

**We thank both referees for their constructive and detailed feedback, which has substantially strengthened the manuscript. We wish to emphasize at the outset that despite the extensive changes visible in the tracked changes document, the core findings of this study are unchanged. However, the added analyses and discussion requested by the referees have required us to restructure the results section to integrate them coherently, meaning that several subsections have been reorganized and rewritten. In most cases, this is a matter of presentation and sequencing rather than of new or different results, and we have flagged the substantive additions explicitly in our responses below, where we address each comment in turn. The linenumbers in our responses refer to the document without tracked changes.**

Anonymous Referee #1

This manuscript introduces a rigorous and well-designed framework for evaluating the physical plausibility of unprecedented heat-wave events generated through ensemble boosting in climate models. The assessment is comprehensive, examining multiple variables that govern heat-wave evolution through comparisons with reanalysis across temporal evolution and bi- and multivariate relationships. The proposed methodology for evaluating ensemble-boosted events using reanalysis analogues is novel and strengthens the utility of ensemble boosting for climate-risk applications.

The study is a valuable addition to the literature. My only substantive concern is that evaluating plausibility exclusively in terms of standardized anomalies may conceal systematic biases in the model mean state or variability. Consequently, agreement in standardized anomalies does not necessarily imply agreement in the absolute physical state, which is often the quantity relevant for real-world impacts and decision making. The authors explain why this approach is necessary for assessing unprecedented events in future climates, and I agree with this reasoning. However, I think the manuscript would benefit from a more explicit discussion of this trade-off and its implications for interpreting the results in light of storylines, stress-testing and decision-making. With this clarification, I recommend acceptance following minor revision.

**We have added additional analysis and discussion along the lines proposed (lines 263–291 and 626–632).**

General comment:

Because events U1 and its reanalysis analog are generated during the historical period, and thus should have comparable background climates, I wonder whether the authors could also reproduce Fig.~2 using absolute fields rather than standardized anomalies.

This would provide a useful illustration of the influence of systematic model biases and complement the anomaly-based analysis. The comparison could then be used to discuss an important limitation of the framework, namely that it primarily evaluates the plausibility of anomalies rather than the plausibility of the absolute physical state. This distinction is particularly relevant for applications such as stress testing and preparedness, where decisions are often based on absolute thresholds.

**We have modified our multivariate analogue method and now compare composite fields of unprecedented and historical heatwaves (modified Sect. 3.1). As suggested, we do this for both anomalies and absolute values, discussing the effect of biases and the importance of bias correction. We also compare the unprecedented heatwaves in more detail to individual heatwaves in recent years (Fig. B3).**

Minor comment:

Figure 3 labels U4 as originating from parent event P1 (2015), whereas Section 2.3, Table A1, and Figure 4 indicate that U4 originates from parent event P2 (2031). I suspect this is simply a labeling error.

**This was indeed a labelling error. The figure no longer appears in the manuscript, but the labelling in the new Tab. B1 and Fig. B2 is correct.**

Anonymous Referee #2

Recommendation: Major revisions

General comments: This manuscript addresses an important and timely question: how can the physical plausibility of unprecedented events generated by climate models and targeted sampling methods be assessed? This question is increasingly relevant given the growing use of climate storylines, large ensembles, rare-event algorithms, ensemble boosting, and related approaches to explore events outside the observational record.

The manuscript presents a comprehensive analysis of month-long Western European heatwaves generated using ensemble boosting with CESM2. The authors assess these events through comparisons with historical heatwaves in ERA5 and less extreme events in the CESM2 large ensemble, considering historical analogues, individual physical drivers, temporal persistence and substructure, and multivariate relationships.

I find the scientific question compelling and the analysis substantial. My main concerns relate to the conceptual framing of plausibility and the independence and strength of the evidence used to assess it. While the analyses provide multiple lines of evidence

consistent with the simulated events being physically plausible, I am less convinced that the proposed framework provides a sufficiently stringent or falsifiable test of plausibility.

I therefore recommend major revisions aimed at sharpening the conceptual framework, distinguishing different levels of evidence more clearly, and developing a stronger synthesis of the main result. I believe that, with these revisions, the manuscript could make a valuable contribution.

