Response to editor comments

The first referee is satisified with the revisions. The second referee was not available to check the revisions, but I have assessed whether their concerns were addressed and believe they are. Overall, the authors have done a nice job of revising the manuscript to improve the clarity and interpretability of the results.

Below I've listed some issues requiring clarification and/or correction that seem to have slipped through, some of which are related to points from the ECR peer-review training group. In addition, I strongly recommend a dedicated proofreading pass to catch the small but numerous writing glitches throughout the manuscript. The goals/results of this study should be very interesting and useful for the community, hence I feel these final efforts are worthwhile. The manuscript should be ready for publication after these points have been addressed.

Thanks a lot for providing another review!

- The peer-review training group had a comment about the mismatch between the "main objective of the study ... to investigate long-term temperature trends" and the "the prominent motivation based on heatwaves and extremes" (i.e., the first two paragraphs of the intro). The revised manuscript now includes some lines in the second paragraph to bring the focus back to trends, which goes some way to addressing this problem. I think it could be fixed completely by doing a bit more reorganizing/refocusing of the first two paragraphs - perhaps lead with the trends and bring in the extreme heatwaves as an important related impact. (The peer-review group had a different suggestion, which would also work but would take things in a different direction: "It would be helpful if the authors could discuss a more concrete example of how their estimated trends allow a better understanding of extreme events").

We agree, that heavily focusing on heat extremes in the beginning of the introduction may be a bit misleading. We slightly changed the first paragraph and have included more text about warm seasons and heat waves. Overall, we believe that the focus on warm seasons (including heat waves) is still a good motivation for the trend analysis, because the heat waves and warm seasons are indeed an important impact of the trends (and were studied in that way also for instance in Teng et al., 2022).

- L101: The use of the word "experiments" here confused me. I believe you've run 3 simulations (or members) of the historical experiment and 3 simulations of the SSP370 experiment, correct? Sometimes, you refer to the historical and SSP370 simulations as different runs, and sometimes you refer to the combined hist+SSP370 simulations as one transient run. Please pick one and stick with this. Table 1 still refers to 1300, 1400, etc.

We replaced “experiments” by “simulations” in most instances to avoid confusion. We also adapted table 1. Thanks for noting the inconsistency!

We also introduce the scenario hist+SSP370 properly when it is first mentioned (line 73-75).

- Section 2.2: The description of these experiments is confusing. I think you should state upfront that there is a piControl and a hist+SSP370, and clarify what you mean by "the same" anthropogenic forcing (L179). Are they all run in AMIP configuration with observed SSTs?

We clarified the description of the experiments and the used forcing. We also clarified the AMIP setup and highlighted the differences to the other nudged experiments.

- Section 2.3: The peer-review training group had some very nice suggestions here, and I feel they should be straightforward to implement (the revised manuscript already reflects some of these suggestions). Readers who do not want to go into the details should be able to read the start of each subsection and still appreciate the study's results. In addition, a few comments from me on the revised version:
* 2.3.1 the info in L209-213 is useful for introducing the method, \lambda is called both the ridge regression parameter and the regularization parameter).

Thanks for the suggestion. Changed.

* 2.3.2 has improved in response to referee 1's comments, although I don't understand L251-252, and the last paragraph remains quite confusing.

The paragraph has been re-written as:

In this study, analogues are selected from the free-running hist+SSP370 simulations to dynamically adjust (1) each hist+SSP370 simulation and (2) ERA5. Each hist+SSP370 simulation is dynamically adjusted using the ``leave-one-out" approach (Deser et al., 2016; Lehner et al., 2017). In the leave-one-out approach, for each month, e.g., June 1900, analogues are selected from all other Junes in the simulation's 1850-2014 period except 1900. The leave-one-out approach is used for the comparison between the hist+SSP370 and piControl-nudged simulations. In the second approach, analogues are selected from the entire 1850-2014 period of each of the hist+SSP370 simulations and used to dynamically adjust ERA5. The resulting three dynamical components of ERA5 are shown in Figure F1 and are averaged to produce the circulation-induced trend estimates in Figure 4.

* 2.3.3 similar to 2.3.1, could use a general description before launching into details.

