Review for "Decomposition of skill scores for conditional verification -Impact of AMO phases on the predictability of decadal temperature forecasts" by Richling et al

Richling et al propose a decomposition of mean skill scores as weighted sums $SS = \sum_i^{\mathcal{D}} W_i SS_i$ of the skill scores SS_i for non-overlapping subsets $\{i \subset \mathcal{D} : \uplus i = \mathcal{D}\}$ of the data, with the weights W_i given by the proportion of the data in each subset times the performance of the reference forecast for each subset relative to that for the full data. The decomposition is straightforward, as it derives from the associative property of addition. The authors use toy examples to examine how the weights W_i modulate the skill score contributions SS_i to the overall skill score over $\mathcal{D}$, and implement this methodology on predictions of 2m air temperature with the MiKlip system conditional to the 3 phases of the Atlantic Multidecadal Oscillation (AMO).

I agree that such an approach could be helpful to provide insights when evaluating forecast mean skill scores, and, despite its simplicity, I'm not aware of such a decomposition discussed elsewhere. However, the paper needs substantial improvement and can be made more concise. I thus recommend the authors to address the following comments before their paper can be considered for publication in GMD.

Major

1. The key to the decomposition in Eq. 2 is that, for the verification score $\mathrm{S}_n = \mathrm{S}(f_n, o_n)$ and the mean score in Eq. 1 (denoted $\overline{S}$ here), we have:

$$\overline{\mathrm{S}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{S}_n = \sum_{i=1}^{K} \frac{N_i}{N}\left(\frac{1}{N_i}\sum_{n=1}^{N_i} \mathrm{S}_n\right) = \sum_{i=1}^{K} \frac{N_i}{N}\overline{\mathrm{S}}_i \tag{1}$$

   with $N = N_1 + \ldots + N_K$. I suggest indicating this in Eq. 1 to guide/justify the abrupt second equality in Eq. 2. Note the use of the overline to indicate the mean over sample. In the paper, both verification score and mean score are denoted the same, which is confusing.

2. Section 2 describing the decomposition of skill scores is too long. It is divided into 6 subsections, which I don't think is necessary. Consider a smaller section showing Eq. 1-3 and giving short descriptions for each term in the decomposition (no need to repeat the expressions that define each term in separate subsections). For example, L118-121 in subsection 2.2.4 can be moved right after Eq. (3), indicating which term in Eq. 3 is referring to, and each term then named and described. Also, statements like (L102-103) "In Sect. 3, this term can be found ...", or (L107-109) "Consequently, a situation..." can be deleted as they do not seem to add much to the description. The bottom line is that the decomposition is straightforward and each term is self-evident, so they do not need much discussion.

3. Section 3 describing the toy example can be shortened too. In particular, Figs. 1, 2 and 4 can be condensed into one easy-to-read 4-panel figure for cases A0-A2 and B0-B2, indicating

on the figure the terms in the decomposition (Eq. 3). For example, panel 1 and 2 can show the two panels in Fig. 1 for SS_1, SS_2 and SS, whereas panel 3 can show the panel in Fig. 2 for W_{ref_1} and W_{ref_2}. Because $W_{freq_i} = 0.5$ for $i = 1, 2$, there is no need to show these, but can be mentioned in the caption too. Finally, panel 4 can show Fig. 4 for the contributions W_1SS_1 and W_2SS_2. Figure 3 doesn't seem to add much insight and could be deleted.

4. It is unclear whether any of the contributions to forecast skill from the 3 subsets is statistically significant. Given the short period available, each subset has about 17 years on average (the actual number of years is different for each subset and determines the frequency weighting). These are small samples and it is unclear whether the results and conclusions are robust. Can the authors comment on this? I suggest adding confidence intervals to the barplots of Figs. 1, 2 and 4 (perhaps condensed into one figure; see comment # 3), and stippling for the maps of Fig. 5b-d. Perhaps this could be done with the bootstrapping method used for Fig. 5a.

5. L249-250 The authors removed the linear trend from the SST average over the North Atlantic region to define the AMO index. It is known that this approach confound the internal variability of AMO with the underlying forcing signal from greenhouse gases and aerosols. A common and easier approach to derive the unforced index is to remove the SST global mean anomaly from that in the AMO region (e.g., Trenberth et al 2006 [doi.org/10.1029/2006GL026894]). Or, perhaps an even more accurate approach is that of Deser and Phillips (2021) [doi.org/10.1029/2021GL095023]. Do the results/conclusions for the MiKlip decadal system assessed here depend on how the AMO index is computed?

