Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., author comment AC2
https://doi.org/10.5194/gmd-2021-276-AC2, 2021

Reply on RC2

Julian F. Quinting and Christian M. Grams

Author comment on "EuLerian Identification of ascending AirStreams (ELIAS 2.0) in numerical weather prediction and climate models – Part 1: Development of deep learning model" by Julian F. Quinting and Christian M. Grams, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-276-AC2, 2021

Response to Reviewer 2

This study offers a well thought out alternative to a previous simple statistical model (logistic regression), and represents a significant improvement over the previous implementation. The Convolutional Neural Network (CNN) allows the consideration of non-local spatially dependent predictor/predictand variables, and is readily more interpretable than the logistic regression model. Generally, this is a strong paper, and subject to the minor comments below, the manuscript should be published. My main concern is with the with the generation of figure 4, which I detail in the minor comments below. Additionally, please add a section on limitations of the method, caveats, and future improvements that could be applied to this work.

Dear Reviewer,

We are very grateful for your overall positive feedback and the thought-provoking comments on our manuscript. We agree that the method still comes with certain limitations, and ideas are around for future improvements. Thus, we will include a section in the revised manuscript discussing these aspects. In the following, we respond point by point to your minor comments. Our responses are highlighted in bold.

Kind regards,

Julian Quinting and Christian Grams

Minor Points

L55-L64 This is a good justification for a computer vision based machine learning approach.

Thank you for this positive feedback. We will add the key-word „computer vision based machine learning approach" to this part of the manuscript so that it connects directly to the next paragraph of the manuscript.

L66-67 this feels like a bit of a misrepresentation, I would change to say "CNNs identifying salient features in the input space which influence the desired prediction."

We fully agree. We will change the manuscript accordingly.

L72. I would say that it is "originally designed" as a semantic-segmentation model, as it's applications are now much further reaching.

Thanks for this comment. We will modify the manuscript as suggested.

L106 How much model degradation occurs without this 5th predictor? Figure 2 seems to indicate that the seasonality is not a big factor, as much as including time-lagged ascent information. Figure 7. Confirms it is not a factor. This seems like something that needs to be explored or commented on further. Is this due to the normalization around the date of interest, and the selection of data around the forecast date. I think it is worth testing whether this variable affects the final skill of the model when you are not selecting data in a 30-day randomized window. Or de-emphasize this line in the introduction in general, as you immediately remove this variable as a predictor.

Thank you for this comment. We absolutely agree that Figs. 2 and 7 indicate that the 5th predictor is of minor importance for the models' skill. When designing the CNN-based models we hypothesized to see a benefit when incorporating the climatological occurrence frequency as a predictor. The results do not confirm our initial hypothesis which we attribute to the normalization of the predictors around the date of interest. So, it seems that the models rather learn the seasonality indirectly from predictors 1-4 than from the 5th predictor. Thus, we will de-emphasize this aspect in the introduction.

L157. The non-linearity is not necessarily required. Has it been tested to use linear activations? This would give you an idea of the linearity of the actually predictor/predictand relationship. You have two competing predictor improvements in this model (compared to local logistic regression) 1) the addition of a spatial component via convolution 2) the nonlinear predictor/predictand relationship. It would be good to test what is a bigger factor for model improvement, my inkling is the spatial information is more valuable.

Thank you very much for this interesting comment. Indeed, the logistic regression model suggests that a linear relationship exists between predictors and predictands. Since we have not tested yet using linear activations, we will follow your suggestion and will test to what degree the models' skill changes with either approach.

L165. The debate over the efficacy of dropout is distracting to this paper. I would take it out.

We will remove this part of the sentence in order to not distract from the main

content of the paper.

L208. Please specify what dataset the MCC threshold tuning (0.05-0.95) tuning was done on.

The threshold tuning was done on the validation data. We will include this information in the manuscript.

L245. Readers would benefit from a quick summary of Quinting and Gram's (2021) logistic regression model.

Thanks for this suggestion. We will provide a more detailed description of the logistic regression models in the introduction of the revised manuscript.

L255 The authors do not define why +- 10% is considered perfectly reliable (nor do they test via any subsampling), either justify this more clearly, or I would suggest adopting the Bröcker and Smith reliability diagram framework (Bröcker, J., & Smith, L. A. (2007). Increasing the Reliability of Reliability Diagrams, *Weather and Forecasting*, *22*(3), 651-661.)

Thank you very much for pointing us to the suggestions of Bröcker and Smith (2007). We will adopt their approach by including consistency bars created via consistency resampling. Indeed, this will provide more quantitative information on the quality of the probabilistic forecasts.

L270-280 Can you justify why this process should be performed on the testing dataset and not the validation dataset? It appears as if this is tuning a hyperparameter, and you are increasing your model bias skill on the testing data. It seems like the thresholds should be determined on the validation data as you don't plan on running the expensive Lagrangian framework model when implementing this CNN in the future. This seems concerning for this figure.

We agree that it is indeed more intuitive to determine the thresholds on the validation data. The issue we encounter here, however, is that due to the low climatological occurrence frequency of WCBs the thresholds become spatially highly variable. Accordingly, two neighboring grid points may be classified as non-WCB and WCB despite exhibiting nearly equal conditional probabilities. Thus, taking the longer testing data period yields spatially less variable thresholds. Further, with the relatively short validation period the results may be affected by possible long-term variations of the WCB occurrence frequency. Both aspect will be discussed in the revised manuscript.

Grammar edits.

L70 missing space "intrusions (Silverman..)"

We will correct for this.

L370 remove “aims to” --> “UNet CNN that identifies”

We will modify the text accordingly.