days have an area $\geq$ 80th percentile and 192 days an area $<$ 80th percentile. Of the 285 days, 167 have at least 5 losses, more than the 118 days with less than 5 losses. Furthermore, the 192 days include only 23 days with ≥ 5 losses, compared to 96 days with < 5 losses. South of the Alps, the 80th percentile also provides the most balanced numbers of days per category while guaranteeing that more days are defined as "hail days" than "not hail days".

Table C.1: Number of days with POH ≥ 80 % area >200 km^2 per area percentile (p) and number of car insurance loss reports (losses).

	p	Area [km^2]	>5 losses area >p	<5 losses area >p	# hail days	>5 losses area <p	<5 losses area <p	# nonhail days
North	70	164	190	224	414	0	0	0
	75	319	184	184	368	6	40	80
	80	**580**	**167**	**118**	**285**	**23**	**96**	**192**
	85	990	146	72	218	44	152	304
	90	1881	120	37	157	70	187	374
South	70	188	119	203	322	0	0	0
	75	316	108	164	272	11	39	78
	80	**499**	**94**	**123**	**217**	**25**	**80**	**160**
	85	755	76	90	166	43	113	226
	90	1330	55	48	103	64	115	230

C.2 Details on resampling considering the seasonality of clustered hail days

Here we describe the methods for determining which hail days we count as within independent clustered hail day periods. From these, we create the average composites of reanalysis variables during clustered hail days. We then explain how the isolated hail days are resampled following the seasonality of clustered hail days. Furthermore, details of the Kolmogorov-Smirnov (KS) test and the modification of the significance threshold through the false discovery rate (FDR) are explained.

The clustered hail days are by nature dependent. We therefore apply a 500-times-repeated resampling to all clustered and isolated hail days such that each of the 500 series contains only serially independent data. Isolated hail days are by nature independent; this category does not need any additional treatment to ensure independence. However, clusters of hail days that are longer than 11 days are further divided into periods of 5 days that have at least 2 days between each other. We call these periods independent. For the clustering period in 2004 (Fig. 5.2), some clustered hail days have the sequence no hail (0) and hail days (1) "11011101". Although all these hail days are by our definition clustered, the central 5-day period that starts and ends with no hail days "01110" contains only 3 hail days, despite being marked as clustered by their attribution to neighboring 5-day periods. If in such cases the algorithm determining independent periods by accident selects a sequence containing only 3 hail days, that choice is corrected by displacing the 5-day period to one day earlier. Consequently, the number of hail days per clustering period is always ≥4. This criterion of independence has the consequence of not including all potentially available clustered hail days. North and south of the Alps, this treatment additionally removes 29 and 13 out of 164 and 102 clustered hail days, respectively.

The resampling vectors north of the Alps each have 32 days and south 21, following the number of independent clustering periods in each study area. The seasonality following which isolated hail days are sampled is defined as follows. The number of clustered hail days within independently clustered hail day periods is counted per 20-day period as shown in Fig. 3, starting at DOY 80–99 and ending at DOY 260–279. Wherever clustered hail days occur, the relative frequencies of clustered hail days per 20-day period are divided by the relative frequencies of isolated hail days. These values are the probability of sampling per DOY of all isolated hail day during each 20-day period. Because the number of isolated hail days per DOY varies, the probability of sampling per each isolated hail day is further divided by the number of isolated hail days per DOY.

