NHESS-2024-371

Reply document to referee comments 3 (RC3)

A modelled multi-decadal hailday time series for Switzerland

Lena Wilhelm, Cornelia Schwierz, Katharina Schröer, Mateusz Taszarek, and Olivia Martius

We thank the reviewer for their useful comments and the positive feedback. By addressing these comments, we can better clarify some crucial points and substantially improve the quality of the manuscript. The suggested changes are addressed in the following document. The comments of the reviewer are shown in black and our replies in blue. We number the specific comments throughout the document (comment 1 = C1, etc.). Removed parts are ~~crossed out~~ and new additions are in *italic*. All line numbers refer to the originally submitted manuscript. Technical corrections are changed directly in the manuscript.

Specific Comments:

C1: The use of "Switzerland" in the title is not consistent with the study area discussed in the text. The radar data used in the analysis cover Switzerland, and parts of France, Italy, Germany and Austria. I would suggest saying "in the vicinity of the European Alps" or something similar.

We acknowledge that this might be somewhat misleading. However, the expression 'in the vicinity of the European Alps' or other demarcations inclusive of adjacent countries are too broad and encompass more than what falls within our study area, which is the operational weather radar network of Switzerland. We do not want to disappoint future readers and want to refrain from listing countries in the title. Therefore, we prefer to retain the current title.

C2: Along the same lines as (1), it is not clear how the central Alps region, which is excluded from the study area, is defined. Some more information and motivation are required here on how the boundary was determined.

Considering regions north and south of the Alps is fundamentally motivated by the major climatic divide the Alps present to the climatological regimes. Previous climatological analyses of the hail frequency in Switzerland (Nisi et al. 2016, 2018, Schröer et al. 2023) show that hail occurrence is very rare and radar quality impaired (Feldmann et al. 2021) in the inner Alpine region. Excluding the central inner Alps avoids biases from these effects and allows for a clear separation of the climatic preconditions north and south of the Alps. The delineation of the central vs. the northern and southern Alpine regions, respectively, is based on the official prognosis regions provided by the Federal Office of Meteorology and Climatology MeteoSwiss. To clarify the motivation and delineation of the regions, we added the following information *(L114ff):*

The Central Alps, delineated from the northern and southern Alps based on the boundaries of the official prognosis regions from the Federal Office of Meteorology and Climatology MeteoSwiss, are excluded from the analysis, as hail rarely occurs there and radar quality may be lower (Feldmann et al., 2021). The study regions so allow for a clear separation of the climatological regimes north and south of the Alps and correspond to the ones in Barras et al. (2021).

C3: Page 2, lines 32-34: Can the authors please briefly elaborate on how the interannual variability differs between areas north and south of the central Alps.

We see that years with many or few hail days are not the same for both regions. Additionally, the standard deviation is lower in the southern region compared to the northern region. We also see differences in the intra-annual variability, namely an earlier peak of the convective season in the north than in the south. Further details about these variations in the temporal distribution of Swiss hail are mentioned in section 3.1 lines 170-184 and in Fig. 2. These differences and previous research (Nisi et al. 2016, Barras et al. 2023) suggest that some weather situations that favor the development of hailstorms in the North are not the same as for the South.

We changed lines 32-34 to:

Swiss hail occurrence exhibits strong year-to-year variability and follows a pronounced seasonal cycle (Schröer et al., 2023). Recent studies (Barras et al., 2021; Nisi et al., 2018, 2020; Schröer et al., 2023) have highlighted substantial differences in both intra- and inter-annual hail variability between the northern and southern sides of the Alps. In the northern region, the peak of the convective season typically occurs in June, whereas in the south, it is observed in July (see Fig. 2). Moreover, the occurrence of hail-prone or hail-sparse years differs between these regions. Understanding the drivers behind this *inter-annual variability and the seasonality is crucial for developing potential adaptation strategies, particularly concerning agricultural hail losses.*

C4: Figure 1: I would suggest adding borders for surrounding countries for clarity and using either a different colour or bolder line for the Swiss border.

We thank the reviewer for this suggestion. Figure 1 will be adjusted accordingly in the revised manuscript.

C5: Pages 2-3, lines 56-60. Please briefly discuss what is meant by “local thermodynamic conditions”. Are the authors referring to the mountain-plain circulation. What thermodynamic conditions are unique to the Po valley and not to areas north of the central Alps?

