Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., author comment AC1
https://doi.org/10.5194/nhess-2021-171-AC1, 2021

Reply on RC1

Jussi Leinonen et al.

Author comment on "Nowcasting thunderstorm hazards using machine learning: the impact of data sources on performance" by Jussi Leinonen et al., Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-171-AC1, 2021

Dear Dr. Rigo,

We thank you for your constructive comments. Please find below our answers to your comments. The original comments are posted in *italic* font and the point-by-point responses are under each comment in normal font. The specific changes made to the manuscript in response to the comment are described in **bold** font.

Best Regards,

Jussi Leinonen (on behalf of all authors)

General comments

The document is well-addressed, easy to follow, and well-documented. My major concerns regard on the type of data and/or methodologies, according to the main objective: "we seek to understand the impact on thunderstorm nowcasting from the new generation of geostationary satellites, which, compared to the previous generation, provide higher-resolution imagery, additional image channels and lightning data ":

- In my opinion, there is a lack of coherence on the considered model: if the authors analyze an American region and all the data are from American sources, why in the case of the NWP they have considered the European one. Could you clarify this point?

3/4 of our team are from a Meteorological Service based in Europe, whose primary focus is on improving weather products in that region. We are carrying this study out in the American region mostly out of necessity rather than out of specific intent to analyze the US thunderstorm environment.

More specifically, to fulfill the objective of working with MTG-like data, we need to carry this study out in regions where GOES-R data is available, as MTG itself is not yet operational. The use of GOES-R limits us roughly to the Americas, from which we chose a part of the US as a study area (as the higher time resolution CONUS product is available there), which in turn required us to use the NEXRAD network as our radar data source. On the other hand, the ECMWF IFS model *is* available in the US, so unlike with the satellite

and radar, we are not constrained to choosing an American model for the source of our NWP data. In fact, we prefer using ECMWF since, as mentioned in Sect. 2.2.4, it should allow us to adapt the knowledge and tools developed in the course of this study to later research in Europe.

We have added some clarifying remarks in the first paragraph of Sect. 2.2.4 to add information mentioned in the response above.

- In a similar way, if the analysis is focused on operational methods, why you do not have taken into account the operational methodology (or a similar one, maybe provided by the NWS or the NOAA) used in Switzerland. The use of different types of techniques can lead to significant errors. Could you justify with at least one example, the similitude of the results of both techniques?

The current operational procedure in Switzerland is using an empirical thunderstorm intensity (TRT rank). Therefore, it is difficult to directly compare precipitation intensity, lightning activity and hail occurrence to the TRT rank. The assumption for the TRT rank is persistence. The paper compares the predictive skills of the ML models to the persistence, and, hence, is a kind of comparison to the Swiss operational algorithm.

In order to keep the conclusions of this paper general and applicable to several environments and operational agencies, and facilitate the replication of the results, we decided to use very generic methodology rather than any particular operational scheme. However, we feel this was not well communicated in the submitted version and **we have added a mention of this in Sect. 3.1.**

- According to figure 3 and other comments in the discussion and the conclusions, how optimistic are you in the improvement of the nowcasting using this technique?

We show in Figure 3 and other results in this study that we can improve results over simple baseline methods. The significance of the improvements in practical use depends on the end user and the application. The results could be applied to several different use cases (e.g. lightning prediction for aviation, flood warnings for hydrological services, or real-time hazard warnings for the general public). Different countries and meteorological services also use different procedures to issue warnings. Therefore, in this paper we would prefer to evaluate and quantify the results and leave it to the individual end users to decide whether the improvements are worth the implementation of the ML prediction scheme. However, **we have made an improvement to Figure 3 that shows the relative error of the rain rate that corresponds to the reflectivity error**, thus providing a more concrete way to judge the improvement.

And which could be the ways for improving the ML technique (e.g. other data sources, other thresholds, the ML itself) in the future?

In the final paragraph of the Conclusions, we already discuss potential future directions for the study, including additional data sources and alternative ML methods (specifically, neural networks). We are not quite sure what the reviewer means by "other thresholds". However, **we added in a mention in this paragraph that the methodology could also be extended for use with other types of hazard**.

Minor comments

- Figure 3: If the methodology (section 3) considers that Maximum reflectivity should exceed 37 dBZ, I cannot understand the Figure, because it shows that most of the time MaxZ do not reach this threshold and even in one case it never exceeds this value.

