Paper egusphere-2025-5146

Advective, adiabatic and diabatic contributions to heat extremes simulated with the Community Earth System Model version 2

Matthias Röthlisberger, Michael Sprenger, Urs Beyerle, Erich M. Fischer, and Heini Wernli

Final author comments

We thank the reviewers for their comments that will help us to further improve the manuscript. Based on the reviewers' suggestions, we plan to implement the following main changes:

- We better explain certain aspects of our rationale that concern the main objective of our study (evaluating the physical functioning of heat extremes in state-of-the-art climate simulations), which led to critical remarks from the reviewers.
- We will add a paragraph in the Discussion section to elucidate the differences between the approaches of Röthlisberger & Papritz (2023b) and Mayer (2025).
- We will redraw Fig. 2 and indicate where the respective ERA5 value falls outside the range of CESM2 values.

Below we provide a one-to-one response to all points raised by the reviewers. The reviewers' comments are in black and our replies in blue. Line numbers refer to the originally submitted version of our paper.

However, before we address the individual comments, we would like to address the main critique by both reviewers, which is that when comparing CESM2 and ERA5, we don't compare like with like because CESM2 and IFS – the model used to produce the ERA5 reanalyses – differ fundamentally in their horizontal and vertical resolution (and in their dynamical core and utilized parameterizations). We agree with this statement – but we draw different conclusions from this fact. Indeed, CESM2, like most other climate models contributing to CMIP6, has a coarser resolution than ERA5. But this is exactly why it is interesting to validate climate models, or more specifically for our study, to address the question we pose in the first line of our abstract "Do heat extremes in climate model simulations form for the right physical reasons?"! Models like CESM2 are the best tools available to inform society about effects of climate change on heat extremes globally and to separate forced changes from internal variability, thanks to running large ensembles. As discussed in the introduction of our paper, the demand is high for projections of heat extreme frequencies and characteristics in a warming climate and several impact studies of heat extremes on, e.g., health and food security rely on such projections. We are therefore convinced that it is very important to compare the statistics of heat extremes (as done in several earlier studies) but also the underlying physical processes in CMIP-style climate models (like CESM2) with observation-based datasets (like ERA5). By doing so, we regard ERA5 as the "truth" and CESM2 as "one of the state-of-the-art global climate models that provides valuable information about future projections of extremes", and we are fully aware

that the two datasets differ in resolution and are not based on the same numerical model. This is exactly what makes the comparison interesting and relevant, and of course also technically challenging. We will emphasize these general considerations better in the revised version of the paper.

Reviewer 1

Recommendation: major revisions

Summary

R1.1 The study by Rothlisberger et al. is a unique approach to understanding potential biases in the Community Earth System Model (CESM) relative to ERA5 reanalysis for the representation of heat extremes. The methodology is interesting and the findings are, so far, a compelling argument for more in-depth process-level analysis of model behaviour.

Thank you for this positive overall assessment of our study!

R1.2 That said, I believe the study requires major revisions to be accepted. The largest potential bias in the study is the use of different horizontal, vertical, and temporal resolutions for the bulk of the comparison between CESM2 and ERA5.

We don't think of this as a bias of our study but rather its research goal, see our general reply on p. 1 of this document.

R1.3 The authors partially address this by a handful of sensitivity experiments with re-gridded ERA5, but the results are somewhat different across those case studies which calls into question the paper's main findings.

First, we would like to emphasize that we performed a systematic set of sensitivity experiment, which are computationally expensive and technically challenging. To the best of our knowledge, this is the first time that a physical processes study with reanalysis data was repeated step-by-step to adjust the temporal, horizontal, and vertical resolution of one dataset to another. In our view, the results of these sensitivity experiments, documented in Sect. 3.4.1, provide the clear conclusion that (away from high topography) the results documented for ERA5 in Fig. 1 do not systematically change when adjusting the resolution of the ERA5 dataset to the resolution of CESM2. We therefore don't understand why the reviewer writes that "the results are somewhat different across those case studies which calls into question the paper's main findings". We will anyway better explain the rationale for and relevance of the sensitivity experiments in Sect. 3.4.1.

R1.4 In a similar vein, there is no apparent significant testing to determine if the differences between these two model-based products is significant, so it remains difficult to determine if the signals shown here are real. I think this work has the potential to be truly impactful in the

field, so certainly recommend that the authors pursue additional analysis as outlined in the specific comments below for this study.

We address the significance testing in the specific comments below.

Specific Comments

R1.5 Definition of heat extremes – is the hottest day each year truly representative of heat extremes in the model? One could imagine that more important are prolonged (3-5 day) events, and/or those that pass a specific threshold (95th percentile per grid cell over time, e.g.).