We rewrote lines 56-60:

~~The region north of the Alps is exposed to frontal systems arising from the west (or north), as there is no barrier by a high mountain range like in the southern region (Schemm et al., 2016). The region south of the Alps is influenced by the advection of moist and warm air masses from the southwest or south and is protected from northern air masses by the Alpine chain (Schemm et al., 2016). In addition, Cacciamani et al. (1995) showed, that in the Po valley, hailstorm formation is almost always associated with some synoptic-scale dynamical forcing and not only depends on pure local thermodynamic conditions.~~
The region south of the Alps is influenced by the transport of moist and warm air masses originating from the Adriatic and Mediterranean Seas (Nisi et al. 2016). These conditions, coupled with local wind systems like mountain-plain circulations and valley breezes, create ideal conditions for convective storm development. Previous studies have highlighted the relevance of anabatic – katabatic wind systems in the southern Prealpine region and specifically in the Po Valley for hail formation (Morgan 1973, Gladich 2011). The valley's unique topography, encircled by the Alpine chain and the Apennines, promotes the convergence and rising of air masses, intensifying atmospheric instability and moisture retention. In contrast, areas north of the Alps are influenced by cooler, drier Atlantic air masses and have a more open terrain that allows air to circulate more freely (Trefalt et al. 2018).

C6: Page 4, lines 123-124: Please provide a reference (or references) to support the assertion about data quality from ERA-5 declining before 1959.

We added the following reference:

Bell, B., Hersbach, H., Simmons, A., Berrisford, P., Dahlgren, P., Horányi, A., et al. (2021) The ERA5 global reanalysis: Preliminary extension to 1950. *Q J R Meteorol Soc*, 147(741, 4186–4227. Available from: https://doi.org/10.1002/qj.4174

C7: Section 2.1, “ERA-5 environmental parameters”. It is not clear to me over what areas the ERA-5 data were extracted for each day. For example, were all the parameters calculated using ERA-5 profiles at grid points that were within or close to those areas where the POH was at least 80% selected, and then the average of those values used? Or where the values calculated using profiles at grid points over the entire northern or southern domains on days when the POH spatial criteria were met?

Since we are recreating hail events in the past where we do not have radar data, we cannot select profiles that fall within the POH threshold. The best model performance was achieved by calculating profiles at 12 UTC (pre-storm environment) at grid points over the entire northern or southern domains and then averaging over the respective region. Instead of computing the spatial mean at 12:00 UTC, we explored various other options, such as selecting all convective parameters at the time and location of maximum CAPE, or other extremal statistics. However, these alternatives did not improve the model's performance. Coupled with the fact that the resolution of ERA-5 does not capture all relevant small-scale processes in Switzerland, we opted for using area mean values for modeling hail events. This suggests that the mean values contain sufficient information for a 'simple' decision between hail and no hail, as opposed to determining more detailed parameters such as hail size. This may be attributed to the reduced noise in the aggregated values.

Trefalt (2017) similarly found that for many convective parameters, the separation of mean distributions of hail vs. no hail was sufficient to distinguish between hail and non-hail days and did not necessarily improve when comparing percentile or extreme value distributions.

We rewrote section 2.1.1 lines 126-145 for clarity:

In models classifying hail events, one typically selects the ERA-5 grid point that is temporally and spatially closest to the hail incident. However, for reconstructing past hail events, this is not possible because there is no information available on the exact location of the hail event prior to the start of the observational period. Therefore, to model the occurrence of a hail day, we calculated ERA-5 profiles at grid points across the entire northern or southern domains at 12:00 UTC and then computed spatial averages of the values. The choice of using the value at 12:00 UTC, which exhibited the highest predictive skill, may be attributed to the fact that most storms in Switzerland occur in the afternoon. Thus, this value likely captures the atmospheric conditions prior to storm formation.

Our objective was not to reconstruct every hail event per grid point and time step, but to reconstruct the occurrence of a hail day versus a non-hail day. Our definition of hail days focuses on days with more than just a single hail cell. The thresholds are set to capture events that in the past led to damages and affected somewhat larger areas (Probability of Hail ≥ 80% over a minimum area of 580 km² for the northern domain and 499 km² for the southern domain, as detailed in Section 3.1).

