Dear reviewer,

Thank you for your time in providing detailed feedback and suggestions for our manuscript "Exploiting radar polarimetry for nowcasting thunderstorm hazards using deep learning". Below, we provide our responses to the comments of the reviewer

The Reviewer's original comments are noted below in *green italics*. Our responses are given below each comment in normal font.

Best wishes,

Authors

General comments

The authors present relevant and new scientific results, which fit well within the scope of the journal.

Overall, the level of is English is good but the clarity and flow of the text can be improved, for example the following sentence is unclear: "... due to high demand in computation time results, are not available in real time." or L122": "The main difference between the predicted thunderstorm hazards is that heavy precipitation is trained ... ". Some formulations are ambiguous or confusing and should be improved e.g. L174 "The average loss of lightning".
Thank you for pointing this out, we proofread the paper again and tried to improve the clarity of the writing; the detailed changes can be found in the attached tracked-changes file.

Based on both reviewer comments, we rewrote the second paragraph in the introduction, consequently this sentence was removed: *"... due to high demand in computation time results, are not available in real time.*"

And we rewrote:
"*The main difference between the predicted thunderstorm hazards is that heavy precipitation is trained ..."*
into
"The main difference between the predicted thunderstorm hazards is that the output of heavy precipitation is accumulated over 1 hour for predefined warning levels, whereas hail and lightning are produced at a 5-min resolution for 12 time steps (1 hour)." (lines 158-160)

I missed the information on how the train-test split was done (randomly, different time periods?). Did you make sure to include all kinds of events (and non-events) in both? Please provide some more details on the learning process (learning rate, stopping criterion etc.) which are crucial pieces of information for this kind of research.

We used the same method as used in the previous research from Leinonen, but this is indeed not mentioned in the manuscript. As it is a crucial part, we added in the methods a new section about the selection of events (*section 3.1 Event selection*). We also wrote in the section 3.2 *Neural Network* (lines 173-176) about the learning rate and stopping criterion.

I would also suggest that the authors better motivate the limited lead time of one hour. Is this enough to act upon? Or is it motivated by an inherent predictability limit to the phenomena you are trying to forecast?

In order to make it comparable with the results of Leinonen et al. (2023) we limited ourselves also to a lead time of 60min. For lightning, the result is useful for beyond 1h. For that reason, it is possible to extend the lead times for lightning in another study. For hail, the convective nature and the small scale structure of the phenomena heavily impede the nowcasting on a local scale to a maximum useful lead time of 10 to 20min. This is rather short. In case of a fast data transfer, a user might react in time to get a benefit (e.g. move property to safety); a fully automated system like hail damage prevention for window blinds would also have time to react. An example of such a system is described at https://www.hagelschutz-einfach-automatisch.ch/eigentuemer-verwaltungen.html.

Moreover is there a reason why rain gauges are not used as an input, or for validation of heavy rainfall?

The nowcasting system of MeteoSwiss is designed for a very high timeliness for the reasons discussed in our response to the previous comment; we start the nowcasting 30s after the last measurement (radar data arrives in time for this). In contrast, data from rain gauges arrive in sufficient completeness around 8 min after measurements, and for the same reason CombiPrecip (product which combines the real-time radar reflectivity and rain-gauge observations) is not used an input. However, CombiPrecip is used as a target.

It would be useful for the authors to also discuss the added value of their DL-based methodology as compared to more traditional methods in terms of computational performance. How long does it take to run this 1-hour nowcast?

We need 8 seconds for a nowcast of one hazard with 12 timesteps on a machine with 4 CPUs (Intel(R) Xeon(R) Gold 6142 CPU @ 2.60GHz) and need 16GB of RAM. We added this to the end of subsection *3.2 Neural Network* (lines 178-179).

Specific comments

L22: The sentence "the initial state of the atmosphere in NWP assimilation is based on previous model predictions rather than the latest available observations, which makes it less suited for

accurately predicting the time and location of convective storms": appears to stem from some misconception. Indeed, the previous forecast is used as a first guess or background field, but this is in fact combined with the latest observations to create an initial state in the data assimilation process. The main reasons why nowcasting is important is because of the faster computational time which allows a higher update frequency, and because a purely (or mostly) observation-driven system will (by construction) be closer to observations for the shortest lead times.
Thank you for pointing out this mistake, we rewrote this paragraph into: *"NWP analysis is a combination of previous model predictions and the latest available observations, and the assimilation creates a physically consistent state of the atmosphere, which typically deviates slightly from the latest observations. Meanwhile, nowcasting algorithms aim to provide their output within tens of seconds up to a minute Pierce et al., 2012).. They typically do not strive for a physically consistent representation of the atmosphere, but do make use of the latest observations, which results in higher performance on the very short and short time scales (i.e. 1~h) and smaller scales (Simonin et al., 2017) (but inferior performance on longer lead times).."*

L37: The rainfall fields on which these nowcasting systems are based typically already make use of dual polarization variables to estimate the rain rate, clutter etc. What you mean is that you explicitly add polarimetric variables in the nowcasting scheme.
In the Swiss weather radar network, dual-pol data is indeed used for clutter suppression. We included this in the *Single-pol Radar* (before named "*Weather radar")* part in subsection *2.2 Data sources and preprocessing*.
To make it more clear that we use polarimetry variables explicitly we changed the text to: "*However, these studies primarily focus on single-polarization radar observations (e.g. precipitation rates based on horizontal reflectivity or the reflectivity itself), and do not utilize polarimetry explicitly, despite polarimetry can provide further information about the micro-physical properties of hydrometeors. Hence, adding polarimetric radar variables explicitly helps considerably to reduce ambiguities concerning the hydrometeor classes and drop size distributions* ."