We agree that different meaningful approaches exist to define and identify heat extremes. In our view, TX1day events are one such meaningful type of heat extremes: It has been used previously in heat extremes literature (e.g., Röthlisberger & Papritz, 2023), is essentially the temperature-analogue to the frequently used WCRP ETCDDI heavy precipitation index "Rx1day" and the daily mean analogue to the widely used TXx index shown in the Summary for Policymakers of the IPCC AR6 report. The index and has applications in extreme value statistics, as it is the annual block maxima for daily mean temperatures. Furthermore, we are confident that biases identified for TX1day events are also relevant when studying, e.g., prolonged 3 to 5-day heat waves, because many of the prolonged events contain a TX1day event. A further reason for limiting our analysis to the hottest day of each year was the computational demand of our analyses: The analysis presented in Fig. 1 required the calculation of 261'999'360 ERA5-trajectories and 108'822'528 CESM2-trajectories, which took roughly 6 months of computation time on our systems. Systematically considering longer time-scale heat extremes would have further increased the computational demand. Note that using a dataset-specific percentile-based threshold would remove some of the differences between the datasets we want to study here.

R1.6 Line 6: "largest daily mean two-meter temperature" – I would specify again here that it's the largest daily mean anomalous 2 m temperature. Though reading further in the methods, I'm confused since the definition is based on absolute temperature, but the values listed in the abstract are anomalies… Why not base the identification of events based on anomalies as well, for consistency?

In L120 we explain that TX1day events are based on absolute temperatures because we are interested in the maximum daily mean temperature that occurred during a particular year. Once these events are identified, we can calculate their anomalies relative to climatology. Some studies identify the maximum anomalies relative to a time-varying seasonal cycle, which is why we here make clear that the events are initially selected based on the absolute temperature. In the revised version we will mention the climatology earlier to avoid confusion about how we defined anomalies.

R1.7 Line 42: Very minor but suggest removing "literally" – unless this is in reference to thermometers actually shattering! In which case, perfect use of the word.

We removed “literally” as suggested.

R1.8 Section 2.1: Although it is useful to note that the setup is the same as used for the CESM2-LE, additional details would make interpreting this paper’s results easier. What horizontal resolution is the model run at? Is this a fully coupled simulation (with ocean actively feeding back with the atmosphere), or AMIP-style?

Yes, this is a fully coupled simulation with 30 vertical model levels, and the horizontal resolution of the model simulation is 1.25° longitude times ~0. 9° latitude. We will specify this in the revised version.

R1.9 Section 2.2: If CESM2 is presumably run at 1 degree resolution, why keep ERA5 data at half degree instead of matching the spatial resolution in CESM2? Similar for the temporal resolution, why not use 6-hourly to match CESM2? I realize the ERA5 analysis was already published, but consistency here would be incredibly beneficial for supporting the findings!

Our answer here is two-fold: 1) In line with our general reply on p. 1, we disagree with this comment. With ERA5, we would like to get close to reality, and it does not make sense to *a priori* reduce the quality of the ERA5 results by degrading the resolution. Quite to the contrary, our goal was to use ERA5 at the highest spatial resolution for which we could still manage the computational costs of our analysis – this turned out to be 0.5°. Knowing that CESM2 has coarser resolution, we are interested in how well it represents the physics of TX1day events. 2) In our sensitivity experiment (Sect. 3.4.1) we then specifically address the question, whether ERA5 results are mainly different from CESM2 results because ERA5 data has higher temporal and spatial resolution – and we show that this is not the case, which shows that “matching the resolution” as suggested by the reviewer, would not fundamentally change the biases identified in our study.

R1.10 I see this is noted in Section 2.3; suggest moving up mention of the interpolation.

Thank you, good point, we will mention the sensitivity experiments earlier in Sect. 2.

R1.11 Line 124: I don’t know if everyone’s familiar with LAGRANTO 2.0 (I am admittedly not!); it would be worth adding a few sentences to explain how this tool works.

We will be happy to explain a bit more about the trajectory tool LAGRANTO.

R1.12 Line 157: I doubt this will change results much, but again I’d suggest keeping the analysis of ERA5 and CESM2 identical whenever possible; in that case CESM2’s average of T for 1980-1983 should also stem from 1979-1987 instead of extending back to 1976.

We see your point; however, recomputing the T’ decomposition for TX1day events in CESM2 for the early and late years would be a substantial additional effort and since we compare with a fully coupled model that generates its own atmosphere-ocean variability,

individual years should not be expected to directly agree with reanalysis. As a first-order test that the effect of this methodological inconsistency is minor, we computed the differences between CESM2 and ERA5 for 1984-2015, i.e., for the years when the time periods for calculating the climatology were perfectly consistent, and the results look very similar to the ones shown in Fig. 2 (see Fig. A1 at the end of this document). We therefore conclude that our results are not affected by this issue.

