Weather Clim. Dynam. Discuss., referee comment RC2
https://doi.org/10.5194/wcd-2022-55-RC2, 2022

Comment on wcd-2022-55

Anonymous Referee #2

Referee comment on "Improved extended-range prediction of persistent stratospheric perturbations using machine learning" by Raphaël de Fondeville et al., Weather Clim. Dynam. Discuss., https://doi.org/10.5194/wcd-2022-55-RC2, 2022

General Quality:

The research presented in this manuscript contributes to the scientific progress of S2S forecasting by leveraging the impact and understanding the dynamics of sudden-stratospheric warmings (SSWs). The authors present a novel three step procedure that combines existing data-driven machine learning techniques to enable a more extensive assessment of SSW dynamics. Using the gained insight they increase the performance of numerical ensemble forecast for lead times above 25 days. Overall, the data preprocessing, the dynamical analysis, the prediction task as well as the post processing are thoroughly motivated and the authors present strong results with according statistics (Cross-Validation) to support their conclusions. As most methods are described clearly and in-detail, the work appears to be reproducible. I have technical questions, regarding the ML forecast, more specifically the neural network, and minor general concerns about the presentation of the results. Should these specific comments be addressed, I recommend publishing this work after minor revisions.

Specific Comments:

1. *ML forecast using a deep neural network (DNN):* In general your work impressively outlines, a reliable application of machine learning techniques in research, as you leverage physical knowledge to present a concise learning task to the network. However, for me some questions and concerns remain. First, the MLP forecasting is not reproducible, due to the very limited description of the training parameters, input and output dimensionalities and other specifics. Even though to me this is not necessarily main body material, including the information in the appendix or supplementary material (maybe even code or a trained model) is necessary for reproducibility. My second concern is the lack of statistics for the results of the MLP. DNN results sacrifice reliability due to cherry picking, i.e. training only one network. I suggest, for example retraining the model several times given different initial parameters, i.e. deep ensemble approach, yielding more substantial results. In addition, this procedure provides you with the lacking ensemble information,

discussed in line 318-320. Lastly, such ensemble approachs as well as the field of Bayesian Learning, I would reframe the statement circa line 317, where you address the default of point forecast for machine learning algorithms, as it is only partly true. Also, I would suggest adding the according citation that supports your argument in line 318-320.

2. *Post-processing Equation:* Overall, the description of calculation procedure as well as detailed equations help to understand and reproduce the results. The improvement I want to suggest, concerns the post-processing used to enhance the numerical ensemble forecast. I think combing existing descriptions with an equation would complete the paragraph.

3. *Visualization:* While the figures in your work provide strong and straight forward visualisation, especially Fig. 3 can profit from improvements, as well as the discussion of Fig. 5. In terms of Fig. 3, I agree with the Comment (RC1) of Reviewer 1 and have nothing to add. Regarding, Fig. 5, during the discussion of the results you do not specifically mention which of the three plots you address, which sometimes makes it hard to follow the conclusions. Thus, some of the visual arguments do not become evident right away. My recommendation is either a more descriptive wording or a more distinct visualisation.

Technical Corrections:

In terms of technical corrections, the preprint would profit from overall larger figures, especially Fig.1 and Fig.3. Even in case of the current figure size, Fig. 1, 2 (left plot), 3, 4 (right plot) and Fig. 6 have fairly small tick labels on both x- and y-axis. Furthermore Fig. 1 would greatly improve if plotted with a joint colorbar, that also indicates lowest and largest values (with increased tick-size for all values).

In l. 324 I think the authors wanted ‘machine algorithms’ to say ‘machine learning algorithms’.

Finally, I agree with the technical comments of RC1.