This is because the threshold for selecting the cases is based on a single-pixel value, i.e. MAXZ exceeding 37 dBZ in a single pixel. Meanwhile, in Fig. 3 we are showing the MAXZ predictand, defined as the mean of the column maximum reflectivity over the 25-km diameter neighborhood. As the mean is smaller than the maximum, in many cases the shown MAXZ does not exceed 37 dBZ. However, thank you for pointing this out. This seeming discrepancy was not well explained in the original text; **we have revised the caption of Fig. 3 and the text of Sect. 4.1 to make it clearer what is shown and why the MAXZ is sometimes below 37 dBZ**.

- In the same way that the previous point: which is the reason of selecting a so low reflectivity threshold (37 dBZ), considering that severe thunderstorms present values much higher than this threshold. Could you explain the motivation of your choose?

The 37 dBZ threshold was selected based on various earlier studies which identify thunderstorms based on reflectivity thresholds between 30 dBZ and 40 dBZ. **We have added more explanation and several references in section 3.1.1 to support this**. Severe thunderstorms can indeed have reflectivity values much higher than this, but one goal of nowcasting is also to identify those thunderstorms that may later become severe, so that advance warnings can be provided. Therefore, we need to choose a threshold at which the thunderstorms are not yet severe but rather have potential to become severe storms.

- Figure 4: I assume that as higher is the value of y-axis, worst is the performance. But, how do you really quantify the quality of the performance? E.g. POD values close to 0 (1) are very bad (good) skill values, or the opposite, FAR close to 1 (0) indicates bad (good) performance.

For variables like MAXZ shown in Fig. 4, which are predicted as a value of the variable rather than a probability of some event occurring (as with, for example, LIGHTNING-OCC), showing the MAE or RMSE error metric is how the performance of the prediction is usually quantified. Beyond such metrics, the performance depends on the specific application and the needs of the end user. As mentioned in our response to the reviewer's earlier question, in order to keep the conclusions of our study general, we have left such considerations outside the scope of the paper.

- Paragraph of L290: how good do you assume is your performance in operational terms with an increase of the MAE of 1.2 dB. Can you explain it?

Please see our response to the previous comment and to the earlier comment from the same reviewer starting with "According to Figure 3…"

- The occurrence of hail is poorly dependent of the occurrence of 45 dBZ, because of different reasons: values are concentrated at low levels, or the freezing height is much higher than the EchoTop45. In my opinion, choosing VIL parameter gives a better correlation (also poor, but less in any case) with hail occurrence. I would like that you provide a clarification of your selection

We do not intend to represent hail directly with the occurrence of 45 dBZ. As indicated in the second sentence of the third paragraph of Sect. 3.2, we refer to the heuristic from the Probability of Hail (POH) metric, which uses the height difference between the 45 dBZ echo top and the freezing level to indicate hail. This difference is undefined if the 45 dBZ reflectivity does not exist in the vertical column. Therefore, we split the prediction into two components: ECHO45-OCC, which predicts whether the 45 dBZ reflectivity occurs, and ECHO45-HT, which is only used if ECHO45-OCC predicts hail, and which predicts the height of the echo top. The freezing level is obtained from the NWP data so there is no need to predict it with the ML model.

In an operational setting, the hail prediction using our scheme would work roughly as follows:

- If the ECHO45-OCC model predicts that a 45 dBZ reflectivity will be present, use the ECHO45-HT model to predict the 45 dBZ echo top height, then subtract the freezing level from this in order to compute the POH.
- If the ECHO45-OCC model predicts that a 45 dBZ reflectivity will not be present, predict that no hail will occur.

We have summarized the above description at the end of Sect. 3.2 in order to clarify this point further.

- Paragraph of L320: The increase of the influence of the NWP in time over the forecast is a well-known fact. Could you be more concise in the weight of this data source in your results? How do you explain the "valleys" in the relationship with radar data? (Fig. 6)

Our best explanation for the valleys is that they are simply noise. Since the models for each time step are trained independently, they may converge to slightly different feature importances. **We now mention this in the paragraph**.

We think that it is useful to mention the increase of the influence of NWP with longer lead times, but **we now also state that it is a known result and support this with a reference**.