R1.13 Figure 2: Some the differences are quite small, as noted in the text around negative delta T' in Fig. 2a, for example. Adding a significance test to determine if the difference is statistically real would strengthen this analysis considerably.

We agree that it is valuable to put the differences between the model and ERA5 into context. In response to your comment, we now add stippling to the region where ERA5 falls outside the range of ensemble members in Fig. 2. This is a simple and transparent approach that has been used in numerous other studies. In contrast to a parametric significance test, it avoids making assumptions about the underlying distribution, which may differ strongly between grid points, and it makes use of the multiple ensemble members.

R1.14 Figure 3: Suggest removing the difference from the bar charts in (a) and (b); it's somewhat distracting to consider separately. Also suggest adding an x label to 3c for clarity.

Thanks for the suggestions, in the revised version we will remove the differences from Fig. 3a, b and add an x-axis label to Fig. 3c.

R1.15 Figure 3c and discussion in the text: Could the legend be moved so as to show the purple line at high deciles? It looks like there's an interesting change in the bias trend from the 9th to 10th decile that could be worth mentioning. Does this show up if different binning is used? Does it indicate something unique physically about these regions?

We moved the legend to not cover parts of the purple line. We don't think that we can provide more explanations about the differences between the 9^{th} and 10^{th} decile.

R1.16 Figure 6: For consistency with other 2-column figures, suggest swapping columns so that the right stays ERA5 and the left stays CESM2.

Thank you, indeed it is better if the panel are arranged consistently across the paper – this will be adjusted in the revised version of Fig. 6.

R1.17 Section 3.4.1: Strongly suggest that any comparison between products be done at the same grid to begin with. The 0.5 degree used here is already not really the "native" grid of ERA5 (should be 0.25).

Again, we disagree (about the first part of the comment), see our general reply on p. 1. If the entire study was done with degraded ERA5, then we would potentially compare CESM2 with "degraded truth". But we would like to compare it with something that is "close to truth", i.e.,

with the highest resolution feasible. The reviewer is correct that this would be ERA5 at 0.25° resolution; however, this would increase the data volume of ERA5 by a factor of 4. Note that we require hourly 3-dimensional fields! When we downloaded ERA5 data we simply could not afford more than 0.5° resolution and also the Lagrangian analyses would have taken at least 4 times longer, which would mean at least 2 years instead of half a year. In this sense, we regard our approach as an ideal compromise, which is "close to the best possible" but still feasible.

R1.18 Table 1: As in other figures, statistical significance would be really helpful here; the values of adiabatic/diabatic T' seem really large when moving from 0.5 to 1 degree resolution, but is it significant?

See our comment to R1.13.

R1.19 Fig 10: Minor point, but since red shading is used it's somewhat hard to see the red star indicating the point of interest. Maybe add an outline to the marker for enhanced visibility? Or a different colour?

Good point, thanks, we will change the colour of the star.

R1.20 Fig 10 & 11: I wonder if it would be more straightforward to combine SH and LH fluxes into evaporative fraction or Bowen Ratio; it would be a bit easier to see the differences if it were just one figure/field.

Agreed. We will combine SH and LH fluxes by presenting the Bowen ration in the revised version of these figures.

R1.21 In general, I'm not sure how much the case studies in Figures 10/11 help illustrate the point, given that as stated in line 403, "the degree to which these case study results can be generalized needs to be evaluated further." More analysis is needed to identify where/when the flux imbalances are key for heat extremes for this section to be impactful in the current study.

We thought that these case studies provide interesting insight into potentially systematic model issues. The case studies have been carefully selected, by focusing on areas where biases in diabatic T' are particularly large. Showing then that flux partitioning is an issue in these areas appears to us as an important result, which is clearly relevant for these areas. With our note of caution in L403 we just made the point that we did not yet investigate this aspect globally. Given the length of the paper, we trust that we can regard this as beyond the scope of this study.

R1.22 Lines 426-427: "However, such effects would likely also depend on the (observed and simulated) base state regarding soil moisture, as further drying in already too dry regions may be underestimated in simulations." – add citations and/or figure refs.

We will add a reference to Ficklin et al. (2016).

R1.23 Lines 434-436: "While the magnitude of TX1day events is modest in these regions, these biases are still worrying, as they point to a rather different physical mechanism leading to TX1day events in CESM2 compared to ERA5 in these regions." – how important are heat extremes of 1 day length over ocean regions though? While interesting that there are differences, I would imagine that ocean heat events are more impactful for SST, as organisms under the surface are impacted vs 2m temperature.