L58: "dataset from Leinonen et al. (2022b)"- can you provide a bit more detail here (which variable? Radar rainfall dataset?)
We specified which exact dataset for training from Leinonen et al. (2022b) was used in section *2 Data*. In subsection *2.2 Data sources and preprocessing* we added the variables that are part of this dataset *Single-pol Radar* (before named "*Weather radar")*.

L71: "Weather radar (R) observations"- please be more specific. I suppose you mean reflectivity, but there are many more weather radar observations (e.g. radial velocity…)
We specified the corresponding variables used in the data source *Single-pol Radar* (before named "*Weather radar")*.

L93: "After hyper-parameter turning, a value of -0.5 for was selected, as Wolfensberger et al. (2021) found that this resulted in the best parameter value" - it's unclear to me whether the

value selection was the result of hyperparameter tuning or just taking the best value from literature? Also, "this resulted in the best …"should probably read "this was the best…"

The value is taken from literature. In the research of Wolfensberger et al. (2021) hyperparameter tuning was performed, which resulted in the value -0.5. We rewrote this so it is more evident that we did not do the hyperparameter tuning ourselves.

L94: Next, the data was transformed by first normalizing it by bringing the mean close to 1 -> How was this "brought close to 1", why not simply set it to 1?
The normalization assumes a climatologic distribution. So on average it is 1, but not for every single field. We changed it into: *"First, the polarimetric data were transformed by normalizing the standard deviation and by shifting the mean to 1."*

L127-128: "The focal loss is an adaptation of the CE and focuses more on the difficult cases" - this is not informative, please be more specific or leave out the last part of this sentence.
We changed this into: "*To be consistent with Leinonen et al. (2023), the focal loss (Lin et al., 2017) is used for lightning. The focal loss is an adaptation of the CE and focuses more on the pixels whose classification is more uncertain ($p_t < 0.5$). In which p_t is the predicted probability of the target."*

L129: "We trained each possible combination of data sources..." - you cannot train data (or sources, or variables), you train a model.
This has been changed to "*We trained the model with each possible combination of data sources three times*"

L130: trained only 3 times: is it enough?
If you want to draw a statistically meaningful conclusion, you need to have more results. However, it is constrained due to the computational time to train all the models with all different combinations multiple times. By already training it a few times, we can get a first impression of the variance, and how robust the results are.

L134: Shapley value, which distributes the total score among its predictors... -> In game theory it is used the score this way, but there is no "score" in the picture here, please rephrase.
We rephrased it to:
"The importance from individual data sources can be assessed using the Shapley value (Shapley, 1951) as a quantitative indicator of the total importance of each data source. The total contribution among the predictors is distributed by assigning a value that represents their marginal contribution."

L138: You convert the probabilities to binary values, why? Since both the network and POH algorithms output a probability, why not use a metric that quantifies PDF overlap/mismatch?
We used multiple metrics to evaluate our model. The cross entropy (for POH and precipitation) and focal loss (for lightning) both also take the probability distributions into account (Figure 3).

Additionally, we did the analysis with binary values; as it was often requested by reviewers of the previous papers, we decided to add it here.

L139: What do you mean by "predictability"?
With this we mean how good the model is in predicting that the rainfall amount is exceeding the different warning thresholds (>10mm , >30mm and >50mm). To avoid confusion, we will change it to skill scores.

Throughout the manuscript, "radar data" is used to refer to reflectivity data. This is confusing, since polarimetric data also comes from radars. Please change this throughout the manuscript (e.g. L189 "radar is the most important source…"-> "radar reflectivity data is …") also in the captions (e.g. Fig. 3)
We changed the name of the "Weather radar" source to "*Single-pol radar".* Throughout the manuscript, where appropriate, we also changed the name from radar to single-polarization radar to better differentiate between the two different data sources.

L200: Please specify which thresholds you mean (probability? intensity?)
Specified to *probability thresholds*

L213: "resulting in penalization"-> what kind? Can you be more specific (double penalty?)
Corrected to double penalization

L215: <FSS for> RPQ and R
Corrected

Typographic, grammar and other small corrections

Abstract and L14: humans -> human lives
Corrected to "*human lives"*

L21: NWP <models> often have
Corrected to "*NWP models"*

L33: In the recent years -> in recent years
Changed to "*in recent years"*

L63: Operational<ly> available products
Corrected to "*operationally available products"*

L186: Lie more apart -> Lie further apart
Altered to "*lie further apart*"

L229: stresses out -> confirms, shows

Changed to “*shows”*

L232: can not -> cannot
Corrected to “cannot”