EGUsphere, referee comment RC1
https://doi.org/10.5194/egusphere-2022-599-RC1, 2022

Comment on egusphere-2022-599

Anonymous Referee #1

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-RC1, 2022

The authors investigate ConvLSTM-based models for wind speed prediction at lead times up to 12 hours, motivated by energy applications and with a focus on Europe. The central contribution of the paper is an investigation of different types of loss functions with the aim to improve predictions of extreme events.

Overall the paper is well-written and easy to follow. While it presents a new and interesting perspective on the training of LSTM models, there are several key issues in my view. Most importantly, in my view, the investigation and evaluation of forecasts of extremes should be better motivated and connected to the intended applications; and the description of technical details of the model need to be improved. These and additional comments are detailed below.

Major comments

- Perhaps the most important issue to me is that the investigation of extremes, and the specific approach taken in the paper, should be motivated better and be connected better to the application:
  - For example, the focus is on the relative rarity at each coordinate (page 7), rather than on the exceedances of thresholds, which is stated as the motivation in the introduction for shutting off turbines at certain thresholds. While I understand that considering relative rarity makes the modeling easier, it should be better explained why this would be practically relevant for wind energy applications. More generally, what is the aim of the proposed models? Should they work for all outcomes, but be better at predicting extremes? Or should they specifically focus on extremes, but not care so much about non-extremes? Relevant literature from statistics on evaluation (e.g. Brehmer and Strokorb, 2019, DOI: 10.1214/19-EJS1622, and references therein) could be consulted as a starting point for discussions on this aspect.

    - What is the potential application of the forecasts as they are produced by the models in the paper in practice? While improving predictions for extremes, the quality for non-extremes will likely get worse. From an application perspective, what would be specific economic situations of users that would motivate the use of the proposed models.
    - Regarding 2., a more practically useful approach to me seems to be a model that predicts probabilities of the exceedance of critical thresholds for shutting off turbines. This would be directly connected to different kinds of potential losses, and probabilities allow for optimal decision making in applications.
- All considered models are relatively complex (ConvLSTM models accounting for the spatial structure9). To be able to compare models in a fair way, more simple benchmark models should be considered as well. In particular, a grid-point wise standard NN model that uses previous time steps locally at the single grid point of interest, or similarly motivated local, per-grid-point LSTM models should be included in the comparison to be able to evaluate whether the temporal and spatial aspects of the proposed models are truly relevant.
- The model description in Section 2.2 is rather short and does not include all relevant details to independently replicate the work. For example, which fields are exactly used as inputs (only wind speed? which previous time steps?), what does "12 hour input and 12 hour prediction" mean (page 8, line 209)? Are the predictions made hourly? Which inputs are exactly used at each time step?
  Some more details are provided in Section 2.2.3, but it does not become clear how you selected the hyperparameters (optimization algorithm, learning rate, ...). Did you try different values and how robust are the results in terms of these choices?

Minor comments

- The last sentence of the abstract should be moved to the acknowledgements section.
- The literature review should also refer to probabilistic predictions, which are of critical importance for extremes. For example, there has been lots of work recently on probabilistic energy forecasting, in particular in the domain of combining physical-statistical hybrids. With regards to post-processing NWP models, the recent of of Phipps et al (2022, https://doi.org/10.1002/we.2736) appears to be relevant in the context of the discussion on page 3, line 68-71).
- page 4, line 107-114: The discussion of data-driven weather forecasting seems to not be relevant for the remainder of the paper, as the applications consider rather different time scales and variables.
- page 6, line 156: Why did you use 1000 hPa wind fields instead of the surface wind fields? Wouldn't this be a more relevant target for energy applications?
- page 7, line 185f: Doesn't the normalization performed here implicitly assume Gaussianity? As an alternative approach, it would have been possible to simply select the per-grid-point quantiles from the climatology of historic observations in the training set.
- There seem to be a lot of options for choosing the weight function in equation (3): Have you tested any alternatives and performed comparisons? How much do the results depend on the specifities of the definition of this weight function?
- page 12: Ferro and Stephenson (2011) argue that SEDI should only be applied to calibrated forecasts to guarantee convergence to a meaningful limit for rare events in the sense that the number of predicted events equals the number of observed events. Is this the case for the models considered here?
- page 12: Would it in principle also be possible to incorporate the SEDI loss in the model

estimation, similar as recently proposed for the FSS in Lagerquest and Ebert-Uphoff (https://arxiv.org/abs/2203.11141)? This again relates to the question of what the actual goal is here for the models.

- Table 2+3: Are the results averaged over lead times?
- page 15, line 342: Why does the ensemble not include all 5 models, i.e. also the MAE and MSE based ones? This 5-model ensemble should be added to the comparison.
- page 20, line 409f: The shuffling procedure does not become completely clear here: Do you shuffle full fields, or also simply grid points within a field?
- page 20 / Figure A1: Why is the evaluation here based on RMSE, rather than SEDI scores as before?