We did not address ocean heat events and their impact on organisms, also because this would be beyond our expertise. However, we think that for a global validation of study of physical processes leading to near-surface atmospheric TX1day events, it is also worth mentioning a few issues over the oceans. We will add a brief statement to make clear that process studies on marine heat waves would need to be studied separately as they most likely differ from TX1day events.

R1.24 Lines 440-441: "Furthermore, we speculate that mesoscale phenomena such as cold pools, which are known to affect temperature variability over subtropical oceans (e.g., Vogel et al., 2021) are resolved even more poorly in CESM2 compared to ERA5." – This is a tough speculation to support based on the evidence so far presented. Suggest removing or strengthening the analysis to support this notion.

Ok, if this is regarded as too speculative, we will omit this statement.

R1.25 Lines 445-449: Given the suggestion of how to strengthen the results here seems relatively clear, is there a reason that this analysis is not undertaken as part of this study? Is the data readily available in the CESM2 experiments?

To the best of our knowledge, no such physical tendency output from different parameterizations is available from any CMIP6 climate model; and even for ERA5 all that would be available are tendencies due to total diabatic heating and radiative heating.

R1.26 In general, the inclusion of multiple ensemble members for CESM2 is a positive feature of the study, but there's rarely mention of spread across ensemble members, statistical significance relative to ERA5, etc.

We agree about the "positive feature". In the revised version of Fig. 2 we will visually highlight the regions where the respective ERA5 value is outside the range of the CESM2 ensembles.

R1.27 Related to the differences in resolution between ERA5 and CESM2, how much of the difference in P (pressure difference of parcel trajectories) is due to the higher vertical resolution in ERA5 compared to CESM2? I would assume that this is a huge factor given the comparatively low resolution in CESM2…

Yes indeed, this might be the case. The differences between ERA5 and CESM2 must be due to a combination of model resolution, model numerics, and model parameterizations. Disentangling the relative importance of these potential factors is extremely demanding. In essence, one would need a CESM2 model run with 100 levels at 0.25° resolution with the numerics and parameterizations from IFS. With a large set of experiments similar to the ones shown in Sect. 3.4.1 but for CESM2, one could then, in principle, quantify how much of the difference is due to model resolution vs. numerics or physics. But the fundamental issues with this idea are that (i) it is currently not feasible (rerunning a large CESM2 ensemble at this resolution with 3D output would require enormous computational and storage resources) and (ii) even worse, it would not provide information about how good the heat extreme processes are in CESM2 as used in the CMIP6 climate projections – which is the goal of this study.

R1.28 Section 4.3: In general, this section is repetitive with the previous two in Section 4. Suggest merging any relevant points and condensing accordingly.

Section 4.3 is short and contains specific recommendations – we think that they are relevant and useful if documented in this way.

R1.29 Data and code availability: "CESM2 data and code underlying this work are available from the first author upon request." This arguably falls short of the current standard for open and reproducible scientist. I suggest archiving all necessary code/data for reproducibility of these results via GitHub, Zenodo, etc.

The volume of the data used in this study is simply too large to be published on an archive like Zenodo. Note that only the 3D data from CESM2 used in our study already amounts to 22.8 TB.

Nevertheless we will more carefully specify where the interested reader finds the code and data to reproduce our results: The LAGRANTO tool used to compute the trajectories is fully described and openly published in Sprenger and Wernli (2015), while python-code to apply the Lagrangian T' decomposition to LAGRANTO-output is published in Röthlisberger and Papritz (2023a).

Reviewer 2

Recommendation: major revisions

This manuscript investigates the causes of the difference in the hottest day of the year in ERA5 and CESM2 using Lagrangian backward trajectories. CESM2 generally overestimates the magnitude of the hottest day. This bias is largely associated with the advection term. Advection is further decomposed into a mean state temperature bias and circulation bias, of which the latter dominates the full advection bias. Even in regions where the ERA5 and CESM2 hottest days are similar, there are compensating errors between the adiabatic and

diabatic terms, suggesting that in those regions CESM2 gets the right result for the wrong reason.

R2.1 Quantifying and understanding model biases is important, especially when the model gets the right result for the wrong reasons. I think this manuscript has the potential to be a valuable contribution to quantifying model biases of heat extremes. My main concern is the usefulness of Eq. 1 for understanding heat extremes.

