

# Reviewer comment 1

## Overall review

This paper by Pfleiderer et al. aims to improve our ability to decompose climate trends into thermodynamic and dynamical components, with a focus on surface temperature trends in the Northern Hemisphere. The first one is to determine whether statistical methods are able to quantify dynamically induced trends in climate model data by comparing their outcomes to a set of nudged climate models experiment, considered to be the ground truth. Once this is validated, the statistical methods and another set of nudged experiments are applied to ERA5 data to actually determine the contribution of dynamical changes to the surface temperature trends in the northern mid-latitudes.

The paper is highly relevant and timely, and it provides an important assessment of dynamical adjustment techniques. Beyond its specific results, the framework developed could be applied to a wider range of climate variables, such as precipitation or extreme events.

The use of nudged simulations with no external forcing is a particularly smart approach to isolate the dynamical influence on surface temperature trends. Since such experiments are difficult to construct for observational datasets (though the AMIP + nudging above 700 hPa approach seems promising), validating statistical methods is crucial, and this paper does so effectively.

The manuscript is generally well written, although it can be hard to follow at times. Some sections, particularly on the analogues method, would benefit from clearer explanations. Also, the two main objectives, though related, are presented somewhat independently and could be more tightly connected in the structure of the paper. For example the authors could emphasize that the first objective is used to strengthen our confidence in the second objective.

Despite some concerns I have about the paper (detailes below), I think this paper is almost suitable for publication in WCD, but requires some work, notably to improve clarity. For these reasons I suggest to **accept this paper with minor revisions**.

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

## Main comments

### 1. Comparability between methods

One of my main concerns is the comparability between the different statistical methods. Indeed, each method uses a different set of predictor variables:

- Ridge regression uses the streamfunction
- Circulation analogues use sea-level pressure

- Direct effect analysis and U-Net use z500

This makes it difficult to assess whether differences in performance are due to the method itself or the choice of input variables. It would be helpful for the authors to comment on this explicitly. If the best predictor was chosen for each method, this should be clarified.

*Our aim with this article is to evaluate how reliable statistical and machine learning methods for trend decomposition are. We did not develop these methods for the task of disentangling circulation induced changes from thermodynamic changes but rather used already existing methods that are likely to be used for the task. Therefore, we prefer applying the methods as they were used beforehand in the published scientific literature, which entails the usage of different proxies for atmospheric circulation.*

*For most of the methods, sensitivity tests with other input variables representing atmospheric circulation showed that the choice of the input variable does not impact the result considerably.*

*In the revised manuscript we compare results with the ridge regression with streamfunction at 500 hPa as input variable to results from the same ridge regression but with geopotential height at 500 hPa (corrected by subtracting the global mean of geopotential height) as input (see figure G1). The results are very similar:*

