Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., author comment AC3
https://doi.org/10.5194/nhess-2021-171-AC3, 2021

Reply on RC3

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-AC3, 2021

Dear Reviewer 3,

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:

My chief comment concerns exercising caution when using blanket statements about drawbacks of "brute-force" strategies in machine learning (e.g., Page 20, starting Line 424), considering a line of thinking in deep learning that more data is better than less, given that a properly structured neural network with representative training, validation, and testing sets should effectively handle counterproductive data sources by properly reducing their weight in very high-dimensional, multivariate space. Granted that decision trees, rather than neural networks, are the focus of this study, but it may be beneficial to mention that the given recommendations may not hold in the case of neural networks when much more data are available, being that such an avenue is proposed as being part of future work.

In the authors' experience, even with neural networks there is a benefit from proper selection of features, given that reducing the weight of useless or redundant features wastes network capacity that could have been used better if such features had not been present to begin with. However, the reviewer has a point in that this remains to be seen in future work. **We have added a mention in the final paragraph of the conclusions, where future directions are discussed, that neural networks may be better able to utilize large datasets.**

True, also, that computer resources can become a limiting factor when all available data are included, so there is an aspect of practicality to consider, but given close margins of

error in the authors' two analysis studies, and lack of input-dependent hyperparameter tuning, the relative skill of the different experiements is questionable. With some tuning, one could imagine different results in Figs. 8 and 9. Many of the "surprising" results might just be owed to noise in the tuning space plus random noise. This last point about random noise holds especially true given that the authors employed an 80/10/10 split for training/validation/testing, rather than a more robust method like k-fold cross-validation, which would better demonstrate that the model is/is not very sensitive to random data assignment. We also have not been shown skill metrics on the training set and validation set, which when compared with the results shown on the testing set would give the reader a sense of variance in the model. If there is high variance in the model, then result comparisons with narrower margins are difficult to trust. The authors should discuss training/validation set results, even briefly, to demonstrate that there is low variance in their model.

We are aware that small differences in the results might be simply noise and therefore have attempted not to overinterpret small differences in the individual results of Figs. 8 and 9, but rather concentrate on the general patterns found in these figures. While there were already some remarks pointing this out in Sect. 4.3, **we have added further discussion to emphasize that the broader patterns are more robust than the individual results**. To support these conclusions, following the reviewer's suggestion, **we have added figures showing the results of the exclusion studies (equivalent to Figs. 8 and 9) for the training and validation sets in the Appendix. In the revised test of Sect. 4.3, we now mention these figures and briefly cross-compare the results between the testing, validation and training sets.**

Specific comments:

Page 2, Line 35: Please include reference to Bedka et al. 2018, "The Above-Anvil Cirrus Plume: An Important Severe Weather Indicator in Visible and Infrared Satellite Imagery," https://doi.org/10.1175/WAF-D-18-0040.1, and Bedka and Khlopenkov 2016, "A probabilistic pattern recognition method for detection of overshooting cloud tops using satellite imager data," https://doi.org/10.1175/JAMC-D-15-0249.1.

These references were added.

Page 3, Line 66: Fix typo, "…as well as an region…"

This was also pointed out by Reviewer 2 and has been corrected.

Page 7, Line 175: Can the authors explain their reasoning for choosing 37 dBZ?

Reviewer 1 also asked about this. 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.**

Page 7, Line 177: If there are more than one pmaxZ with equal dBZ>=37 connected withing 25 km, what determines which pmaxZ get excluded, and how do you ensure the center-most pmaxZ is not excluded?

It is not explicitly attempted to ensure that the centermost pmaxZ is selected; however, the procedure described in Sect. 3.1.1 selects the points with highest dBZ first and therefore the most significant cells tend to be selected except when they are excluded due to being close to even more significant ones.

Page 10, Line 258: The decision on whether to use MSE or MAE should depend on the

importance of outliers in your training and validation sets. If the outliers are "real," that is, if they are not the result of corrupted data and therefore it is important to detect them, then MSE is the correct loss function to use. Otherwise, if the outliers are corrupted data that are unimportant to detect, then MAE should be chosen because MAE gives less weight to outliers. If using MAE rather than MSE, can the authors demonstrate that outliers in their datasets are unimportant.