Note that Eq. 1 follows from the thermodynamic energy equation (see derivation in Röthlisberger and Papritz, 2023b) and it allows attributing the formation of temperature anomalies (i.e., deviations from the local climatology) along the flow of an air parcel to a set of well-defined processes (horizontal transport across climatological temperature gradients, adiabatic warming and diabatic warming), whose relevance for heat extremes formation has been investigated previously by a wide array of studies.

R2.2 Specifically, why is the climatological temperature field used to quantify the advection and adiabatic terms? The intuitive picture I have of advection during an extreme event is the circulation acting on the anomalous temperature field at the time the extreme event, not the climatology. This perspective was also raised by Mayer and Wirth (2025) and Mayer (2025), the latter showing that the Eq. 1 decomposition for the seasonal mean leads to similar results as extremes. This suggests that the differences in the decomposed terms in ERA5 and CESM2 presented here may be dominated by biases in the climatology rather than anomalies therefrom during the extremes. Can the authors better justify using Eq. 1 (as opposed to decomposing based on the anomalous fields as in Mayer) and include in the discussion the advantages and disadvantages of each perspective?

We are grateful for the opportunity to better explain the rationale behind the Lagrangian T' decomposition of Röthlisberger and Papritz (2023b), its advantages and limitations, as well the differences to the perspective adopted by Mayer and Wirth (2025) and Mayer (2025). We split the above comment into several parts and address each of them below.

R2.2a Specifically, why is the climatological temperature field used to quantify the advection and adiabatic terms?

Recall that the specific research question posed by Röthlisberger and Papritz (2023b) was whether T' contributing to heat extremes at a particular location X form because (a) *air from a climatologically warmer region reaches location X*, (b) because air subsides prior to the heat extreme, or (c) because that air is heated diabatically. Each of these mechanisms has been identified by previous studies as being important for heat extreme formation in certain regions of the globe (see, e.g., references in the introductory paragraph of Röthlisberger and Papritz (2023b)). Importantly, Röthlisberger and Papritz (2023b) considered the exact time period during which the temperature anomaly of a heat extreme is forming. This particular research question is reflected in the formulation of Eq. (1), and in particular involves the formation of T' by air parcels crossing *climatological* temperature gradients.

R2.2b The intuitive picture I have of advection during an extreme event is the circulation acting on the anomalous temperature field at the time of the extreme event, not the climatology. This perspective was also raised by Mayer and Wirth (2025) and Mayer (2025), the latter showing that the Eq. 1 decomposition for the seasonal mean leads to similar results as extremes.

In our view, Mayer and Wirth (2025) and Mayer (2025) somewhat erroneously suggested that they provide a different answer to the same research question investigated by Röthlisberger and Papritz (2023b). *Rather, they reach a different answer to a different question.* Contrary to Röthlisberger and Papritz (2023b) (see their research question above), Mayer and Wirth (2025) investigated the question whether heat extremes form at location X because (a) the source-region of the heat-extreme air is climatologically warmer *than that of the "climatological air"* at location X, (b) whether the heat-extreme air subsides *more than the climatological air*, and (c) whether the heat-extreme air is heated diabatically *more strongly than the climatological air*. Moreover, in essence they consider a fixed time-period rather than the time period during which the anomalies form, which leaves a relatively large portion of the temperature anomalies unexplained (Fig. 3d in Mayer, 2025). Note in particular that to construct their "Lagrangian climatology", Mayer and Wirth considered air with all kinds of temperature anomalies (cold and warm anomalies).

To make the conceptual difference in the two research questions fully apparent, consider, for instance, heat extremes in California in Fig. 1 (this study) and Fig. 3 of Mayer (2025). For heat extremes in this region air approaches the particular heat extreme location from a climatologically colder region, i.e., air crosses a climatological temperature gradient from the cold to the warm side prior to contributing to heat extremes. Despite this, Mayer (2025) might still consider "advection" as the main contributor to the heat extremes (at least when neglecting the large unexplained portion of the T' in this region). Their reasoning is in no way mathematically erroneous, it simply doesn't address whether option (a) in the research question Röthlisberger and Papritz (2023b) holds or not. Conversely, in the Röthlisberger and Papritz (2023b) framework, the advective term is negative for TX1day events in California, but that information doesn't help evaluating option (a) in the research question of Mayer and Wirth. Therefore, Mayer and Wirth simply do not address the same question as Röthlisberger and Papritz (2023b).