We added short introductory descriptions for all methods.

* 2.3.4 What are the "8" CESM2 transient simulations? We've only heard of 3 x 2 experimemts = 6 until now?

In order to avoid over-fitting, the UNET is trained on 8 other transient simulations (historical + ssp) from the CESM2 ensemble, i.e. different from the 3 simulations that have been used to build the nudging experiment (members 1300, 1400 and 1500). This has been clarified in the revised version.

- L343-354: I appreciate the additional details here, but it would be clearer if you made these an enumerated list so the reader gets the explanation of each metric right away.

Good idea. We changed it to an enumerated list.

For (i), what is meant by "the fraction"? Is this just spatial, i.e., out of total gridpoints? This should also be clarified in the Table 1 caption. * ah yes, explained in L372 *

We clarified it in the description at the beginning of the results section and it the caption of table 1.

For (ii), is it an area-weighted spatial correlation?

No, nothing is area weighted here. We added a comment on this in line 273.

Other (language, technical, etc.)

- Check peer-review group's various suggestions.

Done

- L16: Suggest "sign" rather than "direction".

Done

- L94-97: I like the addition of some preamble to the Data & Methods section, but this makes it sound like the decomposition methods are only applied to ERA5!

We added a sentence clarifying that the decomposition methods are evaluated on CESM2 simulations.

- L107: Is it just the forcing that follows the CESM2 LENS2 protocol, or also the method of generating initial conditions (for example, do the 3 members include different ocean states, as in LENS2, or just atmospheric perturbations)?

The three initial conditions used for the freely running hist+SSP370 (and their corresponding nudged simulations) are from a piControl run and the year of initialization for each run is separated by 100 years: namely year 1300, year 1400 and year 1500. Therefore, they have different ocean states. We added a comment concerning the initial conditions in the manuscript (line 76-78).

- Figure 2: Can you please explain the various lines in panel a in the caption? The figure label says 50-year smoothed GMST, which I assume are the bold lines. The thin lines are annual means?

Changed the figure label and caption.

- L155: "both of these processes" suggests only 2 processes at play - suggestion "both thermodynamic and dynamic contributions"

Changed

- L172: It's a bit odd to mention ridge regression and DEA results here, before the methods have been described. I see why this point was added here, but it can be done more generally (referencing some of the decomposition methods, for example)

Changed

- L405-6: Last sentence, second part is quite vague and makes the entire sentence a bit confusing.

Changed

- Some broken labels throughout (figures, references)

- I would tend to favour "smaller/weaker vs larger/stronger" trends, as opposed to "lower vs higher" trends.

Agreed and changed

Response to:

Review by Arundhati Kalyan, Anjali Thomas, and Jan Zibell*

*J.Z. declares being employed at the same institute as one of the co-authors of the study.

We copied the review and added our responses in **green**.

In the above study, the authors assess long-term (multidecadal) circulation-induced changes in summer temperature trends in the northern hemisphere midlatitudes (30–60°N). They employ four different statistical/machine learning-based methods – ridge regression, atmospheric circulation analogues, direct effect analysis, and a convolutional neural network – to isolate temperature trends driven by circulation changes over historical and future time periods in free-running climate model simulations (using CESM2) and ERA5 reanalysis. The relative performance of these methods is evaluated against a benchmark comprising nudged climate model experiments (also using CESM2) which include forced and internal components of circulation variability, but no forced thermodynamic component. The four methods are found to be effective on climate timescales, albeit with biases. The paper also highlights regional differences in dynamically forced trends, revealing alternating wavelike patterns of warming and cooling throughout North America and Eurasia. Finally, the authors discuss in depth the challenges and limitations in using statistical methods to decompose circulation vs. thermodynamically driven trend signals.