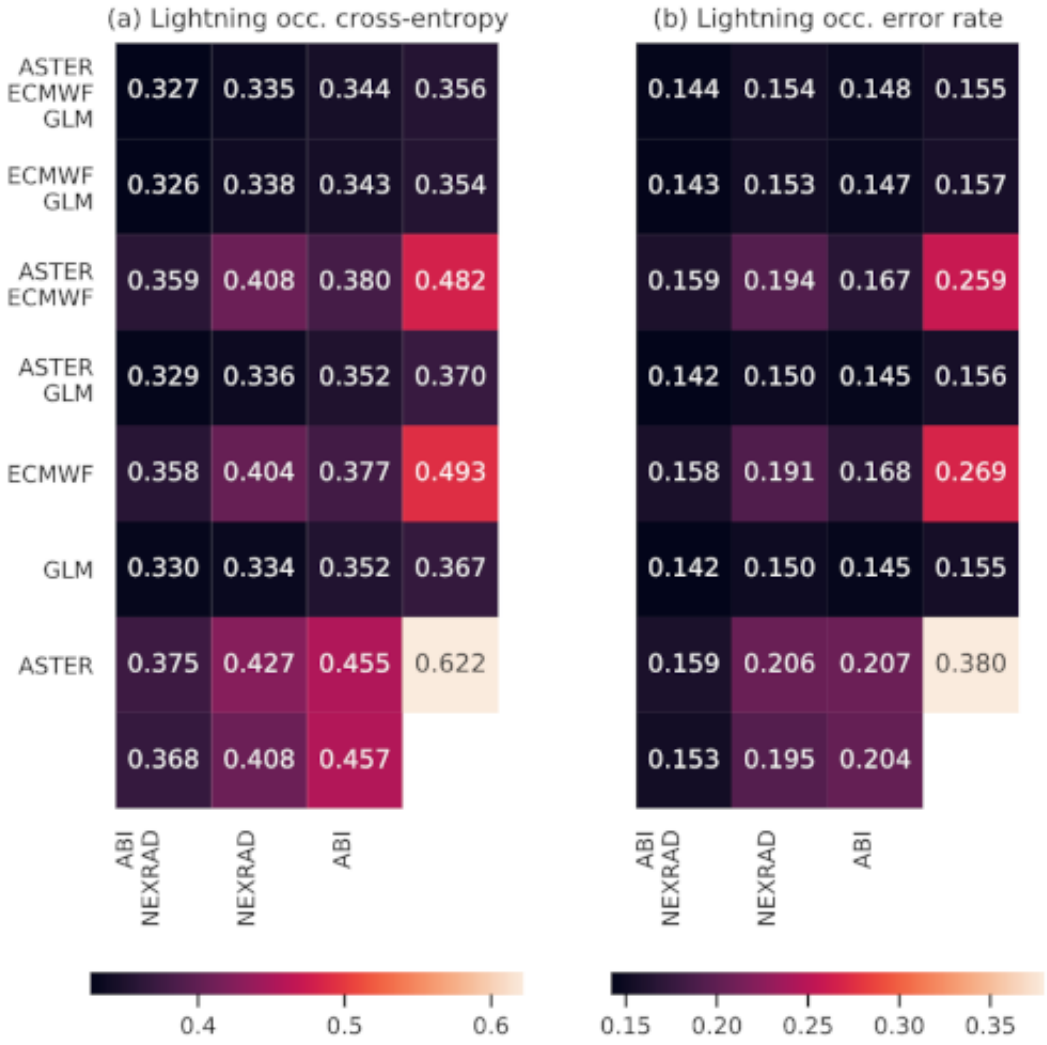
Our study deals with complex multi-source data that are processed with motion-detection and feature-extraction algorithms. Although we made an effort to make this process robust, in such an environment it is to be expected that some outliers will occur due to bad data or failed processing. As the reviewer points out, MSE is more sensitive to these than MAE. Of course, some outliers may also be "real" outliers that result from unusual natural behavior, but in the interest of robustness we believe that it is better to use MAE. One further point that supports the choice of MAE is that a model trained with MAE loss achieved better MSE in the validation set than an equivalent model trained with MSE loss. While we had omitted this mention from the originally submitted version, **we have now added it to this paragraph. We also added some citations on the relative merits of MAE vs MSE.**

Page 10, Line 270: One concern I have is that a given set of hyperparameters is not one-size-fits-all for testing different model setups, which is the main purpose of this article. Changing the data sources in order to assess their importance using the same hyperparameters each time might not be conclusive given that there may be some combination of hyperparameters that results in better performance with one source compared to another. Did the authors use the same hyperparameters for all input sources assessed? That is, when the authors did an "informal manual search of the parameter space," did they do this only for one input source? To be convincing, the authors should search the hyperparameter space for more than one input source (assuming they haven't already) to prove that performance is demonstrably insensitive to the changes, and the relative skill between each case stays consistent.

Doing a thorough hyperparameter search for all input combinations would be very time-consuming and computationally expensive, and our search convinced us that the results were not very sensitive to changes over a reasonable hyperparameter space. However, the reviewer makes a good point that this does not guarantee that this is true for all combinations. To test this, we reran the data exclusion analysis of lightning occurrence from Fig. 8c-d varying two of typically most influential parameters of LightGBM: the depth of the tree and maximum number of leaf nodes. The depth was increased from 5 to 6 (theoretically increasing the predictive power) and the number of leaf nodes reduced from 48 to 24 (theoretically decreasing the predictive power). The results are shown below. These sensitivity experiments show that, even though the individual numbers may change slightly (typically less than 2%, but sometimes somewhat more) the patterns seen Fig. 8c-d remain stable even when the hyperparameters are changed.