C8: Page 8, line 188: Did the authors mean to say “hail potential” and not “convective potential” here?

Indeed, we are trying to quantify the hail potential and not the convective potential. Thank you for this comment, we interchanged “convective” with “hail” in line 188.

C9: Section 4.1: I may have missed this, but I found myself asking “why not just use previous logistic regression models” for this? Or, “why was it necessary to build your own logistic regression model”? Please elaborate on your reasoning here. Also, if there are other logistic regression models out there for predicting the occurrence of hail, how does you model compare?

Previous logistic regression models have either utilized a lower temporal resolution, do not encompass the entire study area and use an older, non-reprocessed version of POH (e.g., Madonna et al., 2018, Mohr et al. 2015), or relied on reports instead of observational radar proxies as input data for the model (e.g., Battaglioli et al., 2023). Given the intricate topography of Switzerland and the distinct synoptic conditions that influence the occurrence of hail-favorable environments both north and south of the Alps, we aimed to construct separate models, each leveraging the most suitable predictors for its respective region. The importance and coefficients of individual convective parameters vary between the models for the northern and southern regions, indicating that the model is capable of discerning the specific factors driving hail occurrence in each region, while also considering the synoptic context.

The reasoning for building our own models is mentioned in the introduction in lines 87-91:

This paper builds on the work of Madonna et al. (2018), however, here we increase the resolution of the analysis to daily, additionally include the South of Switzerland, and extend the time series back to 1959. *Different to* Battaglioli et al. (2023a) who used ESWD severe weather reports, we use radar data to *train the model of* daily hail occurrence of any size at the ground. We use an ensemble of two statistical models (logistic multiple regression and logistic generalized additive model (GAM)) leveraging the best-fitting predictors for each region individually.

A short comparison of our models to other studies predicting the occurrence of hail is given in section 6.2 lines 598-604:

To gauge the predictive capabilities of the models against those in related studies, we will further on use the performance metrics of the ensemble models. Our models outperformed those mentioned in Raupach et al. (2023a) due to lower FAR and higher HSS values. For the different studies, HSS ranges from 0.1 to 0.4 compared to our models' HSS of 0.73 (north) and 0.35 (south). FAR ranges from 0.57 to 0.8 compared to 0.23 (north) and 0.35 (south) for our models. However, there are also studies showing comparable performance skills, like Battaglioli et al. (2023a) using ESWD hail reports and ERA-5 data, López et al. (2007) using radar and radiosonde data and Gascón et al. (2015) using severe storms reports and WRF vertical profiles.

Due to the manuscript's length, we have opted to limit detailed comparisons. However, we are open to incorporating more thorough comparisons if the reviewer deems it necessary.

C10: Page 11, line291-296. The thinking here regarding OMEGA and updraft size and strength is not correct. By extension, reference to the work by Lin and Kumjian is not justified. OMEGA refers to large-scale synoptic/dynamic lift and does not provide information on the properties of thunderstorm updrafts. OMEGA does, however, provide information where large-scale ascent favours thunderstorm formation and maintenance. Please check for this throughout the paper, incl. Section 3.

We thank the reviewer for pointing this out. We fully agree.

Lines 291-296 are changed to:

The vertically integrated vertical velocity (OMEGA_vint) denotes the vertical motion of air throughout the atmospheric column, primarily reflecting large-scale synoptic uplift or descent. The highest probabilities for hail occur when OMEGA_vint values are negative, signifying large-scale ascent. This atmospheric condition promotes the formation and maintenance of thunderstorms, thereby increasing the likelihood of hail. ~~The vertically integrated vertical velocity (OMEGA_vint) represents the vertical motion of air within the full column of the atmosphere. Negative values of OMEGA_vint indicate upward motion of air, which is crucial for the development of thunder- and hailstorms. Highly negative values of (OMEGA_vint) indicate very strong lifting, potentially with a very strong and narrow updraft. However, we do not see the highest probabilities for hail for those cases, but rather for median OMEGA_vint values (see 3). In the context of hail formation, this could mean that a less intense and wider updraft is more favourable than a very strong and narrow one, where hail embryos could be ejected prematurely, as already modelled by Lin and Kumjian (2022).~~

Lines 369-372 are changed to:

Similar to the vertically integrated vertical velocity OMEGA_vint, the vertical velocity at 500 hPa (w_500) is a measure for the vertical motion of air, here for the layer at 500 hPa. Negative values indicate an upward motion ~~and hence measure updraft strength.~~ The highest positive effect is achieved with the strongest negative vertical velocities.