Nevertheless, we have reservations regarding their finding "horizontal transport is attributed the primary role for [heat] extremes globally", as this finding is likely a consequence of the definition of their "Lagrangian climatology" and, in our opinion, should be challenged considering the substantial body of literature that suggests otherwise: (i) Heat extremes consist of hot air, which, unsurprisingly, has a harder time subsiding than air with a near-zero or negative temperature anomaly. Recall that Mayer (2025) compared adiabatic warming of air parcels contributing to heat extremes (i.e., anomalously warm air) with the adiabatic warming of all air parcels arriving at a certain location, including strongly subsiding (and often anomalously cold) airstreams like dry intrusions (Raveh-Rubin, 2017). This reconciles the stark contrast between the "seeming unimportance" of adiabatic warming for heat extremes in Mayer (2025) with the findings of a large number of previous studies highlighting

the crucial importance of adiabatic warming for heat extremes (e.g., Bieli et al., 2015; Black et al., 2004; Fink et al., 2004; Hotz et al., 2024; Schumacher et al., 2022; Sousa et al., 2018). (ii) Sensible heating of near-surface air is strongly influenced by the temperature contrast between near-surface air and the Earth's surface. The larger that contrast, the larger the heating/cooling (heating rates far larger than those contributing to heat extremes are observed, for instance, during cold air outbreaks over warm ocean surfaces, e.g., Papritz and Spengler, 2017). Therefore, diabatic heating of already anomalously hot air through sensible heat fluxes from the surface is unlikely to be large compared to that of climatological or anomalously cold air. Again, the choice of "Lagrangian climatology" by Mayer (2025) likely explains the contrast between their result (diabatic heating is largely unimportant for heat extremes, their Fig 3c) with the result of a plethora of previous studies suggesting the opposite (e.g., Fischer et al., 2007; Hauser et al., 2016; Miralles et al., 2014, 2019; Seneviratne et al., 2010; Wehrli et al., 2019).

In summary, in any Lagrangian analysis that considers "climatological and anomalous Lagrangian air parcel characteristic", the exact choice of the Lagrangian climatology is crucial and directly affects in which ways anomalous Lagrangian characteristics can be interpreted meaningfully. The approach of Röthlisberger and Papritz (2023b) does not depend on a Lagrangian climatology (only a Eulerian temperature climatology) and circumvents the above issues. Moreover, the comparison of "climatological Lagrangian characteristics" of CESM2 and ERA5 heat extremes presented in this study is closer to an apples-to-apples comparison, as we compare heat extremes with heat extremes. We will insert a short version of the discussion above in the revised Section 4 (Discussion).

This suggests that the differences in the decomposed terms in ERA5 and CESM2 presented here may be dominated by biases in the climatology rather than anomalies therefrom during the extremes.

We acknowledge that with the approach of Röthlisberger and Papritz (2023b) we cannot exclude the possibility that some of the biases we identify for TX1day events also emerge when comparing CESM2 and ERA5 mean fields. Yet, for addressing the question "Do heat extremes in CESM2 form for the right physical reasons?" it is still important to unravel these biases. We will mention this aspect in the revised Section 4 too.

R2.3 In addition, I echo the concerns raised by reviewer 1, especially regarding the inconsistent methods used for ERA5 and CESM2 analyses and the lack of statistical significance. There are several differences between the ERA5 and CESM2 analysis method that may be individually small but is unclear how large they add up to collectively. For example, line 533 suggests the modified definition of t_g only has a marginal effect but how much exactly? The most convincing way to resolve these issues is to make the ERA5 and CESM2 analysis like-for-like.

Regarding the definition of the genesis-time, t_g: The original definition of t_g (the last time step when $T'(\boldsymbol{x}(t), t)$ has the same sign as $T'(\boldsymbol{x}(t), t_X)$ when following a trajectory backwards in time) leads to a residual termed $res1$, because $T'(\boldsymbol{x}(t), t)$ is never exactly zero

for a discrete trajectory. In our ERA5 analyses, the magnitude of $res1$ is less than 0.6 K in most regions on Earth, and significantly smaller than the ERA5 TX1day T' (Fig. 1 of this study and Extended Data Fig. 8 of Röthlisberger and Papritz, 2023b).

The refined definition of t_g by Papritz and Röthlisberger (2023) considered the two time steps before and after $T'(\boldsymbol{x}(t), t)$ last crosses 0, and t_g is then identified as the time step (out of these two) for which $res1$ is minimized. In this study, we apply this refined definition of t_g to CESM2 trajectories (due to the 6-hourly temporal resolution) and the original definition of t_g to ERA5 trajectories. Applying the refined definition also to ERA5 trajectories would minimize ERA5 $res1$, but as pointed out above (and shown visually in Extended Data Fig. 8 of Röthlisberger and Papritz, 2023b), $res1$ in ERA5 is small to begin with.

About the more general concerns about not comparing like-for-like we refer to our general reply on p. 1.

Specific comments:

R2.4 Line 163: I suggest removing "original" here because the original resolution is 0.25, not 0.5 deg.