The figure below is equivalent to Fig. 8c-d but with the maximum tree depth increased from 5 to 6:

The figure below is equivalent to Fig. 8c-d but with the maximum number of leaf nodes reduced from 48 to 24:



(The blurriness of the images above is due to the maximum 500x500 pixel size imposed by the comment submission system.)

Page 11, Figure 3: Except for the very start of the 'blue' case, these tracked cells do not depict an active thunderstorm given the defined threshold of 37 dBZ. The purpose of the article is to analyze ML-based nowcasting of thunderstorm hazards, so it would be more relevant to see a figure that better satisfies the authors' definition of an active thunderstorm.

Reviewer 1 had a similar comment so we will reproduce the answer here: 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**.

Page 11, Line 287: "… with MAXZ > 37 dBZ …": Like the previous point, most examples in Figure 3 do not show a MAXZ > 37 dBZ. If the MAXZ threshold was reached at some point prior to t=-60 min, then please explain this in the text and/or figure caption.

Please refer to our response to the previous comment.

Page 14, Line 325: Combining the total source importance in this way seems questionable given that you claim a large selection of well-correlated but poor-performing variables add up to an importance comparable to the much higher skill radar variables – also considering the fact that the NWP variables are likely tapping into the same information, and that signal is amplified by being picked up by many variables. With this in mind, can the authors comment further on the value added by the inclusion of the b) and d) figure panels.

The feature importance shown here is defined as the total gain of a given feature, that is, the reduction in the loss function attributed to the gradient boosting model using that feature. Since, by this definition, the gain is additive, it is our understanding that it is appropriate to calculate the total gain of a group of features by adding together the gains of the individual members of the group.

As for why the individual features seem to be poorly performing, consider the following, highly simplified, example: There are two predictors, A and B, which are very highly correlated. When the decision tree creates a split, it chooses either A or B to use as the basis of the split. Since A and B are highly correlated, the tree might use either A or B depending on which one happens to be slightly better in that particular circumstance. Therefore, the gain is attributed near-randomly to either A or B and thus split near-evenly between them. If the model is instead trained using only feature A, it will make every split based on A, and therefore A will be attributed a much higher gain – indeed, nearly the total gain of A and B in the first case.

Therefore, we do not really claim that poorly performing variables add up to a high importance – but rather that the high importance is split between multiple variables, and thus each of the individual variables seems to be performing poorly. For this reason, we think that the b) and d) panels make it clearer to the reader that the model is actually using the ECMWF variable to make decisions, even though the decisions are split between many individual variables. As for why the ECMWF variables have high importance in the feature importance analysis but low importance in the source exclusion analysis, we have already discussed this at length in the submitted manuscript.

We have edited this paragraph to add some clarity about the points presented above.

Page 14, Line 327: The text seems to suggest GLM has more contribution than ASTER, but the Figure 6b suggests the opposite (or appears to). Can the authors please clarify?

While both contributions are very minor, the reviewer appears to be correct. We suspect, though unfortunately cannot verify, that this was text left over that referred to an earlier version of the analysis. **The wording was changed to "the GLM and ASTER data contribute to a lesser extent".**

Page 17, Line 359: Again, the way Fig. 6b was arrived at seems flawed and maybe suggests more importance assigned to ECMWF features than is the case. Figure 6a shows almost no significant skill in inclusion of the ECMWF variables.

Please refer to our response to the reviewer's comment regarding page 14, line 325.

Page 18, Line 360: "… because the other results in Fig. 8a–b do not suggest in any way … ": As a style suggestion, consider removing "other" and "in any way", as they seem unnecessary and detract from the sentence.

We agree that these can be removed without loss of meaning, and **the sentence has been edited accordingly**.

Page 18, Line 370: "… as can be seen by comparing the columns to each other": Similarly, this phrase is unnecessary.

This has also been removed as suggested by the reviewer.

Page 19, Line 389: It would alleviate ambiguity if the authors could explicitly state why not all panels in Fig. 9 have a bottom right corner showing climatology.

This has been clarified in the caption of Fig. 9.

Page 19, Line 402: Grouping features by data source overcomes the burden of testing all possible combinations of input features, but it doesn't solve the problem of understanding the sensitivity of said combinations (which, as rightly stated, would be implausible to determine in this manner). I would suggest simply making clear that the problem overcome is the former one I mentioned, and that this is a reasonable alternative approach.

This is a fair point, **we have reworded the text here as: "Testing all possible combinations of input features would have quickly become implausible as the number of features increased, but grouping the features by data source allowed us to cover the most realistic situations of missing data, where an entire data source is unavailable…"**

Pag 20, Line 411: "… moderate to high importance …" is questionable. Instead saying "… of some importance …" would be more agreeable.

This was reworded as suggested by the reviewer.