EGUsphere, referee comment RC2
https://doi.org/10.5194/egusphere-2022-599-RC2, 2022

Comment on egusphere-2022-599

Anonymous Referee #2

Referee comment on "Adapting a deep convolutional RNN model with imbalanced regression loss for improved spatio-temporal forecasting of extreme wind speed events in the short to medium range" by Daan R. Scheepens et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-599-RC2, 2022

This paper compares weighting method for addressing the imbalance problem posed by the prediction of rare events in atmospheric data. The weighting methods are reasonable and the computational experiments are appropriate. However, the paper does no go in sufficient details to give insights into whether a weighting schemes is intrinsically more suited than another. In addition, probability-based schemes are presumably subject to uncertainties which are not appropriately discussed.

Major comments

The comparison of WMAE, WMSE (Method 1) and SERA (Method 2) do not seem apples-to-apples. Method 1 still examines all the data points, while Method 2 discards datapoints that are not of interest. In this sense, it seems unfair to compare these two techniques. This issue is not a fatal flaw of the manuscript but should be mentioned and used to qualify the statements that compare both methods.

If balancing data points is important, then errors in the balancing scheme may also be critical. In particular, since probability of rare events are more difficult to estimate than high probability events (higher relative error), how do these errors affect the conclusions of the manuscript? Ideally, these errors should be addressed in the training experiments. If not possible, the manuscript should explicitly discuss this caveat.

The manuscript focuses on describing the results of the experiments rather than explaining the different behaviors observed with the different losses. Could the authors provide insights into why the SERA loss performs better or worse? These explanations would help the readers generalize the present findings. At the moment, it is unclear what would happen if the relevance function was different? What is the effect of the SERA

threshold? What if the SERA loss includes all the datapoints rather than only a few? At least, these details should be discussed. At best, additional experiments would be useful.

The authors go to great lengths to explain what DL architectures are suited for extreme event prediction, while this is not the main focus of the manuscript. It would be preferable to expand on the loss balancing mechanisms for extreme event prediction. Here are a few possible references to discuss.

For capturing the tails of PDF, log transformation has been proposed (Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems"), or by rebalancing the dataset itself which is useful for classification tasks ("A study of the behavior of several methods for balancing machine learning training data") and regression tasks ("Uniform-in-Phase-Space Data Selection with Iterative Normalizing Flows").

It is unclear how the ensemble is constructed. Did the authors average the predictions ? Are all models identically weighted?

The title is not descriptive enough and may not help many readers. The title should explicitly reflect that the prediction of extreme events is addressed by balancing the loss.

Minor comments

P1 L5: "It has become […] challenges". This sentence needs to be in the intro not the abstract and could use a citation

P3 L54: No need for quotes around Weibull

P3 L80: "Given to the model" is colloquial. Please use something like "are processed by

the model”

P3 L84: Remove “excellent”

L 251: citation missing for Pytorch

Sec.2.3: Validation would be better. Verification may carry a different meaning in computational science.

Rephrase paragraph around L265 and possibly try to be more quantitative. What does ``messy” forecast mean?

Table 2: what happens if the architectures become even more complex?

Since the SERA loss depends on a threshold, please indicate what threshold was used as a subscript of `SERA` when results are displayed.

Figure A1 is actually very informative and could be used in the main text. It would be useful if the authors could also show the integral time scale on that graph.

The discussion around frequency bias could benefit from an equation that showcase how the frequency bias is calculated.

L 446: Did you mean “successor” instead of “predecessor”?