The review on the manuscript "Improving object-oriented radar based nowcast by a nearest neighbour approach" by Shehu and Haberlandt

The presented manuscript aims to benchmark the nearest neighbor approach for object-based storm nowcasting in Hannover Radar Range, Germany. The study utilizes the database of storm events that have been compiled with a focus on urban hydrological applications. Hence, it is of particular interest for radar-based precipitation nowcasting, urban hydrology, and water management. The declared scientific question is straightforward, the proposed workflow is reliable, and the obtained results are well delivered and discussed. Still, some comments and questions need clarification and, probably, would require additional experiments and analysis from the authors.

Major comments

Research aim

The decision to predict individual storm characteristics, i.e., area, mean intensity, x and y components of velocity, and lifetime, instead of predicting the entire storm evolution as an integral object should be elaborated. In general, I understand the utility of predicting individual characteristics. However, in this way, we miss the detailed information about storm event spatiotemporal evolution and could not precisely estimate neither location nor intensity-related errors. For example, the spatial structure and distribution of rainfall intensities within the storm event are particularly relevant for urban applications. The example of the possible utilization of the predicted (individual) properties could help to clarify their choice.

Dataset

The authors declare that the compiled dataset includes outliers (L198). That leads to the mixed-use of mean or probability-based (e.g., median) statistics. It is pretty hard to recognize and remember where and why the mean or median statistics are used. Moreover, the authors often describe the need to use mean/median for (not) accounting outliers but rarely communicate the obtained results based on that choice. Thus, it is interesting what is the proportion of outliers in the compiled database and could they be removed for the sake of consistency of mean/median statistics throughout the manuscript.

The compiled database of storm events is based on the open data provided by the German Weather Service. Is it possible to share it? It would serve both manuscript's reproducibility and community interests in the field of storm tracking and prediction.

Baseline

The utilization of Lagrangian persistence as a baseline is reliable, and obtained results are interesting to compare. I recommend authors provide additional information about it (how nowcasted is computed etc.) to account for inexperienced readers.

In Section 4.4., the authors compare the closest single neighbor and 30-member ensemble approach. Do the authors mind finding the single nearest neighbor as a more advanced baseline compared with four and 30-member ensemble solutions? It would then pose an additional research question (partially touched in Sect. 4.4) of an added value of ensemble approach compared to the single neighbor.

Information leakage

In modeling studies, it is particularly relevant to isolate calibration, validation, and test datasets to prevent so-called information leakage -- the situation when the information outside the

calibration set is used for model calibration (training). In the presented study, I suspect four procedures that may lead to information leakage:

1. Normalization of events characteristics.
2. Importance analysis and weights calculation.
3. Optimization of the number of nearest neighbors.
4. Splitting into different event groups.

The authors state that normalization and importance analysis have been done “Before training and validating the k-NN method” (L191). In this way, there is an evident information leakage that connects calibration and validation datasets. Also, it is not clear how calibration and validation datasets have been isolated to find the optimal number of nearest neighbors. I do not think that addressing data leakage would change the results much, but it is vital to ensure methodological reliability.

Splitting the database into three groups according to their duration (L312-317, Table 2) was done before the modeling. In general (and in practice), we do not know a priori if the recently appeared storm will be sporadic or last for a couple of hours or more. So, in my opinion, in making predictions, we should use all the examples from the database to find closer candidates to be used for predictions, not only those from the group of a similar duration. That also would open the new directions of analysis, e.g., how would closest examples change with the storm’s evolution. Is the more mature storm similar to storms with comparable duration, or is there some skew in characteristics similarity?

The minor but also critical comment here is about the research code availability. For sure, open code would ensure research reproducibility and provide information on particular details of the computational workflow.

Training, learning, and cross-validation

The authors use terms of training and cross-validation, but, in my opinion, the presented manuscript does not involve both procedures. The nearest-neighbor model is not trained per se; it only uses a bag with historical examples to find one closer to the “storm-to-be-predicted” based on the similarity metric. In this way, the nearest-neighbor model also does not learn anything as it has no parameters to learn. The only parameter here is the number of the nearest neighbors to use for predictions. However, the choice of that number is entirely subjective (see comment above) and is independent of both the “storm-to-be-predicted” and the available examples and their characteristics.

I would also question the use of the term cross-validation. There is no numerical model to validate as the nearest neighbor approach is instead a database search method than a “pure” numerical model.

The authors explicitly communicate the aim of the study as “... to investigate if non-linear relationships learned from past observed storms can surpass the Lagrangian persistence and extend the predictability limit of different storms.” However, as I mentioned above, the nearest neighbor model does not learn anything. It is also an open question if there are non-linear relationships (and what kind of relationships).

Minor comments

- “Birth” → “initialization”?
- “Death” → “dissipation”?

- L98: “k-NN.” The first appearance needs transcription.
- The orientation feature is in degrees. I wonder how the difference between 1 and 359 degrees is considered.
- Interestingly, the area and number of storm cells do not show similar behavior in importance analysis, but they are highly correlated.
- L261-262: “Only neighbours that display a distance lower than 0.5 are selected for both single and ensemble nowcast in order to minimize the influence of non-similar storms.” Any statistics of that?
- Figure 6: “The weights given here are averaged from the weights calculated at three different lead times and storm durations.” However, the authors then mention (L341-342): “Contrary for Total Lifetime and Area, only for storms that last longer than 3 hours, the method is able to converge and give the most important predictors.” However, we cannot see these results.
- L348-349: “is not completely understood and is not investigated further on for the time being since it is outside the scope of this paper.” But, from the abstract and introduction: “i) what features should be used to describe storms in order to check for similarity?” Thus, it is probably in the scope of the paper.
- L354-355: “Moreover, the important predictors do not change drastically from one lead time or storm group to the other, as seen in the PIC” Could we see it from any table or figure?
- Figures: larger fonts and more vertical space between different types of events would be appreciated.