NHESS-2024-371

Reply document to referee comments 1 (RC1)

A modelled multi-decadal hailday time series for Switzerland

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

The authors thank the reviewer for taking the time to review this manuscript so thoroughly. The constructive feedback and useful comments showed us where we need to clarify points. The suggested changes substantially improve the manuscript, and we addressed all comments in the following document. The comments of the reviewer are shown in black and our replies in blue. We number major comments for referencing purposes throughout the document (comment 1 = C1, etc.). Other comments are addressed by making changes in the manuscript directly. Removed parts are ~~crossed-out~~ and new additions are in *italic*. All line numbers refer to the originally submitted manuscript.

Major comments:

C1: Some of the indices used in this manuscript are quite outdated and have not been used in state-of-the-art meteorological research since the 1970s (e.g., LI, TT, BI, K index). [This list is just a partial list of the ones that are outdated.] These indices are combinations of other variables that are likely in the analysis anyway. (If not, they could easily be replaced with dewpoint at 850 hPa or air temperature at 700 hPa.) This manuscript cannot claim to be using modern methods with such outdated meteorological quantities. I recommend that the authors remove such outdated indices and perform the analysis again. Doing so will improve your paper and give you credibility. This is a potentially state-of-the-art paper, but other meteorologists who read this manuscript will struggle to accept it with such outdated metrics in it.

Doswell and Schultz (2006) have more on the history and inappropriateness of these indices.

Doswell, C. A. III, and D. M. Schultz, 2006: On the use of indices and parameters in forecasting severe storms. Electronic J. Severe Storms Meteor., 1(3), 1–22, https://ejssm.com/ojs/index.php/site/article/view/4.

Thank you for your feedback. We acknowledge your concern regarding the use of certain convective indices, and we appreciate the opportunity to address this issue.

Our study aimed to be comprehensive by including various parameters, both traditional and modern, to statistically model hail occurrence in a region with complex terrain. We also included widely accepted parameters like CAPE, shear, and the individual ingredients of the composite indices (refer to Table A2). Additionally, we tested 132 convective parameters from the thundeR package (including a wide range of CAPE indicators). The model selected the variables independently based on minimizing a loss function. The models presented in our manuscript yielded the best performance.

We agree with Doswell and Schultz (2006) arguing that all diagnostic variables are subject to errors, some of them more volatile than others. We need to stress, however, that our goal was not to forecast individual hail events perfectly, but rather to achieve the best possible reconstruction of hail days over a broad study area in complex topography. We think that the high volatility of “older” parameters is counterbalanced by the combination of multiple variables in the same model, adding to a more complete picture of the atmospheric profile.

While we agree that not all convective parameters in our selection are currently used by forecasters predicting hail storms, we respectfully disagree with the assertion of their outdatedness in literature. We have included citations from the last two decades that that demonstrate the use of these parameters in statistical modeling studies (see Section 4.1, lines 271, 277, 282, and Section 6.2, lines 586-592). We have also transparently discussed the limitations of the selected convective parameters, such as the sensitivity of TT, in Section 6.1. Additionally, recent presentations at the 4th European Hail Workshop (March 5-7, 2024) showed that parameters like TT remain effective in statistical or ML hail modeling studies *when combined with other parameters*, as highlighted by the works of Boris Blanc and Luis Ackermann.

It is crucial to distinguish between the predictive skill of a single variable in a univariate model and the role of an individual variable in a multivariate model. In our models the combination of TT and SLI had great predictive skill, however, if their skill was tested by themselves, they indeed performed worse than e.g. just CAPE. Even with such “simple” statistical models as the ones that we use here, the entirety of which patterns and relationships between parameters the model learns is not fully transparent, especially when composite parameters are in play. In this regard we again agree with Doswell and Schultz (2006).

Lastly, we need to mention that the results concerning trends, drivers of trends, and seasonality are robust to the selection of the model. The following plot shows a comparison of the yearly number of haildays for the study area north of the Alps between our best logistic model (see Sect. 4.1, line 235) with a “simpler” model using only MUCAPE, CIN and wind shear from 0-6km (Fig. 1). The trends in yearly haildays are very similar. Mann-Kendall trend tests show that the tau values are 0.36 and 0.38 comparing the best to the “simple” model, indicating only marginal difference in the strength of the long-term trend.

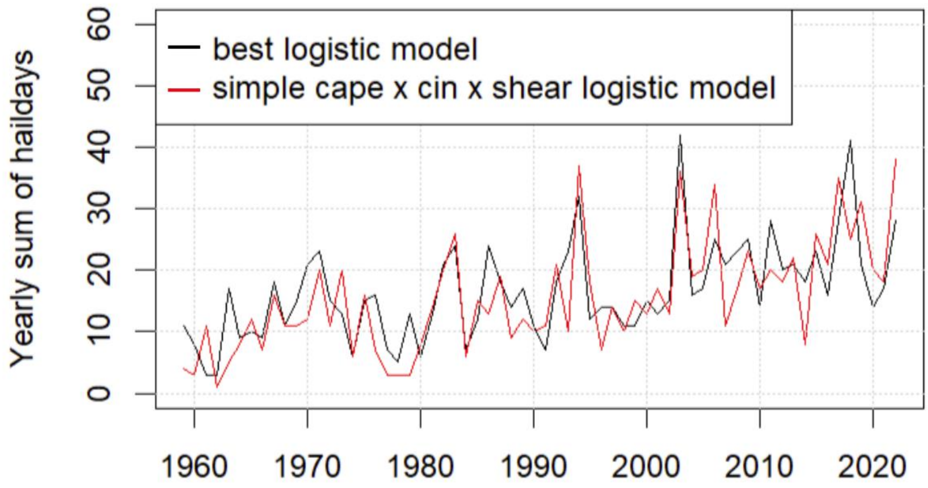

Figure 1: Comparison of the predicted yearly number of hail days in southern Switzerland from 1959 to 2022 from the ensemble prediction (black) vs. a "simple" CAPE, CIN, shear logistic model (red).

Additionally, we compared the yearly number of hail days in southern Switzerland with those in northern Italy as modelled by Battaglioli et al. (2023) (overlapping study area). Despite different models and parameter selection, there was a strong overlap in findings, reinforcing the reliability of our results.

We appreciate your feedback and believe that our comprehensive approach and rigorous validation steps maintain the credibility and relevance of our study. We remain open to incorporating further suggestions and refining our analysis to enhance the robustness and acceptance of our findings.

C2: The authors conflate the various indices, not properly understanding their original intent. For example, at lines 48–49, why “hailstorm”? These ingredients apply to any organized deep moist convective storm (Doswell et al. 1996). The ingredients for hail to be produced from organized deep convective storms are in addition to these. Please clarify. See also line 54.

Doswell, C. A., III, H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients based methodology. Wea. Forecasting, 11, 560– 581.

The authors thank the reviewer for this comment. Indeed, we should differentiate the ingredients necessary for the formation of deep moist convection (DMC) from those needed for hailstone formation.

Lines 48-54 are changed to:

“The development of deep moist convection requires an unstable atmosphere, sufficient moisture at low levels, sufficient vertical wind shear, and an initiation mechanism (Johns and Doswell, 1992; Doswell et al., 1996). For hailstones to form in a storm, three additional ingredients are needed: an embryo particle, typically graupel or frozen drops, an abundance of supercooled liquid water, and sufficient time for the hailstone to grow within the storm’s updraft (Allen et al., 2020; Kumjian and Lombardo, 2020; Kumjian et al., 2021). ~~A combination of multiple mesoscale variables is required to estimate the hail potential of the atmosphere. The development of a hailstorm requires an unstable atmosphere, high lower tropospheric moisture, a sufficient vertical wind shear and a lifting mechanism (e.g. Johns and Doswell (1992); Battaglioli et al. (2023a); Johnson and Sugden (2021); Kumjian and Lombardo (2020); Kumjian et al. (2021). Several studies have shown that both mesoscale and synoptic flow-conditions, including the presence of fronts, and their interaction with complex terrain are significant factors driving hail formation in Europe (e.g. Brooks et al. (2003, 2007); Cacciamani et al. (1995); Costa et al. (2001); Giaiotti et al. (2003); Kunz and Puskeiler (2010); Piasecki et al. (2023)).~~ Regional characteristics such as terrain barriers, local wind systems, and warm water surfaces influence the relative importance of these ingredients necessary for hailstorm development, which is why this study looks at the regions north and south of the Alps separately.”

Furthermore, many of the indices used are not relevant to hail production in storms. Why include them if not relevant?

In our models, we don't specifically differentiate between deep moist convection and hail production; rather, the models learn about both simultaneously. Using low-resolution ERA5 data, indices specific to hail production in storms may not exist, as they would require detailed information on the storm structure and hail embrios. Instead, we focus on variables capturing atmospheric potential for deep moist convection and supplemented with factors essential for large hailstone formation, such as vertical shear and storm organization or information on the freezing level. The model finds a relation to hail day occurrence by using the multi-variate proxy.
We anticipate that ongoing research and access to higher-resolution data in the future could enhance these models.

C3: Line 290–291: OMEGA_vint is the vertical velocity on the scale of the ERA-5 (e.g., 25-km grid spacing). So, it cannot represent the vertical motion on the scale of the convective storms (Doswell and Bosart 2001). And, it is certainly not representing the ejection of hail embryos, as is implied in the manuscript.

