Review of WCD paper “The contribution of circulation changes to summer temperature…” by. Pfleiderer et al.

Overall, the paper is well structured and well documents a thorough study on different methods of estimating the role of atmospheric circulation changes to trends in the northern midlatitudes.

My overall recommendation would be publication after some revisions, which are generally minor.

Thanks for the positive feedback and for pointing out parts of the manuscript that can be improved.

General remarks:

My main query with this paper is the interpretation of the nudged simulations as a “benchmark”. This is an excellent approach to include but I’m not wholly convinced that this is necessarily a gold standard in term of attributing the changes due to circulation.

We agree that the nudged simulations have their limitations and should not be seen as the “gold standard”. In the revised manuscript we discuss these limitations in more detail and alert the reader about these limitations earlier in the manuscript by referring to the section on limitations.

The nudging approach is elegant, and the demonstration in Figure 2 clearly shows that impact of the thermodynamic forcing on the global scale. However, on smaller scales I am less sure that the nudging strictly represents the contemporaneous circulation driven anomalies. A couple of conceptual examples of the potential issues are as follows:

1. In Figure 3 the nudged anomalies are consistently higher than those predicted by the individual methods over the Eurasian continent in the summer. The nudging constrains all seasons (not just summer) so there are likely to be other factors that are modified by the nudging that contribute to these – particularly soil moisture but also other factors such as vegetation, snow melt etc.. These depend on seasons preceding the summer in question. In general these are small but there is the potential for these to have a local influence over time that systematically enhances the temperature response. I suppose these can be summarised as being model “feedbacks” (from other seasons and any associated integrated response) that are explicitly not captured by any of the statistical estimates but are implicitly included in the nudged “benchmark”.

We thank the reviewer for this interesting thought. We did not discuss this aspect so far and are happily adding it to the revised manuscript (line 364-368).

“Additionally, summer temperatures in the nudged circulation simulations might be affected by nudging in other seasons. For example, circulation changes can influence soil moisture in late spring which would then have an impact on summer temperatures. This information is not used

by statistical decomposition methods. Consequently, we have to admit that the nudged simulations are not a perfect benchmark. Further analysis is required to understand how these limitations affect our estimates of circulation induced trends and whether a better suited benchmark test could be designed."

We think that this influence of nudging in the other seasons on our benchmarking test is limited. By nudging the circulation over a longer time we assure that the conditions at the beginning of summer are very similar in terms of SST patterns, soil-moisture and other important pre-conditioning drivers for summer heat. The differences in the starting conditions in early summer between the freely running forced simulation and the nudged piControl simulation are mostly of thermodynamic nature. Concerning soil-moisture for example, with the nudging we assure that the amount of rain bringing storms that pass over a region of interest in spring is the same in both simulations. However, the amount of precipitation from these storms and the amount of evaporation is different in the piControl simulation. These differences in soil-moisture are relatively small. Nevertheless, we agree that the information about these differences are not accessible to our statistical models and could lead to a systematic mismatch between our estimates for circulation induced trends and piControl nudged simulations.

2. The second point is regarding the nudging to winds in the lower troposphere – here the nudging is performed on short timescales and, despite only forcing winds, the dominance of thermal wind balance on synoptic scales means that the nudging will have an effective local temperature forcing. This adjustment will be fast but, for example, any thermodynamic feedbacks that occur between say the land and the atmosphere (for example the strengthening of an anticyclonic high over continents during summer) will show up as being due to the "circulation" when in reality there is a non-negligible impact from thermodynamics. These should not be as well captured by the statistical methods as the information and feedbacks are not directly included but must be elucidated from the data output. Of course, on global scales (i.e. in terms of GMST) this will have no impact but in terms of estimating the "circulation" contribution to local temperature changes, the thermodynamic contribution to thermal wind balance adjustment will be attributed to "circulation".

We agree that nudging the winds in the lower troposphere interferes with small-scale thermodynamic feedbacks and that this has to be considered when interpreting the nudged piControl simulations. Experiments where the wind fields are only nudged from 700 hPa upwards are very similar to the simulations where the whole troposphere is nudged (as used here). We therefore think that these nudging effects do not considerably affect our analysis. In the revised manuscript we briefly discuss the issue.

We would also argue that since the piControl simulations are nudged over long times, this effect is limited. In our simulations, at a given location and time, land-atmosphere interactions should be similar in the freely running forced simulation and it's nudged piControl counterpart, the only difference being that due to thermodynamic changes the interaction might be slightly amplified or dampened. For the intensification of an anticyclonic high due to land-atmosphere feedbacks, only the (potential) intensification of this feedback due to thermodynamic effects would be missing in the statistical estimates.

The raised concern points to an issue that we already discuss in the manuscript which is that the decomposition into "circulation-induced" and "thermodynamic" as we use it here is not very clean and different tested statistical methods treat this decomposition slightly differently. For instance, whether changes in land-atmosphere interaction are part of the "circulation-induced" or the "thermodynamic" contribution differs between the methods. This was a subject of discussion within the author team. Finally, we agreed that despite these differences in the methods, their results are commonly interpreted in similar ways and therefore it makes sense to allow for this inconsistency in the scope of the methods for our comparison.

We extended this part of the discussion adding a schematic figure (figure 5) and a table (table 2) where we summarize the expected implicit treatment of land-atmosphere interactions in the different methods.