The study makes robust choices relating to data and methodology, in that four different decomposition methods are used and multiple statistical evaluation metrics are analysed for each. The authors transparently present a summary of the climate of the wind-nudged simulations which form the reference for their assessment of the decomposition methods. The presentation of challenges (multidecadal timescales, differing climate components included in the circulation-related decomposition, creation of a suitable benchmark, need for multiple models) in the discussion section is strong and may serve as useful reference for future studies. The study completes the stated target of quantifying circulation-driven trends across the NH midlatitudes and validating the different methods against a suitable climate model benchmark. In addition to highlighting features with low and high skill for each method, the study adds some degree of confidence to the findings of earlier studies that examined regional patterns of circulation-driven temperature trends. The article is concise and generally well-written. The title is informative and appropriate to the study, although the abstract could be improved as suggested below. The relevance of the study, objectives, and scientific challenges are introduced well. Even so, there are portions of the manuscript that are hard to follow and hinder conceptual understanding and interpretation of the results. In particular, the explanations of the statistical methods, the illustration of statistical significance/uncertainties associated with the results, and the dynamical interpretation of the study could be improved. We recommend publication of this manuscript in Weather and Climate Dynamics after minor but necessary revisions as outlined in the comments below.

We thank the reviewers for taking the time to thoroughly read and comment our manuscript. The raised concerns and suggestions are very valuable and helpful for the finalization of the paper.

We received these reviews after completing a revised manuscript addressing comments from two other reviewers. Therefore, some of the comments are already discussed in the response to reviewer 1 and reviewer 2. Furthermore, some suggestions have already been implemented in the process of preparing the revised manuscript. We mainly answered comments where we aim to clarify aspects. We also briefly commented on suggestions that we implemented in the final version. In few cases we do not reply, because in these cases the aspects have been largely addressed already at the revision stage. Please reach out to us if you are interested in more detailed answers to specific comments.

Statistical methods: The description of the four methods used is not straightforward and quite hard to follow for a non-expert reader. We are not experts in the statistical/ML methods used in this study and hence, cannot comment on the strengths and weaknesses of the design and implementation of these methods. However, here are a few suggestions to make the methods understandable to a broader dynamics audience, such that we can better appreciate the significance of the study's findings:

- Add an introductory sentence or two in plain language in all of the sections 2.3.1–2.3.4 to explain what the method and/or equation aims to achieve

  Section 2.3.2: Great recommendation! We’ve added the sentence: “It achieves this through the re-construction of monthly mean climate fields using linear regression with coefficients derived from a field representative of atmospheric circulation (here, sea-level pressure).”

- Clearly define all variables and constants introduced in each equation

  We believe that all variables and constants are now introduced.Consider also whether certain details in the method descriptions can be moved to supplementary text

- Here a few line-specific comments:
  - Line 134: Please explain why a 40°×40° rectangle around each grid cell was chosen. Could other sizes change the results?

    This is a good question and usually a sensitivity analysis would be useful. In the case of the ridge regression, the coefficients of grid-cells far away from the location of interest are kept small by the regularization and therefore, the results are not expected to be sensitive to the extent of the region.
  - Line 156: Please discuss briefly how you choose the number of analogues (Ns = 50) and repetitions (Nr = 100). Does changing these choices affect the results?

    Great question, we have added that these numbers were selected to be consistent to how the method was in previous studies. If you are interested, a more comprehensive sensitivity analysis for another domain is available here: https://escholarship.org/uc/item/5mz52654
  - Line 200: It might help to briefly explain the physical meaning of the Yperp component in plain language in this context.

It would be good to see more rigorous discussion of the statistical significance and uncertainties associated with the results presented. The need for this arises due to the following reasons:

The question about the significance of the trends has also been raised by reviewer 1. Please see our response to reviewer 1, which we hope addresses your concerns raised above and regarding the specifics below.

- "… observed trends are falling out of the range of model-simulated expected trends" (Line 35)
- There is "ambiguity" on the magnitude of the historical circulation-induced trends, and all methods likely underestimate these magnitudes (Sec. 3.2)
- In light of the above, it would be helpful to better understand the likely range/distribution of trends for each method and how they differ. Are some methods more likely to include the observed/nudged values than others?
- The use of a very small sample size (only three ensemble members) for the free-running and nudged simulations limits the ability to show robust

uncertainties. Could you provide more information on whether the initial conditions for the three nudged ensemble members were chosen at random or based on certain criteria? Could you comment on how large of an added value more members would be and why you decided on three only?