Doswell, C.A., Bosart, L.F. (2001). Extratropical Synoptic-Scale Processes and Severe Convection. In: Doswell, C.A. (eds) Severe Convective Storms. Meteorological Monographs. American Meteorological Society, Boston, MA. https://doi.org/10.1007/978-1-935704-06-5_2

We thank the reviewer for commenting on this mistaken interpretation. 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 and primarily reflects large-scale synoptic ascent or descent. In our model, the highest probabilities of hail occur when omega_vint values are negative (Fig. 3c), 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 (omega_500) is a measure for the vertical motion of air, here for the level at 500 hPa. Negative values indicate upward motion ~~and hence measure updraft strength~~. The highest positive effect is achieved with the strongest negative vertical velocities (Fig. 6e)"

C4: I have concerns about how the authors are using the various variables. I don't get a sense that there is an attempt to understand the physical justification for them or even appropriateness of them, as the previous comments have indicated. Instead, the paper is excessively couched in terms of statistics. It appears to be one of the reasons that these outdated convective indices are used in this paper.
As an example of this, I have the following comment regarding text at lines 562–564. If I understand the authors' argument here correctly, they are happy using outdated indices and not understanding why they work, as long as they get a time series they can work with. Do I understand that correctly?

We thank the reviewer for their constructive feedback. We appreciate the opportunity to address the concerns regarding our use of various variables and the perceived emphasis on statistical methods over physical justification.

Regarding the specific text at lines 562-564, we did not intend to imply that we are content with using indices without understanding their mechanisms. Rather, we aimed to highlight that our data-driven approach allowed the models to select the best variables for our specific goals and location, even if those variables included some less common indices. This approach does not negate the need for physical justification; instead, it complements it by leveraging statistical strength to handle data limitations, which is often overlooked.
While we tried to understand each predictor's role (see Fig. 3-6 and Sections 4.1, 4.2, 6.1), there are inherent limitations to what we can infer about the model's learning process. As an example, we are providing additional analyses performed to explore the interactions between model parameters in one model. The heatmaps in Figure 2 show hail day probabilities conditioned on specific combinations of predictors with WS_06km in the southern logistic GAM model, given all other predictors are held at their median value. One can now understand that the southern GAM learns for example a non-linear relationship between CAPE and WS_06km (Fig. 2a). Generally, the highest probabilities of hail day occurrence are achieved with high windshear and median values of CAPE or very high CAPE, negative omega_500 (Fig. 2b), with high TT (Fig. 2c), high Td_2m (Fig. 2d), meaning with increasing instability, moisture and with large scale lifting. Since we cannot grasp more than two to three-dimensional interactions, we cannot fully understand what the model learns from the combination of 5 (composite) predictors.

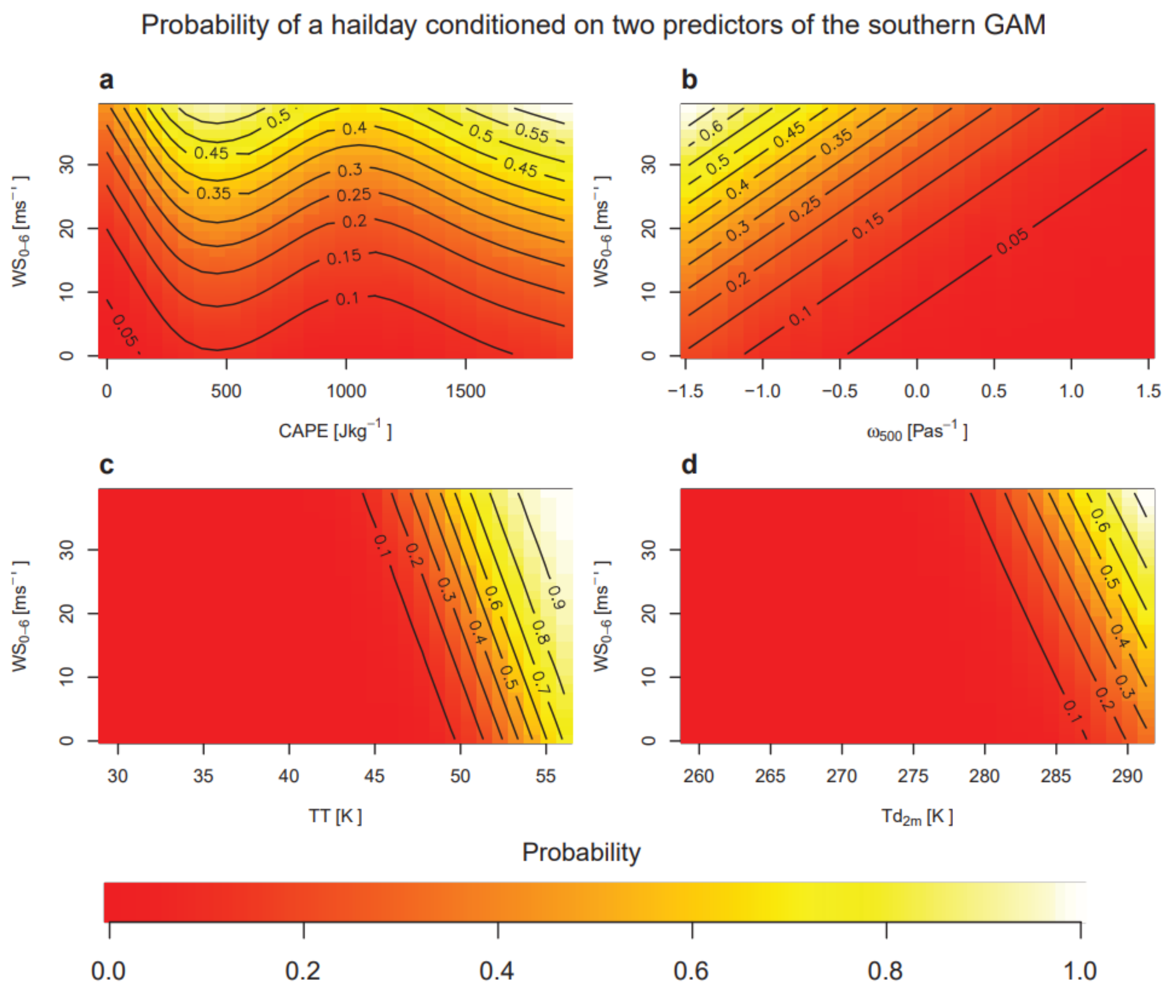

Figure 2: Fitted probability of hail day occurrence conditioned on two predictors of the southern GAM model. The black contour lines and color gradient represent the probability of a hail day, ranging from 0 to 1 (or 0% to 100%).

In this regard we acknowledge that our interpretations are just theories. We have changed lines 562-564 to clarify our argument:

"We do not claim that the combination of LI and TT is better than, for example, CAPE and shear in forecasting individual hail cells or in differentiating between no hail, hail, and large hail. Rather, the specific combinations of around five variables in the statistical models worked best for the reconstruction of hail days in the Swiss study areas using the POH radar proxy and low-resolution ERA5 data. Our data-driven approach identified some less common indices, the statistical models leverage these indices effectively within the constraints of our data, and this statistical approach complements our physical understanding. However, our models should not be transferred to other periods or regions without additional verification. Forecasting applications are much better served by the operational COSMO/ICON weather forecast models than by ERA5.
~~We do not claim that the combination of LI and TT are better than e.g. CAPE and shear in differentiating between no hail, hail and large hail. Rather we want to make clear, that our synergy of around five variables in the 5D models worked best for our exact goal and location, namely, reconstructing Swiss haildays in the last 70 years with the POH radar proxy and low-resolution ERA-5 data. We assume that the models learn how to best deal with the limitations of the data sets and chooses the best variables for our specific questions in a data-driven approach. We did not expect to learn about new processes from the models, but the models seem to agree with previous knowledge. We chose ERA-5, as this is the best available product for multidecadal analyses. That said, one should not transfer our models to the future or to other regions around the globe. For forecasting applications the COSMO / ICON 1x1km analyses are much better suited."~~

C5: There appears to be inconsistency in the messaging about how to interpret the trends. For example, line 618: If "it is not feasible to directly extrapolate our modelled trends into reality", then of what value is this study?
How is this statement consistent with your stated goal of building "a new multidecadal daily hail time series" (line 635)?
Or with, "With this time series we wanted to analyse long-term trends and changes in frequency, seasonality and the variability of model-derived Swiss hailstorms in the past decades." (Lines 640–642)?
Or with, "The final ensemble model reproduces the interannual variability and seasonality of the hail proxies well. The reconstructed time series shows a strong significant positive trend in the number of yearly haildays in both regions from 1959-2022. The trend is also significant and positive when looking at the period of 1979-2022." (Lines 643–645)?
How does the reader interpret the different messages from these different sentences?

Thank you for pointing out the inconsistencies in our messaging about interpreting the trends. We understand the importance of clear communication and will make necessary adjustments.

Our intention was to highlight the inherent uncertainties in using convective parameters from ERA5 data instead of direct observations when analyzing hail trends. Specifically, the variations in trends observed in similar modelling studies presented at the recent European Hail Workshop underscore the need for caution and further investigation before drawing definitive conclusions.

To limit reader confusion, we will remove or adjust the following statements:

Two sentences in Line 617-619 are removed:
~~"Hence, it is not feasible to directly extrapolate our modelled trends into reality. To achieve this, a more extensive period of observational data would be required."~~

Lines 640-642 are changed:
"~~With this time series we wanted to~~ *This time series was used to* analyze long-term trends and changes in frequency, seasonality and the variability of model-derived Swiss hailstorms in the past decades."

Lines 650-653 are adjusted:
"The main purpose of this study is to offer a~~n alternative~~ framework to study intra-annual variability, trends, and changes in the seasonality of Swiss hail occurrence in the past without long-term direct hail observations. ~~Our goal was not to predict every hail event in Switzerland per grid cell, but to reconstruct strong haildays in the past."~~

C6: The paper is in severe need of proofreading, as indicated by the large number of minor comments, grammatical errors, and inconsistencies noted throughout the manuscript. A note to the authors: Many of these issues could have been fixed with more care spent proofreading the manuscript, which I encourage the authors to do for this and subsequent submissions to journals. Peer reviewers should not have to identify these large numbers of issues to improve submissions. Authors should take their own responsibility for the quality and professionalism of their submissions.