The two-sample KS-test measures the largest distance between the two empirical cumulative distribution functions) of both data samples. It has the advantage of being independent of the distribution of each individual data set. Hence, for each variable, the KS-test is applied 500 times, comparing each resampled series of clustered hail days to its isolated-hail-day counterpart (first resampled vector of clustered hail days with first resampled vector of isolated hail days; second with second, third with third, etc.). For atmospheric fields, this procedure yields 500 ks-values and 500 p-values for each grid-point. The p-value indicates how likely it is that the two compared vectors stem from the same distribution (null hypothesis) and that the difference is statistically insignificant. Statistical tests typically consider a significance level α of 5%. Following this method, the null hypothesis is rejected if the chance of accepting it is less than 5%. However, with N repeated tests and an α of 5%, on average N*0.05 (here $500 * 0.05 = 25$) test results will falsely reject the null hypothesis (Type I error). That number itself is drawn from a probability distribution whose mean is N*0.05 and can vary considerably with an increasing N. A solution is to control the false discovery rate (FDR, Wilks, 2016). The FDR-corrected threshold does not define the probability of falsely accepting the null hypothesis (p-value) but the probability that a rejected null hypothesis should actually have been accepted (q-value). In concrete, we follow the procedure explained in Wilks (2016). The p-values are sorted in ascending order and each individual p-value p_i is compared to a threshold p^*_{FDR} that varies according to q (in Wilks, 2016, q is called α_{FDR}), N and i. Assuming statistical independence of each of the N local tests p^*_{FDR} is the largest p_i that is equal or smaller than $(i/N)\alpha_{FDR}$:

$$p^*_{FDR} = [p_i : p_i \leq (i/N)\alpha_{FDR}] \tag{C.1}$$

Acknowledgements

I would like to finish my dissertation by expressing my gratitude to several people, without whom this thesis would neither have been possible nor as enjoyable an experience.

My deepest gratitude goes to Olivia Romppainen-Martius, for accepting me as a doctoral student and guiding me throughout this thesis. With you, I have experienced a working atmosphere that is open, motivating, comfortably ambitious and trustful. Thank you for having the patience of reading and commenting all my texts.

Many thanks go to Alessandro Hering and Urs Germann for their valuable advice, for providing data and their knowledge as radar experts. You welcomed me warmly for several months in Locarno-Monti and made me an integral member of the Radar, Satellite and Nowcasting Division. Grazie mille. I am very grateful to Ulrich Hamann for his invaluable advice and support in developing and writing the machine learning chapter. Thanks go to Joël Zeder as well, for programming the entire data retrieval and preprocessing.

Special thanks go to Russ Schumacher for reading and grading this doctoral thesis. I hope we can meet in person some day.

I am very happy to thank the members of the Mobiliar Lab for Natural Risks, the Climate Impacts Research group and the colleagues at MeteoSwiss for showing research on other natural hazards, the coffee breaks, lunches and the many stimulating and insightful discussions. I would like to especially thank Milka Nicolic for keeping my office clean and for the refreshing chats shared throughout these years.

Great thanks go to all my friends for their support and for sharing fun activities. Particular thanks go to Leonie Bernet and Stéphanie Arcusa for also proofreading parts of this thesis.

Ich möchte herzlich Lucile und Jascha für die schönen zwei Jahre in den gemeinsamen Wohnungen danken. Ich habe die Zeit mit euch sehr genossen. Je souhaite aussi remercier toute ma famille et surtout mes parents, pour tous vos encouragements, votre soutien et la force mentale que vous m'avez donné pendant toute cette période de doctorat. Finally, a million thanks go to you, Lize, for your love, your silly but amazing humor, your encouraging, steady presence and for showing me another level of beauty and happiness in life.

Declaration

under Art. 28 Para. 2 RSL 05

Last, first name:	Barras, Hélène
Matriculation number:	10-060-739
Programme:	PhD. in Climate Sciences
	Bachelor ☐ Master ☐ Dissertation ☒
Thesis title:	Titel of this Thesis
Thesis supervisors:	Prof. Dr. Olivia Romppainen-Martius

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person, except where due acknowledgement has been made in the text. In accordance with academic rules and ethical conduct, I have fully cited and referenced all material and results that are not original to this work. I am well aware of the fact that, on the basis of Article 36 Paragraph 1 Letter o of the University Law of 5 September 1996, the Senate is entitled to deny the title awarded on the basis of this work if proven otherwise.

Bern, June 29, 2021