We use of the nudged simulations as test cases that are supposed to mimic the intended application of the decomposition methods. The decomposition methods are usually applied to reanalysis data. With our test cases, we stick to the application to one climate trajectory as it would be done in reanalysis. We will not be able to give meaningful confidence intervals for the estimated trends in ERA5 but we want to know how meaningful the estimated trend pattern is (and magnitude of the pattern). For this purpose, three members are sufficient. The initial conditions were chosen to be mostly independent in terms of ocean states. Lastly, we would like to comment on the ensemble size in general: We agree with the reviewers that in principle, more members would be better (of course). However, with our specific setup, we believe that three members are indeed enough to sample the thermodynamical component well, because the thermodynamical component for each of the three simulations falls very close to the ensemble mean. More importantly, the dynamical component can be used as a test case for the nudging, and for this application we strongly believe that the three members are sufficient. Yet, of course we agree that more members would be useful for instance if the goal would be to identify potential forced dynamical responses.

- Table 1 would also benefit from the inclusion of confidence intervals and/or p-values to reject the null hypothesis that the correct sign or correlation of temperature trends was obtained by random chance.

  This point was also raised by reviewer 1. Please find a response to reviewer 1 comment 2.

There are instances where we think that the discussions in this study may benefit from a more dynamical perspective:

- It is clear from the introduction and methods that the wind-nudged simulations are viewed as a ground-truth benchmark. The limitations of this approach are discussed as, for instance, in line 335: “However, there may be factors of residual climate variability (such as ocean variability) or feedbacks between circulation and other factors such as land-atmosphere coupling that could still affect thermodynamical processes on climate over land." Aren't there even more concrete examples for a physical relationship that is not a priori captured, such as the time mean thermal wind balance?

This was also the main concern of reviewer 2. Please check the response to reviewer 2.

- Line 36–37: Please consider to slightly reframe "... may indicate ... that a forced change in circulation is missing in the models". This framing could make one think of unresolved/parameterized processes in climate models, such as latent heating in deep convection in the midlatitudes. For those it is the consensus that a forced circulation change *is* missing in the models. Should you actually refer to this, the sentence could be reframed as "that circulation changes due to unresolved processes, which are not captured by the models, turn out to be meaningful / non-negligible".

  Since in this study we do not study the reasons for misrepresented processes in climate models, we actually prefer to keep the statement vague and general.

- It is certainly not the purpose of this study to discuss all limitations of wind-nudging in detail and of course every alternative also has its limitations, but we suggest that the authors at least address the point that in the real atmosphere, the development of a heatwave is not the linear sum of a thermodynamic component plus a dynamic component. There are many non-linear dynamical feedbacks from thermodynamical processes, e.g. from coupling with radiation via clouds, surface fluxes depending on soil moisture, or latent heat release (which, if speaking of heat waves, within a warm conveyor belt may increase blocking intensity (Pfahl et al. 2015)). Possibly, addressing this is related to emphasizing more prominently that the authors regard GMST as the representative variable of thermodynamic changes.

  The reviewers raise an important point that was also raised by reviewer 2. In the revised manuscript we discuss the implications of the highly simplified decomposition into "circulation-induced" and "thermodynamic" in more detail.

- Overall, we suggest that 1) the assumption that a trend can be decomposed into thermodynamical and dynamical components and 2) the use of wind-nudged simulations (as introduced in Sect. 2.1) to achieve this are discussed a bit more critically. This could be done very briefly when introducing the wind-nudging experiments in the Introduction and then in a more elaborate way for instance as a sixth discussion point in Sect. 3.3.

Thanks for this very good suggestion! This is what we ended up doing in the revised manuscript.

The main objective of the study is to investigate long-term temperature trends. Therefore, the prominent motivation based on heatwaves and extremes seems somewhat out of place. Motivating this research with individual events is fine per se, but this study does not investigate trends in heatwaves. It would be helpful if the authors could discuss a more concrete example of how their estimated trends allow a better understanding of extreme events (as indicated in lines 31-34)?

Thanks for this comment. We updated the beginning of the introduction by shifting the focus more towards warm seasons and their impacts.