We apologize for the numerous errors and inconsistencies in the manuscript and greatly appreciate the reviewer's effort in identifying them. Despite multiple rounds of proofreading, it is evident that we missed several issues due to the manuscript's length. For the revised manuscript, we conducted a thorough proofreading and utilized a professional English editing service before resubmitting.

Other comments:

We thank the reviewer for their comments. The authors have adjusted all the mentioned technical corrections in the revised manuscript.

1. Title: The authors may want to be more specific about what kind of "model" is used and what "multidecadal" means (i.e., what years were used). We agree and have changed the title to: "Reconstructing hail days in Switzerland with statistical models (1959-2022)".

2. Throughout the manuscript: Hyphenate "hail-day" when modifying another word (e.g., "hail-day time series", but "the number of hail days"). Adjusted throughout the manuscript.

3. Line 7 and throughout: Use an en dash to separate two years, not a hyphen and spaces. See also lines 78, 149, 199, and 200, for example. Fix throughout. (Note that you do it correctly at line 79.) Adjusted throughout the manuscript.

4. Line 9: GAM is not defined. Definition is added.

5. Lines 13 and 15: Delete "we can see" and reword the sentence. Same with "we can now study".
Lines 13-15 have been changed to:
"The last two decades show a considerable increase in hail days, which is strongest in May and June. The seasonal cycle has not shifted systematically across decades. This time series allows us to study the local and remote drivers of the interannual variability and seasonality of Swiss hail occurrence."

6. Line 14: "however" cannot be used as a conjunction. See also line 87.
https://www.iup.edu/writingcenter/writing-resources/grammar/common-problems-with-however therefore-and-similar-words.html
Unfortunately, the website link is no longer available, but we acknowledge that "however" cannot be used as a conjunction. See previous comment for changes in line 14. We have changed any further misuse in the manuscript.

7. Line 21: "one of the most complicated meteorological phenomena": I'm not sure that this is appropriate. Please reword. Line 21-22 changed to: "*Addressing hail hazard is challenging, as hail is associated with complex interactions of thunderstorm dynamics with microphysical processes that are modulated by synoptic-scale dynamics."*

8. Lines 31–32: This sentence is a tautology: "With hail hotspots in Switzerland, the Alps are affect by hailstorms in Switzerland." Lines 31-32 shortened to: "*The pre-Alpine regions north and south of the Alps are regularly affected by severe hailstorms (Nisi et al., 2016; Fluck et al., 2021)."*

9. Line 47: Why "mesoscale"? We wrote mesoscale because ERA5 cannot capture the microscale processes. We agree that this sentence is misleading and removed it in line 47 following comment C2.

10. Line 51: I am uncomfortable with the word "driving" here. Synoptic and mesoscale conditions often set the stage for convective storms that may or may not produce hail. So, "associated" would be a more neutral and accurate word to use in this sentence. We decided to remove this short paragraph to improve the reading flow and shorten the manuscripts length.:
Line 51 is changed to:
"~~Several studies have shown that both mesoscale and synoptic flow-conditions, including the presence of fronts, and their interaction with complex terrain are significant factors *influencing* the development of hailstorms in Europe (e.g., Brooks et al. (2003, 2007); Cacciamani et al. (1995); Costa et al. (2001); Giaiotti et al. (2003); Kunz and Puskeiler (2010); Piasecki et al. (2023)).~~"

11. Line 55: Why is "(or north)" in parentheses? Either delete the parentheses or delete the

parenthetical. See also lines 61 and 249. Fix elsewhere in the manuscript.
We removed the parenthesis around "(north)" in lines 55, around "(hail producing)" in line 61 and removed "~~(highly)~~" in line 249. We checked throughout the manuscript.

12. Line 57: Air masses are transported, not advected. Quantities such as temperature can be advected. Write precisely. We thank the reviewer for this suggestion and want to write as clearly as possible. Line 57 is changed following a comment from reviewer 3, but we will use "transported" when talking about air masses:
"*Convection in the region south of the Alps is influenced by the transport of moist and warm air masses originating from the Adriatic and Mediterranean Seas (e.g., Nisi et al., 2016) during southwesterly or southern flow conditions."*

13. Lines 68–71: Except for CAPE, these indices are outdated. We refer to comment C1.

14. Lines 78–79. This sentence is poorly written and unclear. Delete the comma after "found". Lines 78-79 are rewritten:
"For instance, Mohr et al. (2015a) used a logistic regression approach to estimate the potential for hailstorms in Germany between 1971 and 2000 and between 2021 and 2050. They find that the potential for hail events is projected to increase significantly in 2021–2050 compared to 1971–2000 in the northwest and south of Germany. ~~*1971 - 2000 and 2021 - 2050. They found, that the potential for hail events will increase in the future (2021–2050) compared to the past (1971–2000), but only statistically significant in the northwest and south of Germany.*~~"

15. Line 85: "Poisson" should be capitalized. Adjusted.

16. Line 106: Why is EchoTop capitalized? The "Z" in dBZ should be capitalized and italicized, as per convention. Changed throughout the manuscript to "dB*Z*".

17. Lines 108, 122, and 123: Change "resolution" to "grid spacing" because these two terms are not equivalent. We adjusted "*resolution"* to "*grid spacing*" in line 122: "In this work, data on model levels (137 levels from 1000 hPa to 1 hPa, 0.5° x 0.5° *grid spacing*) [...]". In lines 123 and 108, we are referring to the temporal resolution, so we cannot simply exchange the words "temporal resolution" with "*grid spacing*".
We did change the word "grid" to "grid spacing" in lines 122, 123 and 108.

18. Line 114: I don't understand the phrase "area up to 140 km". Units are not consistent. Why up? Reword to be more clear. We agree and rewrote lines 114f:
"We ~~here~~ use thoroughly quality-checked and reprocessed POH data ~~of~~ *from* the recently published Swiss hail climatology (Trefalt et al., 2023; Schröer et al., 2023) ~~and consider the area up to 140km around the five radar stations (see Fig. 1) to minimize planar artifacts and ground clutter~~ *and consider areas within a 140 km radius around the five radar stations (Fig. 1). The 140 km radius limitation helps minimize planar artifacts and ground clutter."*

19. Lines 114, 125, 179, and 202: Delete "see". It is unnecessary as the citation to the figure number is sufficient. Fix throughout. Adjusted throughout.

20. Line 136 and throughout: Are colons used in UTC time? Generally not that I've seen. Delete the colons throughout. Adjusted throughout.

21. Line 150: I don't know what "sereal" means. If you mean "serial", I still don't know what that means in the context of this sentence. Please reword. The authors meant to write "several". We removed the word, as it is unnecessary.

22. Line 154: Delete "It is important to note that", which is unnecessary. The sentence is stronger without this phrase. Deleted.

23. Line 161: There should be a brief introductory paragraph after section 3, but before section 3.1, describing what will be discussed in section 3. Give the reader some context for what follows. The same thing is true for section 4. Agreed, we added short introductory paragraphs to the revised manuscript:

Section 3: "*In this section, we first provide an explanation of how hail days are extracted using the probability of hail (POH) radar proxies and then analyze the distribution of the POH time series*

Section 4: "*This section offers an overview of the development of the four statistical models and an evaluation of their performance. We discuss the development and performance of the individual logistic regression models (Sect. 4.1), the generalized additive models (GAMs) (Sect. 4.2), and the ensemble prediction (Sect. 4.3).*"

We also added a description to Section 5: *"In this section, we present the reconstructed time series from the ensemble prediction and discuss its trends (Sect. 5.1), the drivers of these trends (Sect. 5.2), and changes in the seasonal cycle over time (Sect. 5.3). Finally, we compare our time series with qualitative damage data (Sect. 5.4).*

24. Line 173: I would argue that "mesoscale" should be replaced with "microscale". We removed the sentence in line 173 to shorten the text and improve readability, but we agree that microscale is more fitting.

25. Line 176: Is that 27.0 hail days? It should be consistent with the number of decimal places in the other number. Adjusted.

26. Line 182: No need for "as shown in". Just put the citation in parentheses at the end of the sentence. Adjusted.

27. Figure 1: There are stray characters "1e6" in the lower-right corner and top left. I don't understand why it is there. Make its purpose more clear or delete. The stray characters "1e6" in the lower-right corner and top left are there to indicate multiplication by 1,000,000. This notation was used to avoid adding six zeros to each coordinate number, but we agree that this is uncommon. We changed the figure axis to km instead and removed the "1e6".