#### Major comments

##### 1. What would constitute evidence against plausibility?

My principal concern is that it remains unclear what outcome of the proposed analysis would lead the authors to conclude that a simulated event is physically implausible.

The framework appears somewhat asymmetric. Similarity to historical events is interpreted as evidence supporting plausibility, whereas the absence of a close historical analogue does not count strongly against plausibility because unprecedented events need not have historical analogues. Individual drivers within historically observed ranges support plausibility, whereas unprecedented combinations of drivers can be interpreted as “extreme anomalies of common drivers.” Internal consistency between variables also supports plausibility, although such consistency might at least partly be expected for events generated by a physically based climate model.

I therefore encourage the authors to address more explicitly the following question: what findings would cause the proposed framework to reject an event as physically implausible, or at least raise substantial concerns about its plausibility?

**We have revised the manuscript to address this important question more directly. Specifically, for each assessment step, we outline in the beginning of the subsection, what results would support and what results would question plausibility (lines 244–248, 295–301, 369–373, 430–439, 512–517, 543–548, 586–593). We then come back to these criteria at the end of the respective subsection to evaluate the implications of our results.**

##### 2. To what extent is the assessment independent of the model generating the events?

A related concern is the extent to which the analysis constitutes an independent test of plausibility rather than primarily a demonstration of internal consistency within CESM2.

The unprecedented events are generated by CESM2, and several components of the analysis compare these events with the CESM2 large ensemble. It is perhaps not surprising that events generated by a physically based model obey the physics of that model, exhibit internally consistent relationships between variables, and involve

stronger or more persistent combinations of processes already operating in less extreme model events.

The comparison of moderate CESM2 heatwaves with ERA5 is a genuinely useful design element in this context. However, agreement for moderate events does not necessarily demonstrate that the model correctly represents the tails of the distributions, the persistence of circulation anomalies, nonlinear land–atmosphere feedbacks, or the combinations of drivers associated with  $4\sigma$  events.

This concern is reinforced by the authors' own persistence results (Sect. 3.3): CESM2 underestimates the persistence of Z500 (blocking) while overestimating the persistence of low soil moisture, and the authors conclude that the simulated events occur for “partly the wrong reasons.” The argument that these biases compensate is plausible but, as presented, largely qualitative. If the persistence mechanism that produces month-long duration is itself biased, this bears directly on whether the simulated magnitude and duration can be trusted, and it sits in some tension with the optimistic tenor of the abstract and conclusions.

I suggest that the authors distinguish much more explicitly throughout the manuscript between:

- model-internal evidence for consistency;
- evidence constrained by the historical climate record;
- evidence that can genuinely be considered independent of CESM2.

This distinction would substantially strengthen the conceptual basis of the paper.

**Given that our analysis is based only on unprecedented events from CESM2, its findings are not independent of CESM2, which the manuscript is transparent about. We have nevertheless rephrased the discussion of this point to make this limitation clearer (lines 690–694).**

**However, we now distinguish more explicitly between the other two points (model-internal evidence vs. evidence from the historical reanalysis record) throughout the manuscript (e.g., new Fig. 4, lines 319–342, 549–571, Sect. 3.3.2).**

**Regarding the persistence results, we discuss them in more detail. We also discuss why, despite these limitations of CESM2, the overall underestimation of temperature persistence leads us to the conclusion that the simulated unprecedented events, if anything, might represent a conservative estimate for worst-case events (lines 614–624). Furthermore, we have adjusted the tone of abstract and conclusions, highlighting more the limitations of CESM2 (lines 10–12 and 706–715).**

### 3. ERA5 should not generally be treated as equivalent to observations

The manuscript frequently contrasts the simulated events with “historical” events and uses ERA5 as the basis for assessing external consistency. However, the degree to which ERA5 variables are constrained by observations differs substantially between variables.

Large-scale atmospheric circulation and temperature are relatively strongly constrained by assimilated observations. In contrast, quantities such as turbulent surface fluxes, soil moisture, precipitation, and aspects of surface radiation are much more strongly influenced by the underlying forecast and land-surface models. ERA5-Land is itself a model-based product driven by atmospheric forcing.