The use of the dynamical vs. thermodynamical separation of trends could be even further strengthened by a discussion of the geographical variability of the thermodynamically induced trends. In Figure 4, over Eurasia your methods disagree whether the thermodynamical change is rather uniform or dependent on longitude or latitude. Are there any physical arguments in the literature for what the thermodynamically-induced pattern of warming should look like? For instance, it is observed that the Mediterranean region is warming faster (Brogli et al., 2019). Alternatively, it seems also fine to note that this is left for future study or refer to discussions in other studies.

This is a very interesting observation and we will add a comment about this difference in the estimates for the thermodynamic contribution in the discussion where we discuss the differences in how decomposition methods separate between “dynamical” and “thermodynamical” (line 367-368).

Minor comments:

•Line 3–4: “Over the northern hemispheric mid-latitudes, considerable regional differences in summer temperatures have been observed.” → Presumably, you mean differences in summer temperature *changes.*

Has changed.

•Line 5–6: We think the general readability of the abstract would benefit from a brief description of ‘decomposition method’, i.e., what you decompose the trends into. If one is not very familiar with the topic or similar literature, this is not obvious but only (and well) presented in the introduction.

Good idea, done.

•Line 10–11: “Most decomposition methods show skill in estimating the sign of circulation-induced trends but all methods underestimate the magnitude of these trends.” This statement contains the fact that you use the wind-nudged simulations as your benchmark and that you assume that the nudged simulations contain 100% of the dynamical component of the trend. This should be presented more clearly as was done in the Introduction, for instance, around line 60.

As discussed later in the paper, we are convinced that this underestimation of the magnitude of trend patterns is not affected by limitations in the nudging experiments as benchmark.

•Line 16: Consider changing: The intensity of heatwaves “increases globally” to “has been increasing globally”.

•Line 18–19: Consider adding a reference/s here to show that intensification of heat waves occurs in a warmer climate due to thermodynamic factors (perhaps an attribution study?).

•Line 19: “However, heat waves are not only...” is a long sentence and could benefit from restructuring.

Changed

•Line 22: It might be worth mentioning here whether land-atmosphere interactions are more or less important than circulation changes as a factor in driving summer temperature trends.

•Line 23: It might be worth commenting on whether regional trends are more pronounced in the NH midlatitudes than elsewhere. May also be good to include a sentence citing studies that examined trends in the tropics or Southern Hemisphere.

•Line 24–27: This is a long sentence and could be split into two separate sentences.

•Line 31: Consider including more specific references that show why forced changes in circulation are small compared to internal variability.

•Line 48: Consider whether there are any other limitations of the nudged circulation experiments and include that here.

•Line 49: Add “e.g. the circulation not being in thermal wind balance” to clarify that you don’t mean unresolved processes, which are also mechanisms not represented in the models used.

We added a note on the limitations of nudging and refer to the discussion where we add an example of potential inconsistencies in line 378-380.

•Line 49–51: On the other hand, most of statistical decomposition methods ...” can be rewritten/shortened or split into two sentences to enhance readability.

Changed

•Line 54–56: “Moreover, benchmarks for circulation-induced long-term trends have not been available so far, and to our knowledge no systematic comparison of dynamical adjustment methods has been performed.” It would be good to clarify if this applies globally or just to mid-latitudes and to briefly mention if any emerging efforts exist. This would make it clearer why the study is filling an important gap.

This statement is quite general because – to our knowledge – such benchmarks haven’t been used so far in any region.

•Line 57: Add the specific NH latitude range being examined here.

Done

•Section 2: Adding a sentence or two linking each data/methods subsection to the main aim (separating thermodynamic vs circulation-driven temperature trends) would help the reader understand why each method or simulation is being used.

•Line 69: You could specify the ERA5 years here as well.

Done

•Line 72: “First, three standard historical and future forcing experiments ....”. At first reading, this sounds like three different forcing scenarios or the like. Please specify that you mean three ensemble members.

We changed the wording to “simulations”.

•Line 80: How about specifying some of the main features of the nudging, e.g., whether your nudging is done at the model grid or involves some spectral transformations, and to which vertical level it is done? This way the reader gets a good first impression without having to refer to Topal and Ding (2023) to find out what is “similar” to their approach and what is different.