OK, we removed "original" but note that also the 0.25° resolution is not really "original" as the IFS is a spectral model.

R2.5 Line 190: Fig. 2 -> Fig. 2c

Corrected, thank you.

R2.6 Line 392 and Fig. 10/11b: Oklahoma is too large of an area to describe the location compared to the other regions which seem to be more precise (e.g., city). Also, the red star looks farther north than where Oklahoma is.

Indeed, the naming "Oklahoma" was erroneous. We selected the grid point nearest to 39°N/94.5°W, which is very close to Kansas City. We will correct this mistake in the revised version.

R2.7 Line 485: This should specify the vertical resolution used for the Lagranto analysis is not a major reason. It's not clear whether the native vertical resolution of the dynamical core is important.

Indeed, the vertical resolution of the dynamical core is relevant, and this is one of the reasons why we did this study, see also our general reply on p. 1, and the reply to comment R1.27.

References:

Bieli, M., Pfahl, S., & Wernli, H. (2015). A Lagrangian investigation of hot and cold temperature extremes in Europe. *Quarterly Journal of the Royal Meteorological Society*, *141*(686), 98–108. https://doi.org/10.1002/qj.2339

Black, E., Blackburn, M., Harrison, R. G., Hoskins, B. J., & Methven, J. (2004). Factors contributing to the summer 2003 European heatwave. *Weather*, *59*(8), 217–223. https://doi.org/10.1256/wea.74.04

Ficklin, D. L., J. T. Abatzoglou, S. M. Robeson, and A. Dufficy, 2016: The Influence of Climate Model Biases on Projections of Aridity and Drought. *J. Climate, 29,* 1269–1285, https://doi.org/10.1175/JCLI-D-15-0439.1

Fink, A. H., Brücher, T., Krüger, A., Leckebusch, G. C., Pinto, J. G., & Ulbrich, U. (2004). The 2003 European summer heatwaves and drought – synoptic diagnosis and impacts. *Weather*, *59*(8), 209–216. https://doi.org/10.1256/wea.73.04

Fischer, E. M., Seneviratne, S. I., Lüthi, D., & Schär, C. (2007). Contribution of land-atmosphere coupling to recent European summer heat waves. *Geophysical Research Letters*, *34*(6). https://doi.org/10.1029/2006GL029068

Hauser, M., Orth, R., & Seneviratne, S. I. (2016). Role of soil moisture versus recent climate change for the 2010 heat wave in western Russia. *Geophysical Research Letters*, *43*(6), 2819–2826. https://doi.org/10.1002/2016GL068036

Hotz, B., Papritz, L., & Röthlisberger, M. (2024). Understanding the vertical temperature structure of recent record-shattering heatwaves. *Weather and Climate Dynamics*, *5*(1), 323–343. https://doi.org/10.5194/WCD-5-323-2024

Mayer, A. (2025). A New Global Lagrangian Analysis of Near-Surface Temperature Extremes. *Geophysical Research Letters*, *52*(19), e2025GL116696. https://doi.org/10.1029/2025GL116696

Mayer, A., & Wirth, V. (2025). Two different perspectives on heatwaves within the Lagrangian framework. *Weather and Climate Dynamics*, *6*(1), 131–150. https://doi.org/10.5194/WCD-6-131-2025

Miralles, D. G., Teuling, A. J., Van Heerwaarden, C. C., & De Arellano, J. V. G. (2014). Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nature Geoscience*, *7*(5), 345–349. https://doi.org/10.1038/ngeo2141

Miralles, D. G., Gentine, P., Seneviratne, S. I., & Teuling, A. J. (2019). Land–atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges. *Annals of the New York Academy of Sciences*, *1436*(1), 19–35. https://doi.org/10.1111/NYAS.13912

Papritz, L., & Röthlisberger, M. (2023). A Novel Temperature Anomaly Source Diagnostic: Method and Application to the 2021 Heatwave in the Pacific Northwest. *Geophysical Research Letters*, *50*(23), e2023GL105641. https://doi.org/10.1029/2023GL105641

Papritz, L., & Spengler, T. (2017). A Lagrangian climatology of wintertime cold air outbreaks in the Irminger and Nordic Seas and their role in shaping air-sea heat fluxes. *Journal of Climate*, *30*(8), 2717–2737. https://doi.org/10.1175/JCLI-D-16-0605.1

Raveh-Rubin, S. (2017). Dry intrusions: Lagrangian climatology and dynamical impact on the planetary boundary layer. *Journal of Climate*, *30*(17), 6661–6682. https://doi.org/10.1175/JCLI-D-16-0782.1

Röthlisberger, M., & Papritz, L. (2023a). Lagrangian temperature anomaly decomposition for ERA5 hot extremes, ETH Research Collection, doi:10.3929/ethz-b-000571.