This distinction is important because some of these variables play a central role in the process-based plausibility assessment. Agreement between CESM2 and ERA5 for, for example, turbulent fluxes or soil moisture does not constitute the same level of independent evidence as agreement in large-scale atmospheric circulation. (The manuscript already notes at lines 396–397 that ERA5 does not close the surface energy budget, which underlines the point.)

**We agree with this point and now mention it when comparing surface variables between ERA5 and CESM2 (e.g., lines 278–280). We have also added a dedicated paragraph in the discussion section clarifying this point (lines 676–683).**

### 4. Is the process-based analysis sufficiently process-based?

Given the central framing of the manuscript as a process-based assessment, I found some of the diagnostics relatively simple. Regionally averaged Z500, U500, V500, and  $\omega$ 500 are used as proxies for circulation, horizontal advection, and subsidence. The authors themselves correctly acknowledge that 30-day averaging can obscure important short-term processes and that Lagrangian approaches could provide a more direct analysis of the physical mechanisms.

This raises the question of whether a more explicit process analysis is needed to support the ambition of the manuscript. For example, trajectory analysis, temperature-tendency decomposition, circulation-regime diagnostics, or more explicit measures of blocking and persistence might provide stronger physical evidence. I do not necessarily suggest that all of these analyses are required. However, the authors should explain more clearly why the selected Eulerian diagnostics are sufficient to assess the physical plausibility of events that lie so far in the tail of the temperature distribution, especially since the persistence and blocking biases noted in Sect. 3.3 are precisely the kind of dynamical detail that region-averaged Eulerian proxies are least able to constrain.

**This is a fair point. We have added two more process-based analyses along the lines suggested. First, we perform a Eulerian temperature budget analysis to decompose temperature tendencies on the 850hPa level into advective, diabatic, and adiabatic contributions (new Sect. 3.4.1). Second, we calculate a simple blocking index to assess the occurrence of persistent atmospheric blocking (new Fig. 10b in Sect. 3.4.3). These still follow a Eulerian perspective, and we have rephrased the discussion of this limitation (lines 644–652).**

## 5. Interpretation of the multivariate linear regression analysis

I am not fully convinced by the epistemic role of the multivariate linear regression analysis. From what I understand, the regression model is trained on more ordinary events and subsequently applied to highly extreme events outside the training distribution. Agreement between predicted and simulated temperature anomalies may be interesting, but disagreement would be difficult to interpret: it could indicate physical inconsistency, but it could equally indicate that nonlinear relationships become important in the tails or that the simple statistical model is inadequate. Conversely, agreement does not necessarily provide strong independent evidence for plausibility, because predictors and predictands ultimately originate from the same model physics.

These concerns are compounded by acknowledged features of the model itself: Collinearity between SSR and SM in ERA5, sensitivity to regression dilution, wide confidence intervals on the ERA5-trained coefficients, and the omission of nonlinear land–atmosphere effects, which together make me wary of the relatively strong conclusion that the events are “externally consistent from a multivariate perspective.” I suggest that the authors explain more explicitly what this analysis contributes that is not already demonstrated by the preceding process comparisons, and what possible result of the regression analysis would have counted as evidence against plausibility.

**We have revised this part to make the purpose and contribution of the model clearer (lines 417–419, 430–439, 495–503). To study the effect of non-linearities, we also apply a modified version of the model to predict surface fluxes (lines 476–494, Fig. B4).**

## 6. Independence and representativeness of the ten unprecedented heatwaves

Eight of the ten selected unprecedented heatwaves originate from the same parent heatwave ensemble (P1/2015), while the remaining two (U4, U8) originate from a second parent event (P2/2031). The manuscript acknowledges this fact (and shows in Fig. A1 that the TM time series remain distinguishable), but I think its implications

deserve more discussion. Because the eight P1 members are boosted realisations of a single antecedent atmospheric state, they cannot be regarded as eight independent draws; the effective number of independent unprecedented events assessed here is closer to two. This substantially affects the generality of the conclusions.

This issue does not invalidate the analysis, but the manuscript should be much more explicit about whether its conclusions apply primarily to the plausibility of these particular boosted events, or more generally to unprecedented month-long Western European heatwaves. At present the language tends toward the latter while the evidence supports mainly the former.

