EGUsphere, referee comment RC1
https://doi.org/10.5194/egusphere-2022-914-RC1, 2022

Comment on egusphere-2022-914

Mathias Hauser (Referee)

Referee comment on "Understanding pattern scaling errors across a range of emissions pathways" by Christopher D. Wells et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-914-RC1, 2022

Review of "Understanding pattern scaling errors across a range of emissions pathways"

Wells et al. consider errors in pattern scaling for different emission scenarios. They decompose the error into timeseries and pattern errors to better understand their sources. This is highly relevant due to the emergence of climate emulators used to estimate local impacts of emissions - often for scenarios the emulators were not originally trained on. Overall the manuscript is well written and clear. However, there are some points I would like the authors to clarify.

Main points

In the data part I missed that you calculate anomalies of the predictor and target variables and I also see no mention of the reference period. Please also explain how you deal with ensemble members. Do you use one or many per model? How do you estimate the local slope for models with many ensemble members? How do you avoid giving more weight to models with more ensemble members?

Please emphasize more that you use only part of the full MESMER emulator.

You mention several times that using patterns to extrapolate are worse than to interpolate and cite a number of studies showing this. However, I miss a citation of Beusch et al. (2022) who also discuss this. Further, the MESMER emulator has been already extensively evaluated (Beusch et al., 2020a, 2020b) and I think your paper would benefit from discussing this and showing how your paper goes beyond the state of the art.

It's interesting to see that scenarios with a peak in the global mean temperatures show local time lags and would profit from additional predictors. Can you speculate how much the missing MESMER components (i.e. the auto regression) would help alleviate this problem?

Consider changing the way you show significance. I was first confused why you would subtract the standard deviation from your difference signal - I overread the word "magnitude". Therefore I suggest you do one of the following:
(i) switch to showing significance with a test statistic and hatch the non-significant areas in your plots. This should reduce the number of figures and plots without losing (much) information. (E.g. by using a Wilcoxon Mann-Whitney U test and accounting for the large number of conducted tests, by applying the approach of Benjamini and Hochberg (1990), see also Wilks (2016)).
(ii) If you keep your current approach I strongly suggest to make it more clear - add vertical bars in the title of the figures to make it clear that it is the magnitude of the difference and also explain what values larger, smaller and almost equal to zero mean at around L210.
(iii) Instead of subtracting the standard deviation could you divide by the inter model standard deviation. That would seem more intuitive to me.

Minor Points

L131: Why is "pattern scaling more accurate than the timeshift method"? Wouldn't the latter allow for non-linearities?

L143: The intercept will also depend on how the anomalies are calculated (and how ensemble members are treated).

L212: Explain that the pattern averages to 1 globally per design and only because of the investigated variable is tas.

L245: “pattern difference is not as robust between models” that is an interesting way to put it. Isn’t it good for pattern scaling if there are few regions with strong differences?

Figures

General: many of the color scales you show saturate on a large part of the maps. Consider widening the shown range to allow distinguishing the patterns better. Please write the labels and units as “Error (K)” instead of “Error / K”. Then it looks less like a division.

Figure 1: I appreciate that you showcase the different errors in an example. However, I think using a scenario that is symmetric in its global temperature makes it more difficult to understand than necessary. Consider showing a non-symmetric scenario, e.g. just increasing the temperatures from 1°C to 2°C until the end of the century.

You could also consider switching the first and second columns. If I understand this correctly the (current) middle column is the “forcing” for the emulator while the (current) first column is the “response”, so switching them could help clarify this relationship.

Figure 2 and 3: Panels a) and b) don’t have a diverging scale and should therefore not feature a diverging colormap. Please use one with a sequential color map. (If you want to emphasize deviations from 1 you can keep a diverging color map but you should mention this and also use another color map as for c) and d)) Depending on what you decide on showing significance, also consider changing the colormap of d) to indicate it shows something else than c).

Figure 4. I’d be interested to see how similar the pattern in b) and c) are, the saturation in b) makes this difficult.

Figure 7: I suggest you label the “Target” below the axes and to maybe not rotate them by 45° (they might just have enough room) - up to you. Please add % as units to d).

Figure 8: The black vertical lines described on L416 are missing.

Text

L7: delete “multiple”

L7: delete “a few”

L26: Expand “IPCC AR6 WG1”?

L38: Maybe delete “change”

L80: Expand “RCP” and explain what this is.

L85: “than the RCPs” -> “than any of the RCPs”?

L84: “remains to be done” consider rewriting

L100-L104: Make it clearer that these are your two assumptions (e.g. turn it into a list or add (i) and (ii)).

L102: timeseries -> temporal

L103: simply modified -> scaled

Section 2.1: I highly recommend to split this into two sections one on the data and one on MESMER.

L136: against the smoothed -> against smoothed

L137: parameter -> “slope” or “scaling factor”

L137-L138: This … SSP119: The sentence sounds off.

L205: You explain the pattern in panel b) first. Consider reordering.

L349: Upon even only -> Even for

L419: Consider rewriting the introductory sentence.

References

Hochberg, Y. and Benjamini, Y. (1990), More powerful procedures for multiple significance testing. Statist. Med., 9: 811-818. https://doi.org/10.1002/sim.4780090710

Beusch, L., Gudmundsson, L., and Seneviratne, S. I.: Emulating Earth system model temperatures with MESMER: from global mean temperature trajectories to grid-point-level realizations on land, Earth Syst. Dynam., 11, 139–159, https://doi.org/10.5194/esd-11-139-2020, 2020a.

Beusch, L., Gudmundsson, L., & Seneviratne, S. I. (2020b). Crossbreeding CMIP6 Earth System Models with an emulator for regionally optimized land temperature projections. Geophysical Research Letters, 47, e2019GL086812.

https://doi.org/10.1029/2019GL086812

Beusch, L., Nicholls, Z., Gudmundsson, L., Hauser, M., Meinshausen, M., and Seneviratne, S. I.: From emission scenarios to spatially resolved projections with a chain of computationally efficient emulators: coupling of MAGICC (v7.5.1) and MESMER (v0.8.3), Geosci. Model Dev., 15, 2085–2103, https://doi.org/10.5194/gmd-15-2085-2022, 2022.

Wilks, D. S. (2016). "The Stippling Shows Statistically Significant Grid Points": How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It, Bulletin of the American Meteorological Society, 97(12), 2263-2273.