Review of "The ExtremeX global climate model experiment: investigating thermodynamic and dynamic processes contributing to weather and climate extremes" by Kathrin Wehrli et al., Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2021-58-RC1, 2021

The modeling experiments described here shed light on the various roles of land versus atmosphere in extremes, going a step or two beyond what was done in the 1990s and 2000s in the "Koster style" studies of those days. It is interesting, adds to our scientific knowledge of climate variability, and should be published after revision. I do not wish to remain anonymous - Paul Dirmeyer

General comments:

- Realizing this may be difficult without redesigning and rerunning the simulations, but I long to see a bit more separation in the various drivers, e.g., in the atmospheric component, could the roles of dynamics (circulation) versus physics (radiation, clouds, precipitation) be separated? At the land surface, could drivers acting through the energy balance terms versus water balance be quantified separately? Others have delved more into the process level (e.g., https://doi.org/10.1029/2012GL053703), and having models in hand for sensitivity studies enables many possibilities. Likely for "future work", but I wanted to bring up this question.
- I greatly appreciate the message of the paper regarding the role of compensating errors and tuning. There remains among many in the model development community a strange hope that "fixing" one component of an Earth system model (e.g., upgrading the LSM) will somehow solve other problems. But often it just serves to expose those problems even more as the balance of errors has been disturbed. This paper also shines a bit more light on this issue.
- Mainly in §5.1 but also conclusions: The conventional wisdom is that persistent anomalies in the atmospheric general circulation (which may have various causes themselves) establish conditions for heat waves and/or droughts, and then land-atmosphere feedbacks can exacerbate or prolong them. Is there any way to diagnose (confront or confirm) this idea from these experiments? Can the role of climate change
on this evolutionary sequence be investigated here? These analyses are co-temporal and do not seem to account for the evolution over time of heat wave events, although you do consider persistence. It seems the two "approaches" (A) and (B) get at this somewhat (e.g. L343-344) but it is somewhat elusive.

- There are a couple of recent papers that are quite germane to ExtremeX, particularly the notion that heat waves have a mix of land and atmosphere (which may ultimately be traced to remote ocean) drivers: https://doi.org/10.1029/2020AV000283, https://doi.org/10.1002/asl.948.

Specific comments:

L75: Technical point: an ensemble of one is not an ensemble. It is just a single run.

Fig 1: It would be more clear to replot with the X-axis in a time dimension, e.g., label it as the e-folding (relaxation) time scale.

L111: Change "allows to isolate" to "allows isolation of".

L124-125: Which models nudged and which replace soil moisture states? And for those that nudged, what was the relaxation time scale?

L131: I think there was more than one version (combination of inputs) for the LandFlux-Eval data set for ET - which was used?

§2.4: This would benefit from a schematic. Could you reproduce or recreate a figure based on Fig 1 of Wehrli et al. 2019? It would be very helpful. And doesn't differences in the results from approaches (A) and (B) shed light on the nonlinearities in the responses (evidence of feedbacks)?

L285-288: To this list should be added "unrepresented processes" in models, particularly those unresolved due to grid scale: non-hydrostatic atmospheric processes in coarse resolution models, unresolved mesoscale circulations, sub-grid surface heterogeneity.

L288-289: Atmospheric modelers in particular are fixated on 500 hPa geopotential height
errors as a metric of circulation fidelity.

Figs 5, 6 and associated text in §5.1: "Midwest" as a region name does not sit well in the global context, as it is a subregion of the U.S. In the other three cases, "Russia", "Europe" and "Australia" do not designate those entire areas, but a portion within each. Thus, "Midwest" should be replaced with "U.S."

For Russia (line 315) and the U.S. (line 336), how do these areas overlap or intersect the AR6 designated areas? Neither Fig A2 nor any of the other map plots in this manuscript show latitudes and longitudes, so it is difficult to compare by eye.

L353: You discuss results from MIROC, but what about the other two models?

Fig 7: There seems to a growing proportion of contribution from soil moisture as the anomaly periods grow longer (which would be reasonable, as locally soil moisture represents a slower manifold, a redder spectrum than tropospheric variables). It appears this could be easily quantified. Showing the area-weighted average of the metric in the figure (e.g., the SM-dominant percentage, averaged over unmasked areas only) in each panel would show a growing value with warm spell duration in each model, showing the growing relative importance of the land surface states for long-duration events (which would get at the "conventional wisdom" point above, to some degree).

L419-420: Is this true? The atmospheric nudging is very weak in the lower troposphere, and other studies have shown the effect of land surface anomalies on the atmosphere is largely constrained to the boundary layer (e.g., https://doi.org/10.1175/1525-7541(2001)002<0329:AEOTSO%3E2.0.CO;2) except over elevated terrain where heating anomalies from the land surface can get into the upper troposphere directly (https://doi.org/10.5194/gmd-14-4465-2021).

All map figures: Since soil moisture as a climate driver has no meaning over (under) permanent ice, glacial areas like Greenland should be masked from the maps.

Code and data availability: This is not consistent with COPDESS / FAIR data standards to which EGU journals adhere. Public data and/or code repositories should be used and indicated with permanent hyperlinks.