6. The three subsets in the analysis are determined by the phase of the AMO index in the observation-based dataset. However, it would be useful to know how the AMO is represented in the MiKlip system itself. I suggest to include an evaluation of the AMO in the decadal predictions, or provide a reference if this was done elsewhere. The point is, how relevant is the analysis if the forecasting system is unable to represent the AMO?

Minor

1. Please use a continuous line numbering to facilitate future revisions.

2. L4 Aim → The aim

3. L5-7 Consider rephrasing. To be precise, the overall skill score is decomposed into a sum, where each term of the sum is the product of 3 components as described. It is more telling perhaps to say that the overall skill score is decomposed into a weighted sum, where each term is ... (and then can go on about describing the weights and partial skill scores).

4. L10 Atlantic Meridional Oscillation → Atlantic Multi-decadal Oscillation

5. L12 due to performance gain → due to contributions
6. L13 a positive AMO phase → the positive AMO phase
7. L14-16 Delete "sophisticated". Perhaps "insightful" instead?
8. L25 ccore → score
9. Section 3 is described as showing results for "synthetic time series". Unless I missed it, there are no such time series, and the example simply feeds values conveniently into the decomposition of Eq. (3), as per Tables 1 and 2. If that is the case, please clearly indicate so and avoid the somewhat misleading terminology "synthetic time series".
10. I may have missed it, but clearly specify early on in Section 3 that for the toy example $S^{\text{perf}} = 0$.
11. L84 Delete "etc"
12. L87 is → are
13. L112 ajust → adjust
14. L123 degeneration → degradation
15. L128 What synthetic dataset? See item 9 above.
16. L140 Delete "assumption"
17. L150-151 Rephrase or delete "As a first guess from seeing the skill scores ..." Why one would think so? It is trivial that the sum of the skill scores in the subsets is not the skill score over the full set, nor the arithmetic mean in general. What is somewhat less clear is what the weighting is, which is addressed in the paper.
18. L159 we simulate → we consider (?)
19. L193 subset $i = 2$ (SS_1) → subset $i = 2$ (SS_2)
20. middle; → middle.
21. L200-201 Isn't the increase from -0.36 to 0.18, with $\Delta SS_1 = 0.54$ instead?
22. L220 componentes → components
23. L241 Typically, "hindcasts" refer to retrospective forecasts, which are model runs initialized from observation-based climate states, whereas "uninitialized predictions" refer to historical simulations (for past climate) or projections (for future climate) which are not initialized from observation-based climate states and have internal variability that are not expected to match observations. I recommend using this terminology and change "both hindcast sets" to, e.g., "both sets of predictions". This applies to other cases throughout the text. In particular, "initialized decadal hindcast" is redundant and "uninitialized hindcast" should be avoided (L284).

24. L251 time period → period (this applies to other instances in the text)

25. L259 Can the authors be more specific on how they divide the datasets? In particular, is the ensemble mean used to determine the terciles for hindcast and simulations?

26. L265 How is $Y_{j,t}$ obtained. Is it by simply counting the ensemble members in each category (then dividing by the ensemble size) for a given initial year? Please clarify.

27. L291 for Fig. 5b-d, please clarify if these "contributions" refer to $W_i SS_i$ or just SS_i. Clarify also in the caption to Fig. 5.

28. L293 W-EU and C-EU haven't been defined. They are defined in L296.

29. L297 "... with certain AMO phases identified in previous studies". Provide references.

30. L302 "a ... RPSS of 0.3 is achieved" → "RPSS=0.3". At the very least, delete "clearly positive".

31. L307 Given that there is contention on whether AMO may be influenced/determined by external forcing (e.g., Mann et al 2020 [`https://doi.org/10.1038/s41467-019-13823-w`]), and because the AMO phases used here are from the observation-based data, perhaps rephrase to "... uninitialized reference is not influenced by the AMO phases in the observations".

32. L309 Fig. 6a → Fig. 6b (?)

33. L310 Fig. 6d → Fig. 6c (?)

34. L311 Is this contribution statistically significant? A value of 0.08 doesn't seem like a "large amount", even though it is larger than the other two cases. See Major comment #4.

35. L359 Target → The goal (?)

36. L368 Here and elsewhere the authors use terminology like "positive AMO phase initialization". This is unclear. Consider changing to e.g., "forecast initialization during the positive phase of AMO".

37. L380 I may have missed it, but I don't think OHT was defined before.

38. L393 Delete "quite" and "anyway"

39. L391-396 Rephrase. This statement is convoluted and can be made clearer.

40. L406-407 Can the authors expand on how this work relates to: "forecast uncertainty can be quantified and eventually the forecast can be rated as more precised"? I fail to see a clear connection between this work and the quantification of forecast uncertainty.