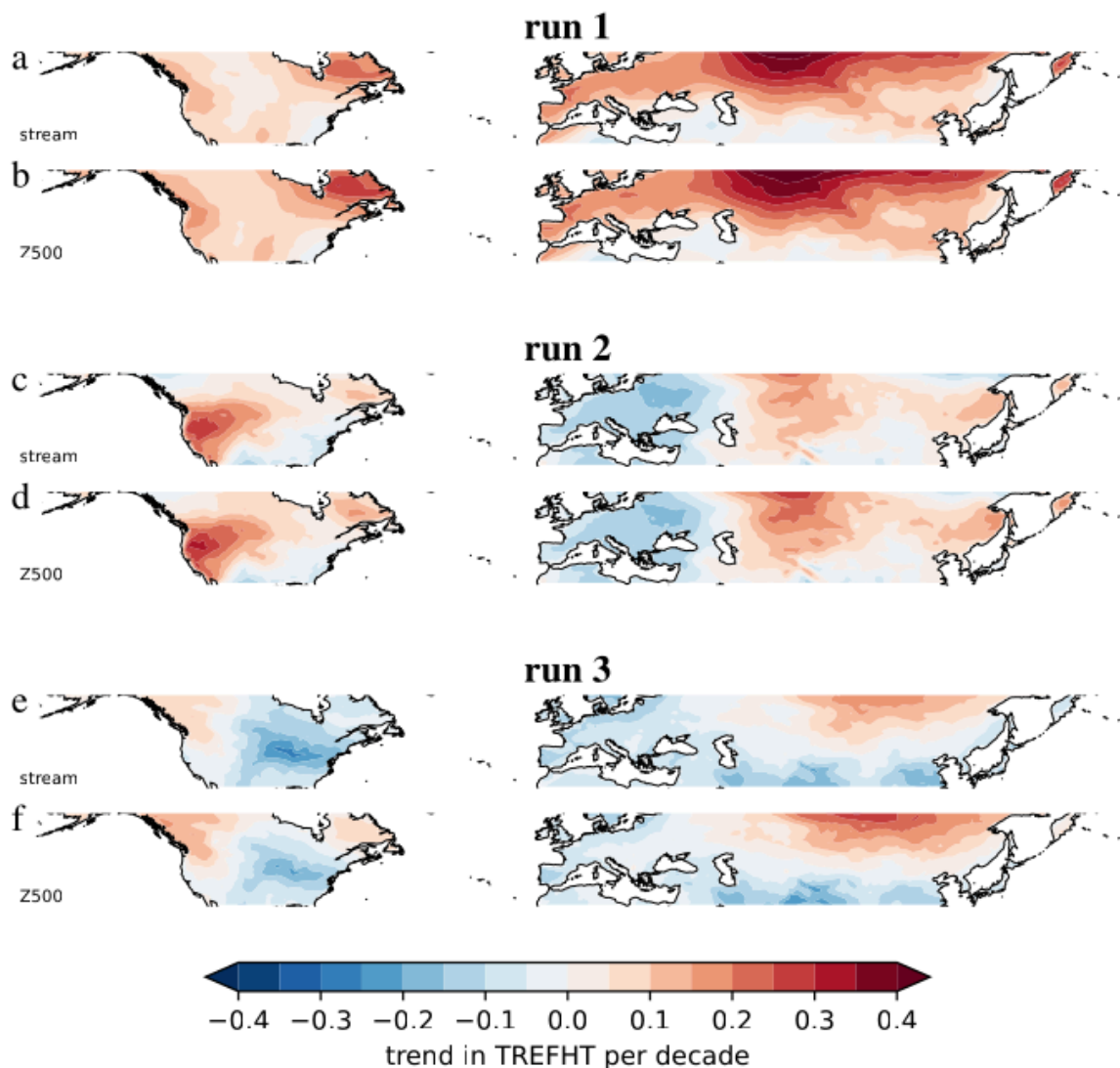


Figure G1. Estimates of the circulation induced trends from the ridge regression over the period 1979-2023 in the freely running forced CESM2 simulations. Estimates from the ridge regression using streamfunction at 500 hPa as a covariate for circulation (a,c,e). The same, but with geopotential height at 500 hPa as anomalies to the global mean geopotential height at 500 hPa (b, d, f).

*In the revised manuscript, we inform the reader about this aspect in line 139-142:*

*“Note that we apply the methods exactly as they were designed and used in other publications and therefore different proxies for atmospheric circulation are used by different methods. We do not expect that the choice of the variable to represent atmospheric circulation affects the results considerably. In figure G1 we show a sensitivity analysis for the ridge regression. In section 3.3 we discuss differences between the methods and how they might affect the decomposition in more detail.”*

## 2. Lack of information on trend estimation, significance and uncertainty

Maybe I have missed it but I couldn't find a mention on how the trends were computed. In addition, such a study would benefit from statistical tests on trend significance and uncertainty, especially for the second objective which aims to provide robust estimates. As all methods provide an estimate of surface temperature directly, trends statistics could be computed for all cases. Moreover, it might make sense to evaluate skill metrics only for statistically significant trends.

*We agree that a discussion on the significance of analyzed trends is lacking. Circulation induced trends are weak compared to thermodynamic trends. To which extent anthropogenically forced changes in atmospheric circulation patterns is subject of debate. It is however clear that a large part of circulation induced trends over a time period of 45 years is a result of internal climate variability. The differences between the nudged piControl simulations (figure 3 in the original manuscript) suggest that in CESM2 most of the circulation induced trends at a local scale mostly reflect internal climate variability. Whether this would be the same in other climate models or in observations is a question we do not address here. Either way, we assume that a large part of the circulation induced trends is driven by internal climate variability and therefore we expect that most individual circulation induced trends are not statistically significant (yet for the thermodynamical trends we expect much less variability and therefore a high proportion of significant trends).*

*The consistent regional patterns we find in the circulation induced trend maps show that the trends are the result of processes in the climate system and we want to quantify these contributions even though from a statistical point of view some individual trends are not statistically significant.*

*In the revised manuscript, we show maps with significance stippling in the appendix (figure B1):*