28. Line 216: Extra space between "hail days" and the period. Delete. Deleted.

29. Line 245: Avoid sentences that start with "this <verb>". The reader often does not know what "this" refers to. Please rewrite to be more specific. The authors agree and change line 245 to: "*The performance metrics suggest that the northern model can distinguish better between hail and no-hail days and misses fewer hail days than the southern model.*"

30. Line 248: Italicize p, as well as other variables throughout the manuscript. Adjusted throughout the manuscript.

31. Line 249: What is "insignificant". It is unclear. Avoid misplaced modifying phrases.
https://opentextbc.ca/advancedenglish/chapter/misplaced-and-dangling-modifiers/
We agree and thank the reviewer for reminding us of the correct use of modifiers. We rewrite line 249: "*Although the month factor was not significant, the model's performance decreased when removing the factor.*"

32. Line 261: Similar problem with "this indicates".

Line 260-261 changed to: "*The model, represented by the red dashed line, matches the marginal relationships of the data represented by the solid blue lines, and hence all predictors are well fitted and do not need further modification.*"

33. Line 294: What is "3"? We changed "(see 3)" to "(Fig. 3)".

34. Lines 294, 316: What does "this" refer to? We appreciate the reviewer's attention to detail in identifying errors in the manuscript.
We rewrote lines 291-296 following comment C3.
Lines 316-317 are rewritten for clarity: "*When the variable combinations from the northern model are applied to the southern domain and vice versa, the coefficients change and the predictive skill declines. This difference in coefficients and predictive skill underlines the necessity of using unique sets of predictors for each domain instead of a single model across all of Switzerland.* ~~When applying the combination of variables of the model north on the region south and vice versa, the coefficients changed and the predictive skill declined. This again justifies the use of an individual combination of predictors for each region, rather than applying one model for the whole area of Switzerland.~~"

35. Line 296: Delete "already". *Line 296 is rewritten following comment C3.*

36. Line 297: Avoid one-sentence paragraphs.
https://warwick.ac.uk/fac/soc/al/globalpad
rip/openhouse/academicenglishskills/writing/paragraphing/
Lines 297-298 are rewritten:
"The vertically integrated specific humidity (q_vint) quantifies the total amount of water vapor available in the atmospheric column and thus indicates the moisture available for hailstorm development. Consequently, a higher q_vint increases hail-day probability (Fig. 3d). ~~The vertically integrated specific humdidity (Q_vint) reflects the total amount of water vapor available for hailstorms to develop and hence, the larger Q_vint, the higher the probability of a hailstorm.~~"

37. Line 311: Why is "recursive" capitalized? Adjusted.

38. Line 311: "e.g." is Latin for "for example". As such, there is no need for "etc.", too. "etc" is removed.

39. Line 313: Delete the two "see"s. Deleted.

40. Lines 314, 330, and 359: "However" is used as a conjunction. Revise.
Line 314 was removed to enhance readability:
"~~Further discussions on the variable selection and their importance will follow in Sect. 6, however, it is important to differentiate between the two domains.~~"

Line 330 is changed to:
"The selection of predictors followed the same procedure as in the logistic regression model.~~, however,~~ For every variable that presented ~~has~~ an effective degree of freedom (edf) > 1, a smoothing spline function was ~~used~~ *applied* to allow for non-linear effects."

Line 359 removed, since we want to be careful in comparing our spatially averaged values to other literature.: "~~Observed values in thunderstorm environments often may exceed 1000Jkg−1., however, *In Europe,* severe hailstorms also occur with less *pronounced* CAPE (Taszarek et al., 2020a).~~"

41. Table 1 caption: Avoid the construction with parentheses.
https://eos.org/opinions/parentheses-are-are-not-for-references-and-clarification-saving-space
Table 1 caption changed to: "Coefficients, standard errors, *z*-values, and *p*-values of all covariates of the logistic regression models *for* north and south. Positive ~~(negative)~~ signs indicate a positive ~~(negative)~~

relationship of the quantitative predictors with modelled hail occurrence *and vice versa.* ~~with hail occurrence relative to the reference category (April) for the categorical predictors.~~ [...]"
All table captions will be adjusted accordingly.

42. Table 1: Italicize variables, such as z and v. Fix throughout. Adjusted throughout.

43. Table 2 caption: Why are bias, precision, and accuracy capitalized? Changed.

44. Lines 328, 331, and 344: There is no need for the apostrophe in GAMs, as it is not possessive. Find all occurrences in the manuscript and fix. Adjusted throughout.

45. Line 338: Be more specific. Most unstable CAPE determined over what depth?
Line 338 changed to: "[...] CAPE is the most unstable convective available potential energy, *computed for parcels departing from model levels below the 350 hPa level.*~~,~~*[...]"*

46. Line 340: d2m is not a sensible abbreviation for dewpoint temperature. I suggest T_{d,2m}, which would also be consistent with w_500. Moreover, the authors use the typical Td in Table A2. Be consistent internally within your manuscript and with common notation in the meteorological literature.
We adjusted the abbreviation "d2m" to "Td_2m" in the text and all tables. All temperature and dewpoint temperature abbreviations are changed to "T_x" and "Td_x" respectively.

47. Line 347: Why is "highly" in parentheses? Either delete the word or delete the parentheses.
We deleted the word ("highly") in line 347, as it does not make a difference if one predictor is highly significant or just significant.

48. Lines 347 and 571: Do you mean "variance" instead of "deviance"? Fix throughout the manuscript, if so. Indeed, we are talking about variance and not deviance. We thank the reviewer for his thorough reading. We changed "deviance" to "variance" throughout.

49. Lines 349–350: Why is GAM being redefined again? Definition removed.

50. Figure 3 and 4 need individual panel letters, and these figure numbers and letters need to be referred to in the text for the ease of the understanding of the reader. The panel letters have been added to the revised manuscript, together with their references in the respective sentences. We thank the reviewer for this suggestion.

51. Line 352: Italicize the y. Adjusted.

52. Line 353: Change "trough" to "through". Adjusted.

53. Line 354: There are long sections of text without paragraph breaks. This paragraph is just one example that needs fixed. But, this problem exists elsewhere, too. We restructured and shortened many paragraphs in the revised manuscript.

54. Line 361: What curve is being referred to? There are no figure and panel-letter citations in the sentences as far as I can see. Each reference to an interpretation of a figure needs to be cited so that the reader knows what they are looking at (e.g., Fig. 2a). Thank you for this suggestion. We added panel letters to the figures and reference citations in the text.

55. Line 370: Change "layer" to "level". Adjusted.

56. Line 371: This text repeats earlier text. It also indicates a profound lack of understanding of what scales vertical motion acts, as discussed previously. We refer to comment C3.

57. At this point, I am frustrated by the lack of proofreading, so I will not repeat comments that I have made before beyond line 372. The previous comments should be applied throughout the manuscript during extensive proofreading. We did extensive proofreading before the resubmission of the manuscript. Additionally, we sent the manuscript to a professional English editing service to ensure that any remaining mistakes are caught. Thank you for all your comments so far.

58. Lines 388–389: What are the citations for this information? We changed the paragraph following a comment (C8) from reviewer 2.

59. Line 392: Delete "We also need to mention that", and revise the sentence. The sentence in line 392 was removed following a comment (C8) from reviewer 2. We acknowledge the need to avoid phrases like "we need to mention" in the manuscript, and we will ensure that future revisions adhere to this guidance. Thank you for bringing this to our attention.

60. Section 4.2: I am finding the manuscript tedious, particularly in this section. Is there anything the authors can do to improve the readability of this manuscript? Options could include shortening the text, putting more content into appendices or supplemental files, and improving interpretation of the results. We concur with the reviewer regarding the manuscript's length and have restructured longer paragraphs to enhance readability. We removed unnecessary sentences to shorten the text. Additionally, we have moved the three appendix tables and the correlation matrices to the supplementary material. To further streamline the text, we have relocated tables detailing the coefficients of the statistical models and Figures 9, 11, 12, and 13 (a, b) to the appendix. These adjustments have effectively condensed the main sections of the manuscript.

61. Lines 461–462: Are two decimal places needed? Adjusted to one decimal place.

62. Line 466: Why is "e" raised to a power rather than "10"? Adjusted.

63. Line 472: No period at the end of this sentence. Adjusted.

64. Line 473: Insert "the" after "Using". Added.

65. Figure 10 caption: "Modelled" needs to be capitalized. "linse" is misspelled. Adjusted.

66. Figures 10, 11, and 12: Why the unusual color bar? There is an abrupt color transition between red and blue in the middle of the bar. Why? The color scheme should have a smooth transition. Moreover, one could ask why a color scheme is needed in the first place? Is the value of the point not sufficient to indicate its value? Why does the dot need a color associated with it? We used the color scheme to show which values are below and which above the mean. We agree, however, that this might be leading to a bias in the readers' interpretations of the result. We removed the colored boxes from the plot in the revised manuscript.

67. Line 487: Delete "it is important to mention that". It is unnecessary. Line 487, "It is important to mention that" is deleted. Thank you for the suggestion.

68. Lines 525–549 is one long paragraph. It needs to be broken up. In fact, given the number of times that I have flagged paragraphs as being too short or too long, the authors should revisit the rules of writing paragraphs.
https://warwick.ac.uk/fac/soc/al/globalpad
rip/openhouse/academicenglishskills/writing/paragraphing/
I also suggest reading Gopen and Swan (1990) "The Science of Scientific Writing"

(https://www.americanscientist.org/blog/the-long-view/the-science-of-scientific-writing) to gain further insight into better structuring your writing, particularly the use of topic and stress positions in sentences and paragraphs. Please indicate in your response that you have read this article and have implemented its guidance in the revised manuscript.
Unfortunately, the link to the first website you gave is no longer working.
The seven structural principles mentioned in the article on the science of scientific writing was considered when restructuring paragraphs and sentences of the manuscript. We also asked the professional English editor to focus specifically on any mistakes regarding the placement of old and new information in the topic vs. the stress positions of sentences and paragraphs.

69. Line 531: I don't understand "shear replacement". Please explain. In many statistical models predicting hail occurrence, kinematic information is often represented by a wind shear predictor. In our models, wind shear was not chosen. Instead, the variable v_500 was included in the southern model, making it a possible "replacement" candidate for wind shear. We agree that this terminology is misleading and have revised line 531 for clarity: "Moreover, ~~the shear "replacement"~~ *v_500*, as well as the explicit *wind* shear variables in the GAMs, did not have high feature importance."

