Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC3
https://doi.org/10.5194/nhess-2021-171-RC3, 2021

Comment on nhess-2021-171

Anonymous Referee #3

Referee 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-RC3, 2021

General comments:

The authors present a well written article that explores the contribution of different data sources in the nowcasting of impactful weather events by way of decision trees. 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. 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.

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.

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

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

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?

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.

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.

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.

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.

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.

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?

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.

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.

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

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.

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.

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