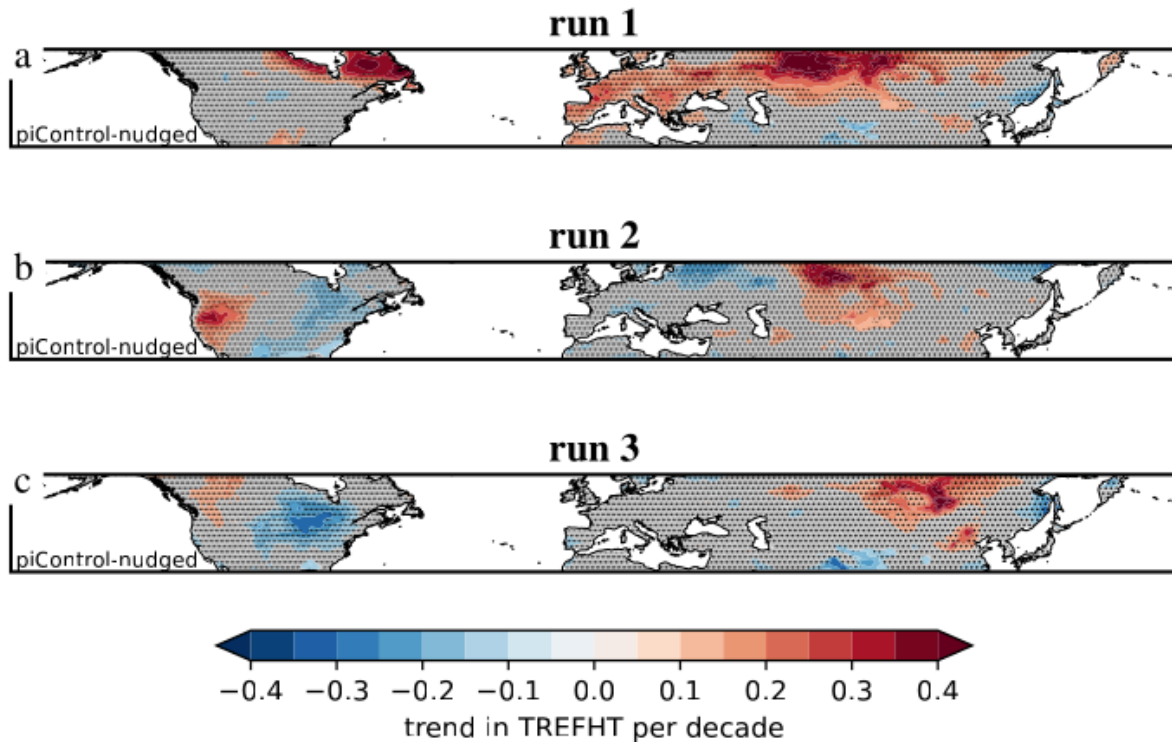


Figure B1. JJA trends in piControl-nudged simulations for the period 1979-2023. The Stippling indicates that we cannot reject the Null-hypothesis of no trend at a 95% level.

We also add a paragraph on the significance of the trends in the results section (line 267-270):

“Note that most of these trends in the atmospheric circulation-induced component are not statistically significant (see figure B1). Since these trends mostly reflect internal climate variability, it is expected that from a statistical point of view, the circulation-induced temperature changes at one location are not differentiable from noise. The spatially consistent trend patterns show that, despite lacking statistical significance, these trends contain helpful information and are worth evaluating.”

### 3. Section 2.3.2 (circulation analogues) lacks clarity

The description of the analogue method is quite confusing. As someone who is not familiar with circulation analogues, I cannot say I have understood what it is from that section. Please revise it to make it clearer. Here are the points that made it unclear to me:

- “Analogues” are not clearly defined when first mentioned (line 155).

We’ve revised to define analogues at first mention. Thank you for the clarification!

- It is unclear what the 80 possible choices for analogues refer to - days, years, months?

We've moved towards an example to emphasize the application of the method on monthly mean fields.

- What are these 50 out of 80 choices? I did not understand this paragraph

This detail is to orient readers familiar with applications of the method in previous papers. The step refers to the strategy of going from the whole record as possible analogues to a subset of the record as possible analogues. We have hopefully clarified that by re-ordering the paragraph and adding detail.

- Line 128: "Once the Euclidian distances are determined" at that point there is no indication that a Euclidian distance is computed, or why and on what it is computed.

Thank you for this point, we had gotten ahead of ourselves a little. We have moved the mention of Euclidean distances from the later paragraphs into the first paragraph so (hopefully) it is now clear what is being done.