70. Line 538: Markowski and Richardson (2010) is a book. As they are not presenting original results, it would be inappropriate to write that they "showed" something. We agree. Line 538 is removed due to comments from reviewer 2 and 3: "~~In fact, Markowski and Richardson (2010) and Dennis and Kumjian (2017) showed *explained* that wind shear primarily drives storm type and hail diameter.~~"

71. Moreover, when citing a book, you should also cite the specific page number to which you are referring to. Otherwise, the reader will not be able to source your information from the entire book easily. We added page numbers 201ff to the reference.

72. Line 539: When you say "significant fraction", please say what the specific number is. Houze et al. (1993) primarily focused on the directional movement of hailstorms in Switzerland, identifying 13 left-movers and 5 right-movers with true-hook structures out of 42 analyzed hailstorms. Since their study did not explicitly categorize hailstorms into supercellular versus ordinary or intermediate types, we agree that it is more appropriate to refer to a more recent and relevant study for specific fractions. Therefore, we have revised line 539 as follows:
"In fact, Feldmann et al. (2023) found that only 10% of severe hailstorms in Switzerland are supercell type storms […]"

73. Lines 534–543: This text is poorly organized. It is not a systematic description of the literature that advances your argument. We agree. Section 6.1 on the interpretation of the lack of windshear in the model is rewritten in the revised manuscript, also following a comment from reviewer 2:
"*The selection of predictors for the logistic regression models and GAMs needs to be discussed, in particular the absence of wind shear from the logistic regression models of both domains. When training our models, wind shear rarely appeared as a skillful predictor, and even when it did, it was not significant in the logistic regression models. Automated feature selection yielded similar results. We see three possible explanations for this. First, shear could also be indirectly represented by the variable v_500 (v-wind component at 500 hPa; southern model) in the southern model.*
Second, there is a nonlinear relationship between WS_06 (wind shear from 0 to 6 km) and the response in the GAM, which the logistic model struggles to fit. Additionally, neither v_500 nor the explicit wind shear variables show high feature importance in all models. This low feature importance has also been seen by Trefalt (2017) for Switzerland and by Mohr et al. (2015b) for Germany. A potential reason for this could be the prevalence of high shear but low CAPE conditions in our domains, which do not lead to hail. Thus, the wind shear parameters may not be effective in distinguishing between hail days and non-hail days in a statistical model, since there is no statistically significant difference in the distributions of WS_06 and WS_36 on hail days vs non-hail days in both domains.

Our third point is that high wind shear might be a less important hail model parameter in regions with complex terrain (Punge and Kunz, 2016). Although large shear values are required to form supercells, which are likely to produce hail, hail also develops in lower-shear environments (Schemm et al., 2016; Trefalt, 2017; Kumjian and Lombardo, 2020; Blair et al., 2021). In fact, Feldmann et al. (2023) found that only 10 % of severe hailstorms in Switzerland are supercell type storms, and Schemm et al. (2016) find average lower-tropospheric shear values at hailstorm initiation locations in Switzerland of less than 10 ms-1. Hail events in low-shear environments can be explained by proximity to mountain ranges, where environmental wind shear is increased by the interaction of the wind field with orography, which is often the case in the Alps (Trefalt, 2017; Kunz et al., 2018).
In such complex terrain, shear might be driven by local conditions, such as Alpine pumping, which are not resolved by ERA5´s resolution. Alpine pumping arises from differential heating and cooling of air masses over mountains and plains, which drive daytime winds from plains to mountains and nighttime winds in the opposite direction (Lugauer and Winkler, 2005)."

74. Line 544: What is "Alpine pumping"? Provide a citation and explain the result to nonspecialists. Or, delete. We added a citation and changed lines 544ff to: *"In such complex terrain, shear might be driven by local conditions, such as Alpine pumping, which are not resolved by ERA5´s resolution. Alpine pumping arises from differential heating and cooling of air masses over mountains and plains, which drive daytime winds from plains to mountains and nighttime winds in the opposite direction (Lugauer and Winkler, 2005)."*

75. Line 545: What is "orographic enhancement" (note correct term), and how does it create higher shear? Provide a citation and explain this mechanism to nonspecialists. After discussion with the coauthors, the sentence in line 545 was removed.

76. Line 562: This is the first occurrence of "5D". It is not defined, and the authors use it twice on this page. Define and use throughout, or delete. We deleted the "5D" throughout, as the term is not important.

77. Line 564: Can you please provide an argument for this assumption that "models learn how to best deal with limitations of the datasets"? We refer to comment C4.

78. Lines 564–568: This text seems like just a list of random thoughts put here, without context, without evidence and without organization. We refer to the modifications implemented in response to comment C4.

79. Line 575: Why is the BI only defined here, even though it has been used throughout the manuscript? Define upon first usage. Definitions were added to the manuscript.

80. Line 590: Why is b before a? Put citations in alphabetical order. Adjusted.

81. Lines 610–611: Why not provide what the increase is? Can you normalize all the different studies to a percentage increase by year, as they all use different time periods? We rewrote lines 606-611 added quantitative information on the increase:
"*Our modeled trends from the ensemble predictions align with the findings of Madonna et al. (2018), Rädler et al. (2018), and Battaglioli et al. (2023a). Madonna et al. (2018) reported an approximately 40 % increase in estimated hail days when comparing the periods 1980–2001 and 2002–2014. Similarly, Rädler et al. (2018) found a 41 % relative increase in hail cases per year during 1979–2016 in western and central Europe. Battaglioli et al. (2023a) identified an 8 % per decade relative increase in hail hours in northern Italy and parts of southern Switzerland for the period 1950–2022. In our study, we observe significant positive trends in both the northern and southern domains, with a 45 % increase in modeled hail days in the northern domain and a 48 % increase in the southern domain comparing 1960–1989 to 1990–2019. This translates to a relative increase of 7.5 % and 7.9 % per decade, respectively."*

Unfortunately, we are unable to normalize the different studies to a percentage increase per year due to the lack of complete data from the mentioned studies as well as the differing modelled periods. We believe that performing such a normalization is more appropriate for a review paper.

82. Lines 618–619: Why is a more extensive period of observational data required? You are already using a decadal-length time series of data from different datasets. In response to comment C5, we have removed the sentence in line 618-619.

83. Line 661–662: The author should make this code freely available via a public repository. The authors decided to share the code on request, since the code only uses very common packages from R to build and diagnose statistical models.

84. Line 669: AIC, BIC, and VIF are not defined here. Definitions added.

85. Line numbers are missing after line 665. Line numbers are added.

86. After equations for AIC and VIF: The comma should be on the same line as the equation. Adjusted.

87. Table A1 caption: Are TP, FP, FN, and TN needed? If not delete. If so, why not express the equations under “Explanation” in terms of these variables? The equations for performance metrics are calculated from contingency tables and typically use letters in the equations referring to TP, FP, FN and TN. It is important that the orientation of the contingency table is the same every time, otherwise the metrics are wrongly calculated, which is why using TP, FP, FN and TN leads to less confusion. We will express the metrics in terms of TP, FP, FN and TN instead.

88. Table A2: Why is deg01 used as an abbreviation for the height of the zero-degree isotherm? Why not use a more understandable expression, consistent with others? For example, z_{0°C}? The abbreviation for the zero degree level is not deg01 but deg0l (“l” as in level). All abbreviations come from the naming convention of the ERA5 files. We changed “deg0l” to "z_0°C" to make reading easier.

89. I am also similarly concerned about a number of other abbreviations that are not standard: FF, lfc (this and many others should be capitalized), mn2t, msl_mean, OMEGA (why all capitals?), etc. These inconsistencies should be made consistent, and standard abbreviations/variables used. Fix throughout the manuscript. We adjusted all necessary abbreviations in all text, figures and tables.

90. If Taszarek is a coauthor, then he should not be listed in the acknowledgements as providing data, but in the Author Contributions in terms of data curation. Indeed, we thank the reviewer for finding this.

91. Also, Taszarek’s data should be made available through the Code and data availability section.
https://www.natural-hazards-and-earth-system-sciences.net/policies/data_policy.html
Agreed, we added a reference to thundeR data.

NHESS-2024-371

Reply document to referee comments 2 (RC2)

Reconstructing hail days in Switzerland with statistical models (1959–2022)

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 for referencing purposes throughout the document (comment 1 = C1, etc.). Technical corrections are changed directly in the manuscript and are referenced with the starting line number, where removed parts are ~~crossed out~~ and new additions are in *italic*. All line numbers refer to the originally submitted manuscript.

Specific comments:

C1: line 28: You could add Manzato et al. (2022) here in addition to the Augenstein presentation. Their trend (Fig. 3) is not significant (but rather negative than positive) and at least it is published.

We thank the reviewer for this suggestion. The citation is added to the revised manuscript.

C2: 100: The intro is nicely structured and compact.

The authors are grateful for the positive feedback on the introduction.

C3: 91-94, 130-145, 162-169: I think there are some minor inconsistencies in your approach or how you explain it. You mention POH specifically includes small hail but then you use thresholds for POH and affected area which were tested for car insurance. To my knowledge, damage to structures and cars is dominated by severe hail (>2cm). Hence, to the reader it remains unclear what hail sizes the trend you find is representative of. It might be worth looking into the trend for more or less severe events (e.g., higher values of POH or larger areas?). For instance, it's good that you mention that the sensitivity to the area threshold was tested, but you could elaborate on this more and test other POH thresholds if feasible. I think such tests and discussions would improve the robustness of your results but I will leave it to you.

You are raising an interesting point. POH, per definition in the literature, captures hail of any size. However, preliminary comparisons with crowd-sourced hail size information do show some correlation between the POH value and hail stone size up to a POH of approximately 80%. Further on, as you rightly point out, damages to cars occur typically only for hail stones that are larger than approximately 2cm. We include this information in the discussion.