Thanks, we have added that the nudging involves regular relaxation procedure applied onto the horizontal winds (and all the levels are described). We now just say that the nudging procedure was used in previous studies such as Topal and Ding (2023).

•Line 80: Is CAM6 an abbreviation of something? If yes, please mention.

Thanks, implemented (Community Atmosphere Model Version 6)

•Line 81: “These simulations will be henceforth referred…”. Simplify this sentence for readability.

Done

•Line 87: The phrasing of thermodynamic forcing being represented by “surface temperature” somewhat suggests that using this metric of forcing is not a choice. In reality, temperature change is non-uniform throughout the atmosphere with implications for the midlatitude circulation. We suggest that the authors instead say ‘commonly approximated by GMST’ or similar.

Done

•Line 89: Good point, but it would help to add a short explanation of why it is hard to evaluate these methods in a coupled system.

We discuss this issue in depth in the revised manuscript (see response to other reviewers’ comments on this aspect)

•Line 95: Will the residual internal variability (e.g. from the ocean) influence the evaluation of the decomposition methods, and how do you account for that?

Is discussed in the following paragraph.

•Line 111: “However, we assume that the effect…”: This assumption is reasonable, but you might want to add a reference or a short justification for this assumption.

•Figure 2: It is not clear what is meant with the cooler versus warmer histograms in panels b), c). Please clarify.

•Line 115: Consider splitting the paragraph into two shorter ones: one describing the experimental setup and another explaining its implications and limitations. This would improve readability.

•Line 120: Briefly define AMIP in the text for clarity.

Done

•Line 132: Could split into two shorter sentences for clarity.

•Line 216: Consider providing more details of the transient CESM2 simulations used to train the UNET.

Done

•Lines 219–220: Clarify why the training is done on CESM2 first and then fine-tuning on ERA5—why does this improve performance or robustness?

Done

•Line 220–222: Rephrase to sound more concise and formal.

•Line 227–237: Consider presenting these two sets of bullet points together instead of separately to make the text more concise and easier for the reader to associate each skill metric with what it represents.
Good idea, done.

•Section 3.1: The discussion mentions how the methods differ (DEA captures magnitude, UNET conservative), but the rationale behind these differences could be explained more clearly. For instance, why does UNET underestimate magnitudes?

This is a typical pattern in statistical or machine learning methods that are trained with a loss function that minimizes mean squared error (= the bulk of the distribution, which pays a price at the tails).

•Line 276: This is a long sentence. Consider splitting it into 2–3 smaller sentences to enhance readability.

•Line 289–291: In "The ridge regression … up to 0.6 K/dec", change "where" to "were", and add "*suggest* stronger circulation induced trends …"
•Fig. 4: Which wavelength could approximate the wave-pattern change that you find? Can you relate this to other studies?

•Line 305: This paragraph sounds like a re-introduction of dynamical adjustment from zero. A bit of repetition is appreciated for the flow, but at the current stage this introduction of dynamical adjustment is even clearer than in the introduction (using even more references). Please consider streamlining this or, otherwise, stating more clearly if you mean something different than in the introduction or possibly moving some of this material into the introduction.

•Section 3.3 is purely a discussion. Why not make it a new section called Discussion? There is no new result in this section.

Good idea, changed.

Technical comments:

•Line 35–36 and onward: Check for the use of citet vs. citep and citep[][]{} throughout the paper.

•In Figure 3, the kernel density maps could be enlarged with axes labels shown.

Removed

•Figure A2 is not explained or referenced in the text. Change.

Removed

•Figure B2 is not explained or referenced in the text. Change.

Checked

References:

Brogli, R., N. Kröner, S. L. Sørland, D. Lüthi, and C. Schär, 2019: The Role of Hadley Circulation and Lapse-Rate Changes for the Future European Summer Climate. J. Climate, 32, 385–404, https://doi.org/10.1175/JCLI-D-18-0431.1.

Pfahl, S., Schwierz, C., Croci-Maspoli, M. et al. Importance of latent heat release in ascending air streams for atmospheric blocking. Nature Geosci 8, 610–614 (2015).

https://doi.org/10.1038/ngeo2487