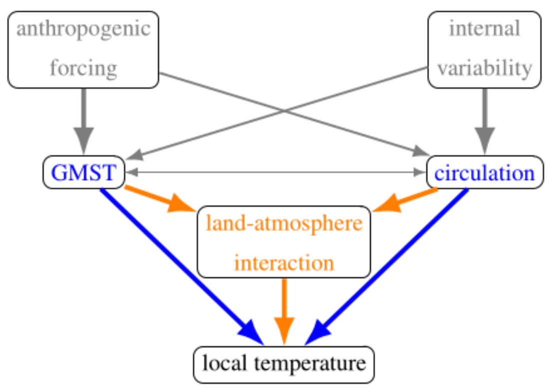

Figure 5. Conceptual illustration of causal relationships influencing local temperatures in a forced climate as in figure 1 but with an additional driver of local temperature. In this study we aim at decomposing the influence on local temperature into thermodynamic (GMST) and circulation induced contributions (in blue). Land-atmosphere interactions is a driver we do not explicitly model (in orange).

Table 2. Expectations on how land-atmosphere interactions might influence the decomposition into "circulation-induced" and "thermodynamic" contributions.

method	what is variability in land-atmosphere interactions attributed to?
ridge	In multiple linear regression, attribution depends on the collinearity of covariates (e.g., circulation) with the second-order effect (land–atmosphere interaction), determining whether it is assigned to GMST or circulation changes. This partitioning may vary by region. If there is no strong collinearity between the prevailing atmospheric circulation at the daily time scale and land-atmosphere interactions, which are typically changing at longer time scales (Merrifield et al., 2017), we expect that effects due to land-atmosphere interactions remain in the residuals.
analogues	It is presumed that circulation analogues occur over a range of land surface states. The circulation-induced component of temperature is defined as an average across this range, which leaves the influence of the land surface on the atmosphere predominantly in the residual thermodynamic component (Merrifield et al., 2017). It is possible for the land surface to induce a temperature anomaly and associated circulation pattern (e.g., a thermal low) and the analogue method could interpret this situation as circulation-induced rather than thermodynamic. Nudging all vertical levels of the atmosphere suppresses influence from the land surface to the atmosphere, so land-atmosphere interactions are likely to remain in the thermodynamic component of the piControl-nudged runs used as a benchmark in this study (Merrifield et al., 2019).
DEA	Because the approach removes the total effect of GMST without conditioning on land–atmosphere interactions, it may also eliminate the mediating effect of GMST operating through this pathway (but this effect is likely small and confined to trends that are colinear with GMST), whereas the mediating effect of atmospheric circulation is expected to be retained, given that the linear model has sufficient expressive capacity to capture these complex relationships.
UNET	The SLP is used as a predictor of the circulation. However, this variable may contain surface imprints which might affect the "circulation-induced" component. Therefore, land-atmosphere interactions may be partly predicted by the UNET architecture.
nudged simulations	Nudging is expected to separate the land-atmosphere interactions into a thermodynamically-driven, and a circulation-driven component (i.e., atmospheric imprints of land-atmosphere interactions are expected to be captured through nudging).

Neither of these examples particularly undermines the nudged simulations but do highlight how they are fundamentally different from the statistical approaches, as they implicitly include more thermodynamic effects adn feedbacks.

This may explain why the distibutions of the trends are systematically underestimated (e.g. the distrubutions on the right of Figure 3 and in Figure D2) in all the statistical approaches as they do not include the feedbacks and adjustments that are implicit in the nudged runs. At present there is only a discussion of the limitations of the statistical methods but I think discussing the limitation fo the nudged approach would also be useful to include.

We added a discussion of the limitations of the nudged circulation experiments and changed the framing accordingly (see previous comments). Concerning the systematic underestimation of circulation induced trends in statistical decomposition methods, we are quite confident that they are mainly due to the underdispersiveness of the models. We cannot exclude that the mentioned shortcomings are relevant.

Line 369-374: "Concerning the land-atmosphere interactions, we conclude that the effect on our estimates of circulation induced trends is diverse between methods. This increases our confidence in the signals all methods agree on (e.g. circulation induced warming over Europe). At the same time, there is no systematic (and consistent) difference in how statistical decomposition methods might be affected by land-atmosphere interactions in comparison to how land-atmosphere interactions might affect the nudged simulations. Therefore, the effect of land-atmosphere interactions cannot explain the systematic underestimation of the magnitude of circulation induced trends in statistical decomposition methods (as compared to the nudged simulations)."

I am sure the authors can directly discuss and address these differences and I think this would strengthen the interpretation of the results.

Minor comments:

Figure 3: This is a bit messy in the version I have– more details on the KDE plots on the right would be usefukle (along with axis labels etc).

We agree that axis labels in the KDE plots are required. We did not find a solution to include in a size that is still readable and removed them from the plot. KDE plots are still shown in the appendix.

Section 2.3.4: I don't quite follow why the SLP would not be detrended as the "forced response" is small. It this important? Surely it would be better to include, unless the results are sensitive to this? I also may have misunderstood this, in which case a brief clarification might help.

The objective is to estimate the part of daily temperature variations which can be explained by the large-scale circulation (using the SLP as a proxy of this circulation).

As we are working with historical+SSP runs, we need to account for the forced response in the temperature which is not insignificant. Thus, for the training, we remove from the temperatures the mean seasonal cycle and the estimation of this forced response (with the method described by Rigal et al, 2019) as they are not circulation induced. We did not pre-process the SLP data by removing an estimate of the forced response because it is small in the SLP (Figures 2 and 3 show that the three piControl-nudged experiments do not exhibit significant common trends).

I want to end on a positive: I think this is a very interesting paper!