C11: Page 16, line 347: Did the authors mean to say "variance" and not "deviance"? I'm thinking the authors are referring to the variance explained or the coefficient of determination (i.e., R^2) here? Check elsewhere in the text for this.

Indeed, we changed "deviance" to "variance" throughout.

C12: Page 17, lines 399-403: There may be a logical explanation for this seemingly counterintuitive result. One of the conditions for thunderstorm formation is instability and sufficient static energy (referring to the equation for moist static energy = CpT + Lq +gz). One of the reasons thunderstorms tend to be more prevalent during daylight hours and warmer months is because of the higher temperatures. Given that surface temperatures and the height of the freezing level are positively correlated (see Tables A1 and A2), what we could be seeing is that the model is identifying the freezing level as a proxy for surface temperatures. In that framework, the negative linear correlation between deg0l and hail threat makes sense. That said, it is not clear why the model selected deg0l over t2m.

We highly appreciate the reviewer's insights regarding the interpretation of this predictor. It is also unclear to the authors why t2m was not selected instead of deg0l if the model captures this relationship. It could also be that deg0l variable below 2500 m.a.g.l. is positively correlated with the lower to middle tropospheric moist static energy and that low deg0l values indicate low MSE. Even with such "simple" statistical models, comprehending the entirety of what the model learns and how individual predictors interact in their linear combination remains challenging. While the authors made efforts to understand the specific combinations of variables in the four models, they still find aspects of the models to be opaque.

We added the reviewers idea to lines 399-403:

The highest probabilities of hail are achieved at freezing levels between 2500m.a.g.l. and 3500m.a.g.l.. *Punge et al. (2023) also found that excluding freezing levels above 2400m.a.g.l. (and below 4845 m.a.g.l.) lead to a reduction of false hailstorm detections in higher elevation in South Africa when looking at characteristics of hail hazard based on satellite detection of convective storms.* A negative linear relationship is evident among freezing levels below 2500m.a.g.l., indicating that lower values of deg0l correspond to a reduction in hail probabilities and not an increase. This unambiguous relation has also been seen by Kunz (2007) before. This might suggest that our model does not learn about the melting or growing of the hail embryos from the freezing level, but that it associated with ~~connected to t~~he height of the cloud base and therefore the width of the updraft (Mulholland et al., 2021). The deg0l *may also be a proxy for the surface temperature, as both are positively correlated (see Tables A1 and A2).*

C13: Section 4.3: Can the authors please provide some quantitative information that supports the choice of the ensemble model over the other models in the manuscript. Are the differences sufficiently significant to warrant the use of the ensemble model?

Quantitative information on the predictive skill of the models is given in tables 2 and 5 as well as in section 4.3 lines 447-448. Our reasoning for the ensemble model is to leverage the strengths of each model per region, aiming for the most accurate reconstruction of past hail events. For instance, while the logistic model (LM) in the south demonstrates higher hail event detection (POD of 0.61) compared to the southern GAM model (POD 0.57), it also exhibits more false alarms (0.36 compared to 0.35). Similarly, do the GAMs exhibit a better score in bias (measure for systematic errors in predictions) but perform worse in the area under the receiver operating curve (AUROC, measure for the model´s discrimination ability) compared the LMs. These disparities are attributed to the difference in predictor combinations per model, but also the inherently different nature of the models, e.g. GAMs being adept at capturing non-linear relationships. Overall, the combined ensemble model outperforms individual models across most skill metrics. Moreover, the findings regarding trend and seasonality remain robust across all models (LM, GAM, and ensemble prediction).

As the paper is already very long, we would like to minimize detailed model intercomparisons. We have revised the introduction of the section (refer to response to C14).

C14: Section 4.3, lines 410-413: The authors only very briefly touch on how the ensemble model was produced. Unfortunately, the text on lines 412-413 was not adequate to provide a clear indication of exactly how the model was constructed. Please elaborate.

The term ensemble “model" may be misleading, as it implies the creation of a new model, whereas we utilize the outputs of the two respective models per region to generate a new prediction.