**This point is well taken. We have made several changes to the manuscript to address this:**

- 1. We have added more boosting simulations, now including all 20 boosted heatwaves that exceed the historical record.**
- 2. Because all of these heatwaves still originate from the same two parents, we additionally include the most extreme heatwave from the other three parent events which are less extreme in magnitude. These so-called “precedented” heatwaves are added to some of our analysis to increase the generalizability of our results, demonstrating that our conclusion also applies to heatwaves from other parent events (lines 156–162, Sects. 3.3 and 3.4)**
- 3. We have added a more explicit discussion of why this limited set of parent events can still provide a good picture of unprecedented heatwaves generally (lines 214–227).**
- 4. We have modified the new Fig. B1 to demonstrate the distinct behaviours in TM and Z500 and the substantial similarities in SM (lines 157–159).**

Specific comments

Lines 37–44: The discussion of plausibility is useful, but I suggest introducing the issue of falsifiability or rejection criteria already here. The reader should understand from the outset what kind of evidence could count against plausibility.

**We now mention this point briefly here (lines 48–51) and come back to it throughout the manuscript (see response to major comment 1)**

Lines 46–56: The distinction between statistical and process-based approaches could be sharpened. In particular, it would be useful to explain whether the proposed framework is intended as an alternative to statistical model evaluation or as an additional layer of evidence.

**We added some more context about previous studies and now clarify that our approach is intended as an additional layer of evidence (lines 52–66).**

Lines 121–140: The standardisation approach is sensible, but its limitations deserve more discussion. Standardising relative to a time-evolving climatology may obscure physically important changes in the underlying climate state; the later discussion of soil moisture and the shift from energy-limited to moisture-limited regimes (lines 283–288) illustrates precisely this issue. In addition, please make explicit that the requirement of a centred 31-year reference window restricts the ERA5 record used to 1965–2009 (lines 132–133). This excludes several of the most recent and most relevant record heatwaves (2015, 2018, 2019, 2022, and the 2025 event mentioned in the introduction) from both the reference climatology and the pool of historical analogues. Since “historical precedent” is one of the three pillars of the framework, the exclusion of the most extreme recent precedents is a material limitation that should be stated clearly and, ideally, mitigated (e.g. by discussing the trade-off against a longer but less-constrained reanalysis such as 20CR, which is already mentioned at lines 264–265).

**We have made several changes to the manuscript to address this point:**

- 1. We have added additional analysis that compares the unprecedented heatwaves to recent extreme heatwaves since 2011 (Sect. 3.1, Appendix B).**
- 2. In the modified Figure 4, we explicitly look at the effect of comparing different climate states and find the effect to be smaller than the effect of comparing events of different magnitudes (Sect. 3.2).**
- 3. We have rephrased the existing discussion on the trade-off between anomalies and absolute values in lines 626–632.**
- 4. We have added discussion about the choice of reanalysis in lines 670–675.**
- 5. We have extended our anomaly calculation to include the year 2010 (this does not change our conclusions).**
- 6. We have slightly modified our definition of historical heatwaves, now including all 30-day heatwaves in ERA5 that exceed the 95th percentile, yielding 11 historical heatwaves (this does not change our conclusions).**

Lines 186–204: Please clarify more explicitly what outcome of the multivariate regression analysis would be interpreted as evidence against plausibility.

**We have done so (see response to major comment No. 5).**

Lines 215–223: This is an important conceptual paragraph. I suggest distinguishing more explicitly between internal model consistency and external empirical support. The

current terminology of “internally” and “externally” consistent is useful, but “externally consistent” may overstate the independence of ERA5-based evidence for variables that are strongly model dependent.

**We have rephrased the terms in this paragraph (and throughout the manuscript) as suggested (lines 228–237).**

Lines 224–267: The analogue analysis is interesting, but the methodological choices underlying the multivariate distance metric deserve more attention. How are the 13 variables weighted? How sensitive are the identified analogues to the selected domain, variables, and metric? Given spatial autocorrelation and very different process relevance among variables, it is not obvious that equal contributions to a Euclidean distance provide the most physically meaningful measure of similarity.