Röthlisberger, M., & Papritz, L. (2023b). Quantifying the physical processes leading to atmospheric hot extremes at a global scale. *Nature Geoscience*, *16*, 210–216. https://doi.org/10.1038/s41561-023-01126-1

Schumacher, D. L., Hauser, M., & Seneviratne, S. I. (2022). Drivers and Mechanisms of the

2021 Pacific Northwest Heatwave. *Earth's Future*, *10*(12), e2022EF002967. https://doi.org/10.1029/2022EF002967

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., et al. (2010). Investigating soil moisture-climate interactions in a changing climate: A review. *Earth-Science Reviews*, *99*(3–4), 125–161. https://doi.org/10.1016/j.earscirev.2010.02.004

Sousa, P. M., Trigo, R. M., Barriopedro, D., Soares, P. M. M., & Santos, J. A. (2018). European temperature responses to blocking and ridge regional patterns. *Climate Dynamics*, *50*(1–2), 457–477. https://doi.org/10.1007/s00382-017-3620-2

Sprenger, M., & Wernli, H. (2015). The LAGRANTO Lagrangian analysis tool - Version 2.0. *Geoscientific Model Development*, *8*(8), 2569–2586. https://doi.org/10.5194/GMD-8-2569-2015

Wehrli, K., Guillod, B. P., Hauser, M., Leclair, M., & Seneviratne, S. I. (2019). Identifying key driving processes of major recent heat waves. *Journal of Geophysical Research: Atmospheres*, *124*(22), 11746–11765. https://doi.org/10.1029/2019JD030635

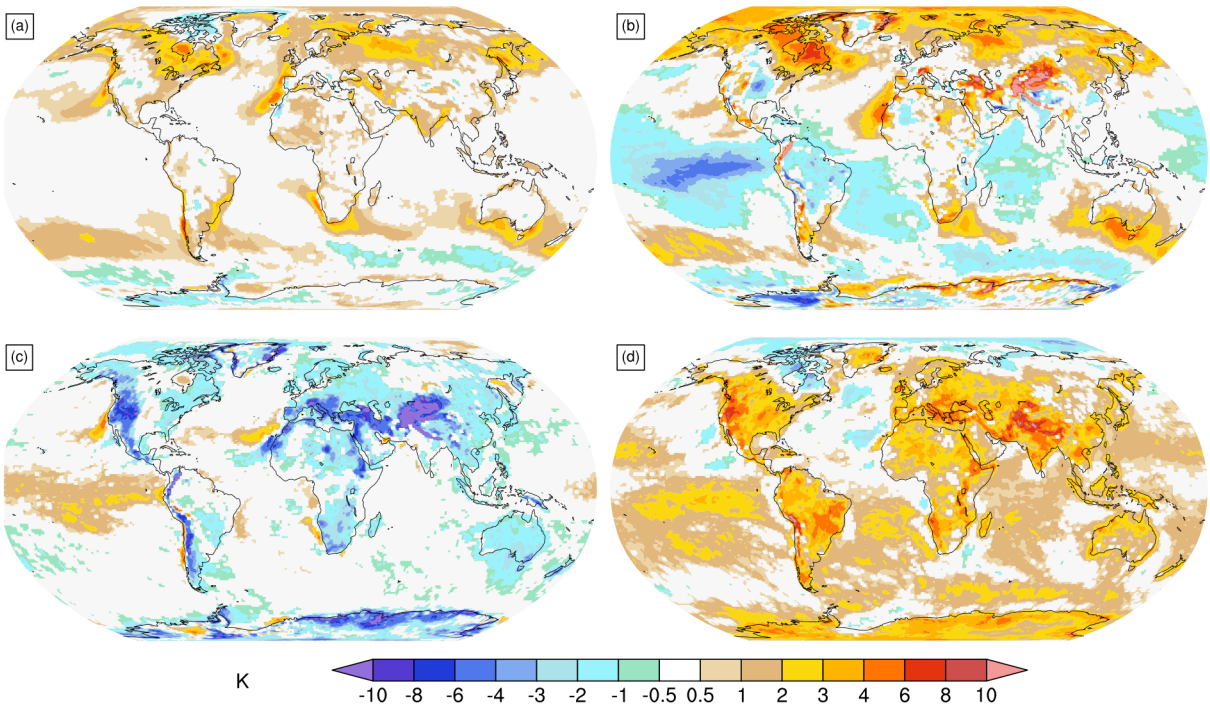

Figure A1: As Fig. 2 but only considering the years 1984-2015, see reply to comment R1.12.