To produce an ensemble prediction from two models, we simply average the predicted probabilities of the logistic regression model (LM) and the Generalized Additive Model (GAM). For instance, if the LM predicts a probability of 30% for hail on a given day and the GAM predicts 70%, the averaged probability would be 50%. Following this, we conducted sensitivity tests to determine the optimal threshold for distinguishing between hail and no hail. Typically, the threshold is set at either 50% or closer to the a-priori probability, which in our case is approximately 15%. Through these tests, we identified the thresholds with the best predictive skill to be 40% for the northern model and 42% for the southern model.

We rewrote lines 410-413 for clarity and changed the title of section 4.3 to “*Ensemble prediction*”:

For the final time series, we create an ensemble prediction out of the best logistic regression model and GAM outputs for each domain. With this we aim to leverage the strengths of each model, thereby reducing residual errors for a "best" time series reconstruction that represents Swiss hailday occurrences. This ensemble prediction is generated by averaging the predicted probabilities from both the logistic regression model (LM) and the Generalized Additive Model (GAM). For instance, if the LM predicts a 30% probability of hail on a given day and the GAM predicts 70%, their averaged probability would be 50%. We again conduct sensitivity tests to determine the most effective thresholds for discriminating between hail and no hail. These thresholds are identified as 40% for the northern model and 42% for the southern model. Overall, the combined ensemble model outperforms individual models across most skill metrics. ~~*For the final time series, we build an ensemble model out of the best logistic regression and GAM for each domain. With this, we aim to even out any leftover residual errors for a "best" final model that represents Swiss hail occurrences. The ensemble was produced by averaging the predicted probabilities for each method for each day and then using the mean threshold of both methods of p(hail) < 0.4 for the northern model and p(hail) < 0.42 for the south.*~~

C15: It is interesting that the SWISS index was not selected despite being developed for this region and having been found previously for correlating well with the occurrence of large hail. Do the authors have any thoughts on why this might be?

In univariate logistic regression, the SWISS index indeed showed promising predictive power in the northern region. However, when multiple variables (4-5) were included in a model simultaneously, the SWISS Index was rarely featured in the best combination models. Notably, its inclusion sometimes led to slightly higher false alarm ratios. It's important to mention that variables well correlated with hail (or haildays) may not always perform optimally in multivariate models, as their importance can change when interacting with other predictors within the linear combination.

Technical Corrections:

All technical corrections are adopted in the revised manuscript. We thank the reviewer for their thorough review and all the suggestions.

L 20: Say, " Addressing the hail hazard…".
L 22: Are you referring to spatiotemporal dimensions?
L 35: Suggest saying, "…is essential for adopting potential adaptation…"
L 40-41: Say, "…, we require a hail time series that is longer than what is currently available.".
L 45: Say, "…ERA-5 is considered one of the most reliable…".
L 48: Say, "…moisture, sufficient vertical wind shear..".
L 55: Suggest saying, "…frontal systems approaching from the west (or north), because unlike over the southern region there is no mountain barrier.
L 79: Suggest saying, "…, but the increases are statistically significant only in the northwest and…".
L 92: POH has not yet been explained. Spell out and provide reference.
L 106: Replace "dBz" with "dBZ".
L 112: Replace "shielding" with "blocking".
L 116-117: Suggest rewording to, "Comparing POH data with car insurance loss data, Nisi et al. (2016) found that a threshold of POH ≥ 80% was best associated with the occurrence of hail on the ground".
L 126: Replace "incident" with "event".
L 148: Replace "reaching" with "extending".
L 150-151: Suggest saying, "Radar-based measurements compliment the archive for more recent periods (i.e., since 2002)."
Page 6, "POH time series". For context, please provide the areas of the northern and southern regions.
L 174: Say, "…, hail over our domain is…".
L 243: Suggest saying, "…model performance is undertaken…".
L 267: Say, "…is connected to environments favouring hail follows".
L 315: suggest saying, "…variables of the model for the northern region on the southern region and vice versa.".
L 577: Say, "…did not present as a skillful predictor".
L 579: Suggest saying, "To address this problem…".
L 585: Say, "Comparison with other studies".
L 614: say, "…, that this study's modelled trends…".
L 653: Suggest replacing "strong" with "severe"?