**We have substantially modified this part (Sect. 3.1). We have replaced the previous analogue analysis by a comparison of the spatial composite fields of unprecedented and historical heatwaves. The findings from this are similar, as the previous analogue-based approach tended to select a very similar subset of historical heatwaves depending on the exact metric used. This new approach is now more robust and demonstrates our point that the spatial dimension between the heatwaves is similar in anomaly space. We also perform a similar analysis for the absolute values where there are more discrepancies as expected (see response to comment above).**

Lines 269–315: I found this one of the most interesting parts of the paper. The result that unprecedented temperatures emerge from unusually extreme combinations of familiar drivers could potentially be elevated to a more central position in the manuscript.

**We have explicitly added this result to both abstract (lines 7–9) and conclusions (lines 700–705).**

Lines 279–288: The discussion of soil moisture, evaporative fraction, and the changing land-atmosphere regime is particularly interesting. It also exposes a limitation of comparing standardised anomalies between historical and future climate states (the same anomaly implying energy-limited behaviour historically but moisture-limited behaviour in the future). I encourage the authors to discuss this more explicitly.

**We have added some more analysis of this effect by now comparing moderate heatwaves in two different climate states and discuss this in more detail (modified Fig. 4, Sect. 3.2, lines 325–335, 343–351).**

Lines 289–297: The authors acknowledge the limitations of the Eulerian analysis. Given the process-based framing of the paper, I wonder whether this should motivate additional analysis rather than being discussed primarily as a limitation.

**We have added additional analyses to address this (see response to major comment 4).**

Lines 298–315: The comparison between ERA5 and CESM2 should acknowledge explicitly that the different ERA5 variables have very different degrees of observational constraint. In particular, agreement in turbulent fluxes and soil moisture should not be interpreted in the same way as agreement in temperature or large-scale circulation.

**We now do so (see response to major comment 3).**

Figure 4: This appears to contain one of the central scientific results, but the figure is visually dense and does not immediately communicate the main conclusion. I encourage the authors to consider a simpler synthesis figure demonstrating that unprecedented temperatures emerge from extreme combinations and persistence of largely familiar physical drivers.

**We have moved panels a and b to the appendix (Fig. B2), and now only show a modified panel c in the main text, which has been complemented by additional moderate events and clearer labelling (new Fig. 4).**

Figures 2 and 3: These figures contain a very large amount of information ( $\approx 26$  panels each). I wonder whether all variables are needed in the main manuscript. A more selective presentation of the dynamically most relevant variables might improve readability, with the complete figures moved to supplementary material.

**We have removed the maps for SHF, LHF, PREC, TX, and TN, reducing the number of panels to 16 (new Figs. 2 and 3). Given the detailed analysis of all drivers in Figs. 4 and B2, the omitted variables in Figs. 2 and 3 provide no relevant additional information and thus they are not included in the appendix.**

Technical comments

The manuscript is generally well written, and I found relatively few purely technical issues.

Table A2 (heatwave H3): the listed dates are inconsistent — day 1 = 1973-08-11 but day 30 = 1973-08-09, so the end date precedes the start date. Based on Table A3 (the analogue for U3 runs 1973-08-11 to 1973-09-09), “day 30” should read 1973-09-09. Please correct.

**This is a good catch; we have fixed it.**

Please use terminology such as “observational record,” “reanalysis,” and “observations” carefully and consistently. ERA5 should not generally be described as observations.

**We have removed all references to “observations” or similar when referring to ERA5, consistently using “reanalysis” instead.**

Consider defining more clearly and consistently the terms “physical conformity,” “internal consistency,” and “historical precedent.” At present there seems some conceptual overlap between them.

**The overlap between these terms is inherent in their definitions, but we have added some more explanations (lines 228–242).**

The number of abbreviations and variables introduced early in the manuscript contributes to the cognitive load. Consider whether all variables need to be introduced and analysed in the main text.

**We believe that all variables should be discussed in the main text. However, to reduce the cognitive load, we have removed the abbreviations H1–H10, instead simply referring to them as “historical heatwaves”. Similarly, we have removed almost all references to U1–U20, instead referring to them as “unprecedented heatwaves” wherever possible.**

Consider moving some methodological detail, particularly material already documented in previous ensemble-boosting publications, to an appendix or supplement.

**We have moved details on the ensemble simulations and return period calculations to Appendix A.**