Lines 89–94 are changed to:

"We build on Madonna et al.´s (2018) work, but in this study, we increase the resolution of the analysis to daily, we additionally include the South of Switzerland, and we extend the time series back to 1959. Unlike Battaglioli et al. (2023a), who used ESWD severe weather reports, ~~we use radar data to model daily hail occurrence of any size at the ground~~ *we use Swiss radar data as proxies to model hail day occurrence. Furthermore,* we employ *an* ensemble of two statistical models, a logistic multiple regression and a logistic generalized additive model (GAM), to leverage the best-fitting predictors for each domain individually. ~~We include smaller sizes as we work with the POH radar product, which is a proxy for hail of any size at the ground. Additionally, even small hail can be damaging to agricultural produce (Katz and Garcia, 1981)."~~ […]"

Lines 126-145 are rewritten following a comment of reviewer 3:

"Statistical models classifying hail events typically select the ERA5 grid point that is temporally and spatially closest to the hail incident. However, such a selection is not possible for reconstructing past hail events because no information is available on the hail event prior to the observational period. Therefore, to model the occurrence of a hail day, we calculate ERA5 profiles averaged across the entire northern or southern domains at 12 UTC. The values at 12 UTC exhibited the highest predictive skill, which may be attributed to the fact that most storms in Switzerland occur in the late afternoon (e.g., Nisi et al., 2016, 2018). Thus, the 12 UTC value is most likely to capture the atmospheric conditions before storm formation.

Our definition of hail days focuses on days with more than a single hail cell. The thresholds are set to capture events that led to damage 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).”

Lines 162-169 are changed to:

“To identify hail days in northern and southern Switzerland, we use daily POH data from 2002 to 2022 during the hail-prone months of April to September. We use the same domains and area thresholds as Barras et al. (2021). The daily area of POH ≥ 80 % is extracted separately for the domains north and south of the Alps (Fig. 1). *To qualify as a hail day, the daily maximum POH must reach or exceed 80 % over an area of at least 580 km2 in the northern domain and 499 km2 in the southern domain. Barras et al. (2021) determined that these thresholds correlate best with days when car damage was reported across Switzerland from 2002 to 2012.* ***This definition implies hail large enough to cause damage to cars, approximately 2cm in size.*** *The sensitivity of our model’s to this threshold was tested by varying the area threshold. We found no significant impact on misses or false alarms, consistent with earlier studies indicating low sensitivity to area thresholds (Madonna et al., 2018).* ~~For a day to be categorized as a hailday, the daily maximum POH needed to equal or exceed 80% for at least 580 km2 in the northern domain and 499 km2 in the southern domain. Barras et al. (2021) found that these thresholds best correlate with days with car damage reported across Switzerland from 2002 to 2012. We tested the sensitivity of our model results to this threshold by varying the area threshold but did not find a significant impact on misses or false alarms, confirming earlier findings of low area threshold sensitivity (Madonna et al., 2018)~~”

The findings regarding *the sign* of the long-term trends, drivers of trends, and seasonality remain consistent regardless of the chosen POH area threshold, as anticipated, given our focus on haildays rather than specific hail sizes. Should we attempt to derive specific hail sizes instead of hail days from varying POH area thresholds (which one shouldn´t do due to the nature of the radar product), the results in trends could show some variability, particularly for very small hail sizes which may be influenced by the increasing melting levels observed in recent decades.

Due to the length of the paper, we prefer not to delve further into the sensitivity analysis in the text. However, should the reviewer deem it necessary, we can include additional comparisons, such as using the maximum expected severe hail size (MESHS) radar product to reconstruct trends in hail sizes.

C4: 235 and 299-307: I was wondering the whole time why no kinematic information is included. I would recommend adding a brief sentence here saying something like: „The lack of a kinematic predictor in the northern model will be discussed further later on.“

We added a sentence to line 307:

“[…] *The lack of a kinematic predictor in the northern model is discussed further in Sect. 6.1.* […]”

C5: 290-296: Convective updrafts are not resolved in ERA5, so large OMEGA cannot be caused by this. In other words, I don’t think this can be linked to the results of Lin and Kumjian (2022).

We thank the reviewer for commenting on this mistaken interpretation. 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 and primarily reflects large-scale synoptic ascent or descent. In our model, the highest probabilities of hail occur when omega_vint values are negative (Fig. 3c), 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 (omega_500) is a measure for the vertical motion of air, here for the level at 500 hPa. Negative values indicate upward motion ~~and hence measure updraft strength~~. The highest positive effect is achieved with the strongest negative vertical velocities (Fig. 6e)"

C6: Also, while I agree that the role of synoptic lift is likely included in this predictor, the inclusion of omega is surprising because I would have expected orographic lift to dominate over Switzerland? Perhaps because the larger scale lift favors move widespread convection and hence more POH area it is still a useful predictor?

We agree that orographic lift is important for the initiation of deep moist convection in the Swiss Alps. While ERA5 data can capture some aspects of orographic lifting, its ability to accurately represent the small-scale mountain-valley circulations and convective initiation with their connection to the diurnal cycle and differential heating may be limited by its resolution. Since omega is the vertically integrated vertical velocity (from model level data), it combines information about the whole atmospheric column. Synoptic lift is included in this predictor, however, if extensive enough, orographic lift might also contribute to the integral. Schemm et al. (2016) also showed, that around 30-40% of Swiss hail storms form in a pre-frontal environment. However, in the northern logistic regression model the existence of fronts might already be captured by the Boyden Index (BI). The predictand omega_vint had a high skill only in combination with moisture predictands, e.g. q_vint, the vertically integrated specific humidity. So, the model might benefit from the combination of information on large-scale lift, fronts and moisture as well as instability.

C7: 362-364: Lin and Kumjian (2022) saw increasing hail potential until around 2500 J/kg. CAPE doesn't even reach such values in your Fig. 6 and the curve flattens at 500 J/kg already, so I don't think these results should be linked, at least not with further explanation.

We fully agree and removed "~~This relation has also been found by Lin and Kumjian (2022)"~~ in line 363.

C8: 376-395 Some of your interpretations here were a bit confusing to me. I think this is a very active research topic but strong storm-relative winds have been shown to promote wider updrafts (Dennis and Kumjian 2017, Peters et al. 2020) and are hence important for hail. Also, weaker low-level winds have been suggested to be better, especially in north-south direction (Dennis and Kumjian 2017, Nixon et al. 2023). So your results or their interpretation are counter-intuitive, which should be clarified (you write that strong storm-realive winds are bad for hail but low-level shear good, at least that's how I understand your text, maybe you got them mixed up?). One explanation could be that the typical environments in Switzerland are different compared to these studies in the US. So the sensitivities to kinematic variables might be different. Which is worth to be discussed.

We thank the reviewer for this comment. Since we averaged parameters over a large area and did not use a single-grid approach, direct comparisons to individual storms in idealized simulations are not feasible. We chose not to compare our study to modeling studies in the U.S. due to the large methodological differences.

One explanation for the negative relationship of WS_36 is that days with high WS_36 often coincide with a front being near or over Switzerland. If the front is close, it might be associated with Föhn winds. Given the size of our domain, this could lead to a mix of prefrontal and Föhn environments.

If the reviewer finds it necessary, we can include additional discussion on this topic in Sect. 6.1.

I think this is a very active research topic but strong storm-relative winds have been shown to promote wider updrafts (Dennis and Kumjian 2017, Peters et al. 2020) and are hence important for hail

While an increase in 3-6 km shear can affect storm-relative (SR) winds, the impact depends on both the storm-motion vector and the orientation of the shear. As the 3-6 km shear increases, the storm-motion vector may also increase, potentially resulting in only a minimal net increase in SR winds despite significant increases in shear. Therefore, we refrain from speculating about SR winds unless we directly analyze a 3-6 km SR winds parameter rather than shear. Additionally, since we also considered non-supercell hail, findings from U.S. studies may not be directly applicable for comparison.

Also, weaker low-level winds have been suggested to be better, especially in north-south direction (Dennis and Kumjian 2017, Nixon et al. 2023).

We agree, but since our study does not include a predictor on low-level storm-relative winds, we again prefer not to make comparisons with those studies.

We adjusted lines 376-395:

"Notably, the deep layer shear WS_06 exhibits a nonlinear relationship to the response variable. WS_06 has its most negative effect at values around 0–10 ms-1, transitioning to a positive effect above 15 ms-1 (Fig. 5d). The curve flattens at very high wind shear values, suggesting that higher shear does not further increase the probability of hail. Additionally, the confidence intervals of smoothing functions widen significantly towards the tails of each covariate distribution.

