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 comparibility 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:

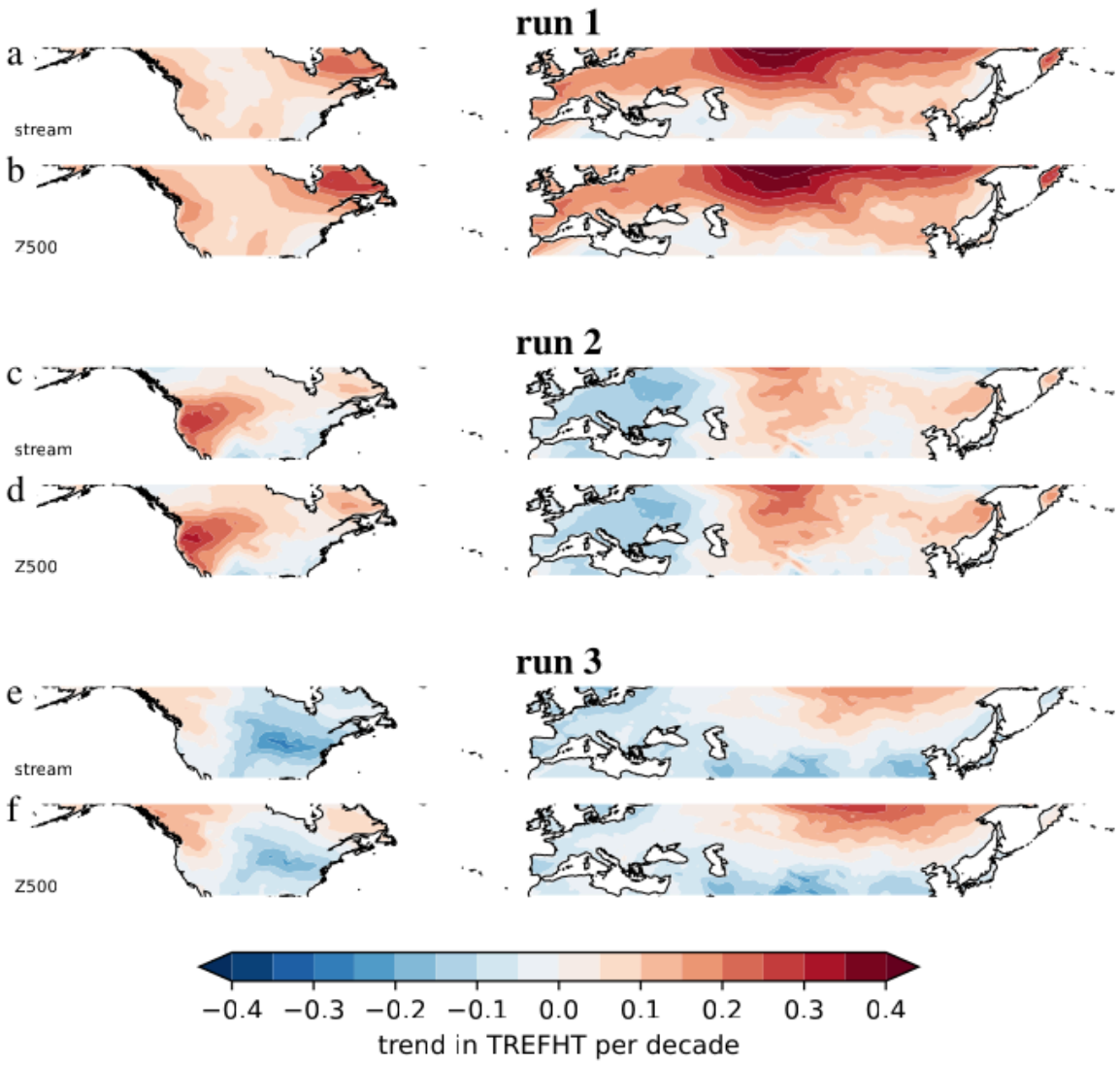

Figure G1. Estimates of the circulation induced trends from the ridge regression over the period 1979-2023 in the freely running forced CESM2 simulations. Estimates from the ridge regression using streamfunction at 500 hPa as a covariate for circulation (a,c,e). The same, but with geopotential height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

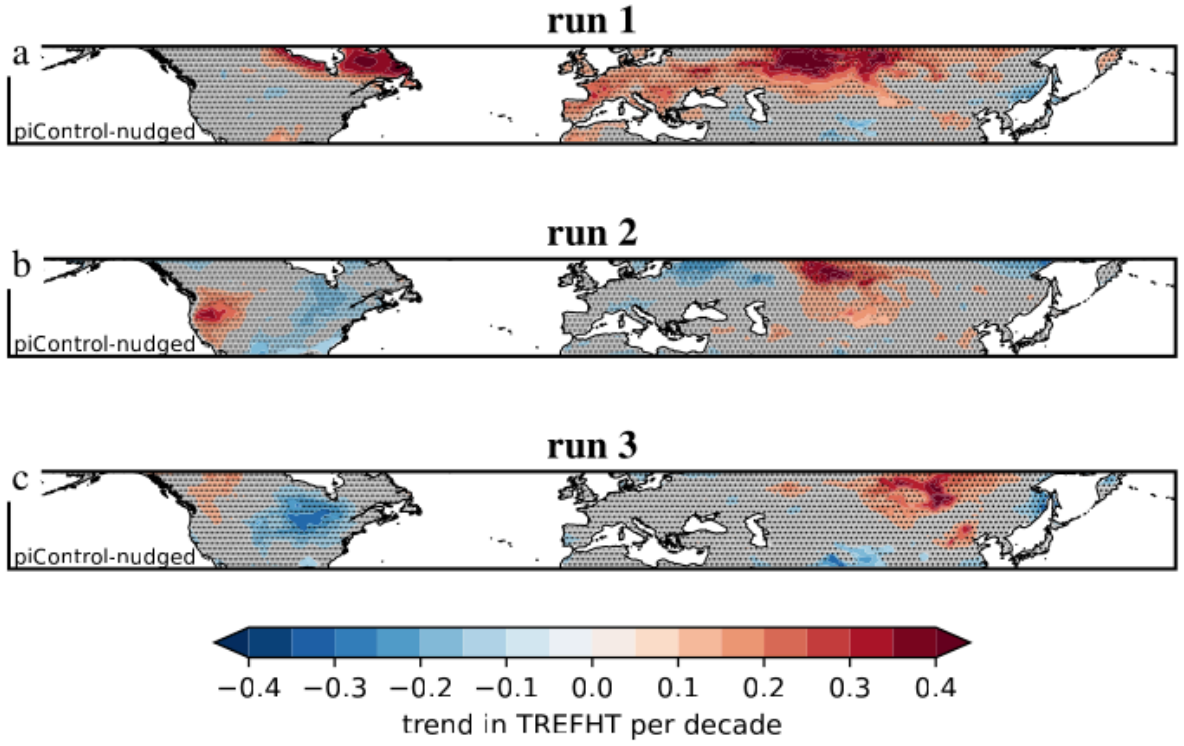

Figure B1. JJA trends in piControl-nudged simulations for the period 1979-2023. The Stippling indicates that we cannot reject the Null-hypothesis of no trend at a 95% level.

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(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(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 (R2) 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 R2 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_orth instead of Y_perp

  Done

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

  Done