- Line 169: analogues are now defined but this should be done earlier

Done, thank you!

Maybe this is also the case for the UNET paragraph, but as I am more familiar with UNETs it was easier to follow.

#### **4. Are the UNET Predictions Truly Circulation-Induced Temperature Changes?**

- Line 213 describes the UNET model as predicting a temperature field with an estimate of the daily non-stationary normal removed. However, this doesn't necessarily isolate the circulation-induced component. The paper assumes that the resulting anomaly is circulation-induced, but this should be justified more clearly.
- If the previous point is justified (which I am sure it is) why not use the method from Rigal et al. (2019) directly to estimate circulation-induced temperature changes?
- Is the UNET performing well to reproduce this anomaly field?
- Why not use the UNET to predict the nudged experiment directly, which serves as the ground truth for the comparison later

Also, it is unclear what "CESM2 transient simulations" refers to. Do these include historical + SSP runs?

The aim of the UNET approach is precisely to estimate the part of daily temperature variations which can be explained by the large-scale circulation (here assessed from daily SLP fields). The mean seasonal cycle of the temperatures is not circulation-induced, so it is relevant to remove it and focus on temperature anomalies ( $T'$ ): this is why we write the UNET model as  $T' = f(\text{SLP})$ .

The UNET is then trained to learn the link between SLP and T'. As we train the UNET on historical + SSP runs (which we call 'transient', i.e. non-stationary), we need to account for climate change in the  $T' = f(\text{SLP})$  relationship. Here we detrend the temperatures but not the SLP, assuming that, in this model (CESM), the forced response in the SLP is small compared to the daily variability --- which seems to be a reasonable assumption as the 3 piControl-nudged experiments do not exhibit significant common trends (see Fig 2 and 3). The detrending is made following the method described by Rigal et al. --- estimation of daily non-stationary normals --- which is convenient as it allows us to remove both the mean seasonal cycle (first point) and the climate change signal at the same time.

## Minor comments

- Table 1: I fail to understand how R2 values can be positive. Could you please explain?

*We use the coefficient of determination “R2” for the evaluation of our results and in the revised method we clearly define it. It informs about how much of the variability in our benchmark for circulation induced trends is explained by the estimates from tested methods. It usually ranges from 0 to 1. Cases where it is negative indicate that just taking the mean of the data would perform better than using the tested model. This is the case in figure 3h for example. We discuss the interpretation of R2 in the paper (line 256-258):*

*“(iii) The coefficient of determination ( $R^2$ ) is a widely used metric for spatial comparisons, as it accounts for the variance at each location and indicates how much of the observed variability is explained by the prediction. Yet, it is -in contrast to Pearson correlation- sensitive to any bias in the estimated average (Kvålseth, 1985); and hence it is possible that a statistical method shows a good spatial Pearson correlation in its estimates but a poor  $R^2$  score.”*

- Figure 4: How are thermodynamic trends obtained? Are they estimated directly (e.g., from the ridge regression method) or as a residual (total trend minus dynamical trend)?

*It depends on the method: In the ridge regression and DEA it can be directly estimated from the model. With the analogues and UNET it is the residual.*

- Line 185: To be consistent with the text, maybe consider using  $Y_{\text{orth}}$  instead of  $Y_{\text{perp}}$

Done

- Line 270: “to weak trends” should be corrected to “too weak trends”.

Done

## Reviewer comment 2

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”.

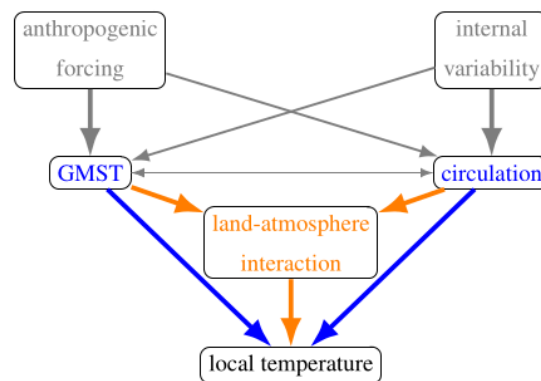
*We agree that nudging the winds in the lower troposphere interferes with small-scale thermodynamic feedbacks and that this has to be considered when interpreting the nudged piControl simulations. Experiments where the wind fields are only nudged from 700 hPa upwards are very similar to the simulations 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 its 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 and feedbacks.

This may explain why the distributions of the trends are systematically underestimated (e.g. the distributions 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 of 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 useful (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. Is 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!