GAMs are not limited by multicollinearity between model terms, which is why both WS_36 and WS_06 were selected in the northern model. The model preferred including both WS_36 and WS_06 over either one of them, as the individual predictors otherwise became insignificant and less important. Surprisingly, WS_36 has a negative linear relationship with hail in northern Switzerland. To gain a deeper understanding of how the WS_36 and WS_06 model terms interact, we further examined contour plots depicting conditional probabilities based on pairs of model predictors (not shown). The highest probabilities of hail are achieved with high WS_06 but low WS_36 in the northern model. ~~Typically the difference in 3km to 6km is much smaller than the deep layer shear WS_06. Hence, the model might indirectly learn, that the low-level inflow (0-3km) are more important to hail growth (and storm dynamic, as the inflow dominates storm dynamic) than the 3-6km layer shear. High 3-6km wind shear rather influences the stretching of the hodograph, making for stronger storm relative winds, and impacting downdraft formation and storm motion. With strong storm relative winds, the hail embryos might also be ejected out of the hail growth zone too quickly to experience significant growth (Dennis and Kumjian, 2017). This could explain why the model learns that high shear at this level leads to less probability of hail when differentiating between hail and no hail (and not hail size). We also need to mention, that we are just taking into account bulk shear which doesn´t include the curvature of the hodograph and the rotation of the storm. Dennis and Kumjian (2017) show that the biggest hail growth~~

~~volume is achieved with a straight, long hodograph because of a strong and in the 395 shear direction elongated updraft, rather than with a curved hodograph.~~ *Trefalt (2017) also found higher WS_06 and lower WS_36 on hail days vs. on non-hail days in northern Switzerland. This untypical relation could stem from the unique environmental conditions in Switzerland compared to the idealized modelling studies conducted for individual hailstorms in the United States (Dennis and Kumjian, 2017; Nixon et al., 2023). It is plausible that the sensitivities to kinematic variables differ between regions due to varying atmospheric dynamics and topographical features. This is further discussed in Sect. 6.1. Conditional probabilities of hail based on the various predictors (not shown) indicate that WS_06 and WS_36 have very low importance in the GAM models compared to SLI and TT."*

C9: 397: What do you mean by „circulate"? Most large hail seems to follow a single up-down trajectory while curving around the updraft (e.g., Kumjian and Lombardo (2017), Pounds et al (2023). No re-circulation with repeated ingestion into the updraft seems to happen. How this is in non-supercell storms is still unclear, but I don't see a reason to assume differently.

Also, I'd suggest rephrasing to „deeper hail growth zone" or „longer residence time in the hail growth zone".

We thank the reviewer for their comment and agree with the misleading wording. We rewrite line 397 for clarity:

"A lower freezing level suggests a greater potential for hail formation due to a longer residence time of hail embryos in the hail growth zone and less melting of hailstones before they reach the surface. ~~A lower freezing level may translate to a higher potential for hail formation because of the longer hail growth zone where embryos are circulating through areas of high supercooled liquid water and are able to grow bigger."~~

C10: 398-403: Punge et al. (2023) also found that excluding freezing lvls<2400m helped reduce false hailstorm detections in higher elevation in South Africa. Might be relevant here.

We thank the reviewer for this citation. Lines 398-403 are changed to:

"The probability of hail is highest at freezing levels between 2500 m.a.g.l. and 3500 m.a.g.l. (Fig. 5f). *Punge et al. (2023) also found that at higher elevations (≈ 2000m) in South Africa only a very small fraction of satellite based hail detections and hail damage claims occurred at freezing levels below 2400 m.a.g.l..* The model fits a negative linear relationship for freezing levels below 2500 m.a.g.l., indicating that lower values of z_0°C correspond to lower hail probabilities. This relation has also been seen by Kunz (2007) *and Trefalt (2017)* before. The negative relation suggests that our model does not learn about the melting or growth of hail embryos from the freezing level but ~~thet it is associated with connected to the height of the cloud base and therefore indirectly with the width of the updraft (Mulholland et al., 2021).~~ *instead uses it as a proxy for surface temperature, as both are positively correlated (Table S3 in the supplementary material). Thus, the negative effect of low freezing levels on hail probability could be related to lower surface temperatures."*

C11: 470: I liked that you followed closely and compared your results to Raupach et al. (2023).

The authors appreciate this comment!

C12: Section 5.3: April and September don't have a good sample size and I don't see any increase for either month, but in the text you say "This leads to more events specifically at the beginning of the hail season (April-June)". Perhaps May-June would be more accurate?

Indeed, this sentence is misleading. We thank the author for his thorough review. Further examinations showed that the fraction of haildays occurring until a specific month did not show a stronger increase in April and September compared to May-June across different decades.

Lines 498-508 are changed to:

"*This section addresses the seasonal analysis of hail occurrence over time. The last two decades exhibit a marked increase in hail days per month, which is strongest in May and June (Figure 10 blue and purple curves). We excluded years 1959 and 2020– 2022 to ensure consistency in the number of years per decade. Although the monthly curves display considerable variability, 445 their difference is not significant, and no systematic shift is evident, as illustrated by the cumulative distribution function (CDF) 23 plots in Fig. C2). However, our analysis is confined to the months of April to September and cannot support any statements about potential changes in hail days preceding or following the period modelled here.* ~~This section addresses the seasonal analysis of hail. Boxplots in Fig. 13 (a,b) show the modelled haildays per month for the whole time series period of 1959- 2022 for both domains. One can see a clear seasonal cycle of hailday occurrence. There is strong year-to-year variability in the monthly number of haildays (see e.g. circles for month July). In Fig. 13 (c,d) the mean number of haildays per month is plotted for each decade in differently colored curves. Decade 1960´s includes the years 1960-1969. The years 1959 and 2020-2022 are excluded, as we wanted to have the same number of years in each decade. In the last two decades (blue and purple curve), there is a strong increase in haildays. This leads to more events specifically at the beginning of the hail season (April-June) and to a shift of the peak of the convective season towards earlier months for both regions. These results are also evident when looking at the seasonal cycle by week or day of the year. Due to the large variability, the differences in the monthly curves are not significant and no systematic shift can be seen (see also cdf plots in Fig. A3). However, as our analysis was limited to the months of April to September, we cannot make assumptions regarding potential changes in hail events in the months preceding or following the modelled period.~~

C13: 525-549: I agree with these explanations, but are your hail day thresholds appropriate then, since they are trained with vehicle damages and hence larger hail ?

We thank the reviewer for their comment. Following comment C3, we will remove our assertion about differentiating between any hail and large hail. However, the low feature importance and low significance of shear in nearly all tested models shows that in our study area, high shear might not be as important for hailstorm formation.

Lines 525-546 are rewritten:

"The selection of predictors for the logistic regression models and GAMs needs to be discussed, in particular the absence of wind shear from the logistic regression models of both domains. When training our models, wind shear rarely appeared as a skillful predictor, and even when it did, it was not significant in the logistic regression models. Automated feature selection yielded similar results. We see three possible explanations for this. First, shear could also be indirectly represented by the variable v_500 (v-wind component at 500 hPa; southern model) in the southern model.

Second, there is a nonlinear relationship between WS_06 (wind shear from 0 to 6 km) and the response in the GAM, which the logistic model struggles to fit. Additionally, neither v_500 nor the explicit wind shear variables show high feature importance in all models. This low feature importance has also been seen by Trefalt (2017) for Switzerland and by Mohr et al. (2015b) for Germany. A potential reason for this could be the prevalence of high shear but low CAPE conditions in our domains, which do not lead to hail. Thus, the wind shear parameters may not be effective in distinguishing between hail days and *non-hail days in a statistical model, since there is no statistically significant difference in the distributions of WS_06 and WS_36 on hail days vs non-hail days in both domains.*

Our third point is that high wind shear might be a less important hail model parameter in regions with complex terrain (Punge and Kunz, 2016). Although large shear values are required to form supercells, which are likely to produce hail, hail also develops in lower-shear environments (Schemm et al., 2016; Trefalt, 2017; Kumjian and Lombardo, 2020; Blair et al., 2021). In fact, Feldmann et al. (2023) found that only 10 % of severe hailstorms in Switzerland are supercell type storms, and Schemm et al. (2016) find average lower-tropospheric shear values at hailstorm initiation locations in Switzerland of less than 10 ms-1. Hail events in low-shear environments can be explained by proximity to mountain ranges, where environmental wind shear is increased by the interaction of the wind field with orography, which is often the case in the Alps (Trefalt, 2017; Kunz 485 et al., 2018).

In such complex terrain, shear might be driven by local conditions, such as Alpine pumping, which are not resolved by ERA5´s resolution. Alpine pumping arises from differential heating and cooling of air masses over mountains and plains, which drive daytime winds from plains to mountains and nighttime winds in the opposite direction (Lugauer and Winkler, 2005).

We also tested combinations of shear and CAPE, such as WMAXSHEAR, an important parameter for differentiating between severe and nonsevere weather (Brooks et al., 2003; Craven and Brooks, 2004; Kaltenböck et al., 2009; Púcik et al., 2015; Tuovinen et al., 2015), but for both domains, WMAXSHEAR was not selected in combination with other variables."

C14: 600: Raupach et al. had to cover a much larger region and different climate zones. Perhaps it is fair to mention that the skill could be linked to such differences?

We appreciate the importance of recognizing such distinctions and aim to maintain humility in our interpretation of the models. Lines 599-602 are changed to:

"Our models outperformed those mentioned in Raupach et al. (2023a) due to lower FAR and higher HSS values. (Raupach et al., 2023a) mention HSS ranges from 0.1 to 0.4 compared to our models' HSS of 0.73 (north) and 0.55 (south). FAR ranges from 0.57 to 0.8 compared to 0.23 for north and 0.35 for south in our models. *However, Raupach 535 et al. (2023a) conducted their study over a much larger area encompassing diverse climate zones. Over our domains, hail is a comparatively frequent event with an a priori probability of approximately 15 % in the sample, which mitigates some of the statistical intricacies."*

C15: 647 and abstract: I don't see a longer peak in Fig. 13. (see also comment on September above)

We agree. Lines 646-648 are changed to:

"The increase in hail days in the last two decades is strongest in May and June. However, the seasonal cycle shows no clear shift towards an earlier start or earlier end, and differences in monthly distributions across decades are not significant. ~~Furthermore, we can see an increase in haildays in April-June in the last two decades and a slightly earlier and longer peak of the hail season compared to earlier decades. Still, there is no clear shift of the whole season towards an earlier start and earlier end, and differences in monthly distributions of all decades are not significant.~~"

Line 13-15 are changed to:

"The last two decades show a considerable increase in hail days, which is strongest in May and June. The seasonal cycle has not shifted systematically across decades. ~~In the last two decades, we can see~~

~~an increase in haildays at the beginning of the hail season and an earlier and longer peak, however, there is no systematic shift 15 in the seasonal cycle.~~"

C16: Overall, I liked your thorough discussion and conclusions.

