Review: WCD 2022-55

Title: Improved Extended-Range Prediction of Persistent Stratospheric Perturbations using Machine Learning

Authors: Raphaël de Fondeville[1], Zheng Wu[2], Eniko˝ Székely[1], Guillaume Obozinski[1], and Daniela I.V. Domeisen

General Quality:
In general the revision improved the quality of the paper, which I suggest be accepted subject to technical corrections. The authors resolved all specific comments in my review and improved the visualisation quality.
With the final edits, this work is an important contribution to the scientific progress of S2S forecasting and demonstrates reliable integration of physics-based machine learning into climate research. The presented three step procedure provides novelty in that it combines existing data-driven machine learning techniques with current physical research findings to enable a more extensive assessment of SSW dynamics. Moreover, the procedure facilitates increasing the performance of numerical ensemble forecast for lead times above 25 days.
Overall, each research step in this work is thoroughly motivated and discussed, enabling reproducibility. The authors present strong results with according statistics (Cross-Validation) to support their conclusions.

Technical Corrections:

1. *ML forecast using a deep neural network (DNN):* Although, the added sentences (l.322-325) regarding Bayesian NNs and deep ensembles provide the appropriate background, I still suggest refraining from strong statements, i.e. 'whose link with numerical ensembles is still not well understood. ' without providing respective citations. I suggest adding context literature and maybe softening the statement as it is a topic of ongoing research. See for example [Abdar et. al. 2021] (https://www.sciencedirect.com/science/article/pii/S1566253521001081), who discuss the topic in detail also theoretically referring to an MLP with one hidden-layer such as yours.
2. *Visualisation:* While Figure 3 has undergone major improvements, unfortunately you have now dropped the PC labels for the individual plots. I suggest adding PC labels or adding A,B,C,D in the Figure and according assignment (A is PC1, etc.) in the caption. Otherwise, all visualisation adjustments are satisfactory
3. *Citation of Kretschmer et.al 2017:* Thank you for the clarification. I think that the statement of the authors is not completely true. The approach described in Kretschmer et al. discusses multiple testing and accounts for it as described in the SI (p2, Robustness of the causal precursor detection scheme): *"Note that the causal discovery algorithm involves multiple hypotheses testing such that the significance parameter α should be considered as a hyper-parameter of the algorithm. As a conservative confidence estimate for a predictor (which is not really relevant for prediction), one can use the maximum p-value among all tested conditions in the PC algorithm. We also tested whether the obtained predictors are significant if we additionally control the false discovery rate (FDR) and found the lag-1 SPV index and v*T*100 over Eurasia (Fig. 3a) to be still significant (adjusted p <0.01) but the other two region (Fig. 3a) slightly dropping in significance (adjusted p ≈ 0.15)."* Thus, I suggest softening the statement but do not see this as a requirement for the publication of this manuscript.