wcd-2022-55 - Response to Reviewers

The authors are extremely grateful to both reviewers for having carefully read a second time their manuscript. Thanks to their help some typos and imprecision were corrected increasing the quality of the publication. All modifications are highlighted in red in the revised version of the paper and Figure 3 has been updated following the reviewers comments.

1 Reviewer 1

- The authors have responded seriously and thoroughly to all of my comments. Figure 3 in particular is much more illuminating. However, this has created one new confusing point: the time-delay embedding is said to contain N = 60 lags, but the horizontal axes in Figure 3 stretch for over 200 days. This is in contrast with the first version of Figure 3, which spanned only 60 days. Has the lag time increased in the new draft, say, to 240 days?

  Thank you so much for spotting the mismatch! It is still 60 days but Figure 3 axis was showing 'time steps' (every 6 hours) instead of actual number of days. We have fixed this issue in the new version.

2 Reviewer 2

- 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/S156625
  who discuss the topic in detail also theoretically referring to an MLP with one hidden-layer such as yours.

  Thank you very much for the suggestion, we softened the claim and added references. The sentence reads now 'Modifying machine learning algorithms to output probabilistic forecasts is possible but requires either advanced techniques such as Bayesian computation or models ensembling. The link between numerical ensembles and probabilistic forecasts is an

active field of research (Collins et al., 2012; Rougier and Goldstein, 2014), thus in this exploratory study, we focus on classical ML algorithms, leaving probabilistic modeling for future work.'

- 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

  Thank you for the spotting this omission, we have added the forgotten labels.

- 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 $p \approx 0.15$)." Thus, I suggest softening the statement but do not see this as a requirement for the publication of this manuscript.

  Thank you very stressing all these details from their work. In their light, we agree that the sentence's claim should be lowered. The sentence now reads : 'their approach is efficient and provides convincing results but cannot scale to very large problems such as ours where we jointly analyze multiple levels and, being a two-step procedure, their methodology requires controlling for multiple testing, which, if not properly adjusted, is susceptible to selection bias in high-dimensional setups.'