We are grateful for the reviewer's positive feedback.

C17: Acknowledgements: You mention that parameters from thundeR were tested. Maybe I missed this in the text, but some discussion of it might be insightful, no?

We conducted tests on thundeR parameters for two reasons:

1.) Some of our parameters were derived from ERA-5 data at a lower resolution (0.5x0.5°) rather than the highest available resolution (0.25x0.25°). We aimed to determine whether this resolution difference influenced model performance and whether calculating profiles based on e.g. the pixel with the highest CAPE in space and time would enhance performance compared to our method of spatially averaging profiles. However, neither scenario resulted in improved performance.

2.) We wanted to investigate if incorporating additional convective parameters, initially not included in the parameter selection (e.g., MUCAPE above -10°C), would significantly enhance model performance. However, our findings did not support this. It's worth noting, however, that we did not dedicate the same level of effort to train and refine these models as we did for our original selections. It's plausible that with extensive further work, models with similar or better performance could be identified. We believe that ongoing research and the availability of more and higher resolution data in the future could further improve such models.

Due to the manuscript length and potential reader confusion, we would like to not include this discussion in the manuscript text.

C18: Also, since M. Taszarek is in the author list, I'm not sure if it's necessary to acknowledge his contribution. Your choice though.

Thank you for bringing this to our attention.

Technical Corrections:

L 27: I'm not sure if this is the standard formatting in egusphere but with some references you use brackets around the years (when the reference starts wiht something line „e.g.") and in some not. Also, it is common to put a comma after „e.g." I think.

Thank you for this comment. All references and the "e.g.," are checked again in the revised manuscript.

L 43: Just a minor thing, but I think „ERA5" is more commonly used? Adjusted in all sections.

L 48: I suggest removing „a" before „sufficient" Adjusted.

L 68: I suggest writing either "both thermodynamic" in the brackets or "...can be grouped into thermodynamics (instability and moisture) and kinematic conditions." to make clearer that „thermodynamic“ refers to both instability and moisture

Thank you for this suggestion, we changed line 68 to:

“The parameters and indices can be grouped into three categories: instability and moisture, which are both thermodynamic, and kinematic conditions.”

L 195: Suggest removing extra brackets in the exponent. Adjusted in all sections.

Fig. 3: Units are missing here in some of the following figures. Added in the revised manuscript. Thank you for bringing this to our attention.

L 455: Add a space before „respectively“. Adjusted.

L 564: „choose“ Adjusted.

The references to Augenstein et al. (2023) and Mohr et al. (2015 a,b) need some fixing.

Manzato, A., S. Serafin, M. M. Miglietta, D. Kirshbaum, and W. Schulz, 2022: A Pan-Alpine Climatology of Lightning and Convective Initiation. *Mon. Wea. Rev.*, **150**, 2213–2230, https://doi.org/10.1175/MWR-D-21-0149.1.

Nixon, C. J., J. T. Allen, and M. Taszarek, 2023: Hodographs and Skew Ts of Hail-Producing Storms. *Wea. Forecasting*, **38**, 2217–2236, https://doi.org/10.1175/WAF-D-23-0031.1.

Peters, J. M., C. J. Nowotarski, J. P. Mulholland, and R. L. Thompson, 2020: The Influences of Effective Inflow Layer Streamwise Vorticity and Storm-Relative Flow on Supercell Updraft Properties. *J. Atmos. Sci.*, **77**, 3033–3057, https://doi.org/10.1175/JAS-D-19-0355.1.

Punge, H. J., Bedka, K. M., Kunz, M., Bang, S. D., and Itterly, K. F.: Characteristics of hail hazard in South Africa based on satellite detection of convective storms, Nat. Hazards Earth Syst. Sci., 23, 1549–1576, https://doi.org/10.5194/nhess-23-1549-2023, 2023.

We thank the reviewer for his thorough proofreading. All references are fixed.

NHESS-2024-371

Reply document to referee comments 3 (RC3)

Reconstructing hail days in Switzerland with statistical models (1959–2022)

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 would 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 are excluded from the analysis because hail rarely occurs there (Van Delden, 2001; Giaiotti et al., 2003; Nisi et al., 2016) and radar quality may be lower (Feldmann et al., 2021). The central Alps is delineated from the northern and southern pre-Alps by the boundaries of the official prognosis regions from the Federal Office of Meteorology and Climatology MeteoSwiss. This selection of the study domains allows the climatological regimes north and south of the Alps to be separated and corresponds to those 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 noticed 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 a strong year-to-year variability and follows a pronounced seasonal cycle (Schröer et al., 2023). Recent studies (Nisi et al., 2018, 2020; Barras et al., 2021; Schröer et al., 2023) have highlighted substantial differences in both inter- and intra-annual hail variability between the northern and southern sides of the Alps. In the northern domain the peak of the convective season typically occurs in June, whereas in the south, it occurs in July (Fig. 2). Moreover, the occurrence of hail-prone and hail-sparse years differs between the two regions"

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 is 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.~~
Convection in the region south of the Alps is influenced by the transport of moist and warm air masses originating from the Adriatic and Mediterranean Seas (e.g., Nisi et al., 2016) during southwesterly or southern flow conditions. These air masses create ideal conditions for convective storm development, when coupled with local wind systems such as mountain– plain circulations and valley breezes. Previous studies have highlighted the relevance of anabatic–katabatic wind systems in the southern Pre-Alpine region and specifically in the Po Valley to hail formation (Morgan, 1973; Gladich et al., 2011). The southern domain is shielded from northern air masses by the Alpine chain, whereas the northern domain is regularly exposed to frontal systems originating from the west or north (Schemm et al., 2016).”

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 days, 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:
"Statistical models classifying hail events typically select the ERA5 grid point that is temporally and spatially closest to the hail incident. However, such a selection is not possible for reconstructing past hail events because no information is available on the hail event prior to the observational period. Therefore, to model the occurrence of a hail day, we calculate ERA5 profiles averaged across the entire northern or southern domains at 12 UTC. The values at 12 UTC exhibited the highest predictive skill, which may be attributed to the fact that most storms in Switzerland occur in the late afternoon (e.g., Nisi et al., 2016, 2018). Thus, the 12 UTC value is most likely to capture the atmospheric conditions before storm formation.
Our definition of hail days focuses on days with more than a single hail cell. The thresholds are set to capture events that led to damage 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 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:

"We build on Madonna et al.´s (2018) work, but in this study, we increase the resolution of the analysis to daily, we additionally include the South of Switzerland, and we extend the time series back to 1959. Unlike Battaglioli et al. (2023a), who used ESWD severe weather reports, we use Swiss radar data as proxies to model hail day occurrence. Furthermore, we employ an ensemble of two statistical models, a logistic multiple regression and a logistic generalized additive model (GAM), to leverage the best-fitting predictors for each domain individually. Our statistical models are tailored to Switzerland. Our goal is not to build a model for forecasting, but we want to produce the best possible reconstruction of past hail days in Switzerland from environmental predictor variables. The statistically modelled time series will then be used to study long-term trends and changes in frequency, seasonality, and the variability of model-derived Swiss hailstorms in past decades."

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 use the performance metrics of the ensemble predictions. Our models outperformed those mentioned in Raupach et al. (2023a) due to lower FAR and higher HSS values. Raupach et al., (2023a) mention HSS ranges from 0.1 to 0.4 compared to our models' HSS of 0.73 (north) and 0.55 (south). FAR ranges from 0.57 to 0.8 compared to 0.23 for north and 0.35 for south in our models. However, Raupach et al. (2023a) conducted their study over a much larger area encompassing diverse climate zones. Over our domains, hail is a comparatively frequent event with an a priori probability of approximately 15 % in the sample, which mitigates some of the statistical intricacies. Other studies have demonstrated comparable performance to ours, such as Battaglioli et al. (2023a) using ESWD hail reports and ERA5 data, López et al. (2007) using radar and radiosonde data, and Gascón et al. (2015) using severe storm 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 and primarily reflects large-scale synoptic ascent or descent. In our model, the highest probabilities of hail occur when omega_vint values are negative (Fig. 3c), 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 (omega_500) is a measure for the vertical motion of air, here for the level at 500 hPa. Negative values indicate upward motion ~~and hence measure updraft strength~~. The highest positive effect is achieved with the strongest negative vertical velocities (Fig. 6e)"

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. Thank you!

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 the 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 probability of hail is highest at freezing levels between 2500 m.a.g.l. and 3500 m.a.g.l. (Fig. 5f). *Punge et al. (2023) also found that at higher elevations (≈ 2000m) in South Africa only a very small fraction of satellite based hail detections and hail damage claims occurred at freezing levels below 2400 m.a.g.l..* The model fits a negative linear relationship for freezing levels below 2500 m.a.g.l., indicating that lower values of z_0°C correspond to lower hail probabilities. This relation has also been seen by Kunz (2007) and Trefalt (2017) before. The negative correlation suggests that our model does not learn about the melting or growth of hail embryos from the freezing level but *instead uses it as a proxy for surface temperature, as both are positively correlated (Table S3 in the supplementary material). Thus, the negative effect of low freezing levels on hail probability could be related to lower surface temperatures.*

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 days. 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 combining the best logistic regression model and generalized additive model (GAM) outputs for each domain. The ensemble prediction is generated by averaging the predicted probabilities from both the logistic regression model (LM) and the generalized additive model (GAM). We again conduct sensitivity tests to determine the best 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 ensemble prediction outperforms individual models across all skill metrics.*”

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, the SWISS Index was rarely featured in the best combination models. Notably, its inclusion sometimes led to slightly higher false alarm ratios. It is 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"?