Dear Editor,

Thank you for the reviews you forwarded on our manuscript “Constructing Extreme Heatwave Storylines with Differentiable Climate Models”.

Following advice from the reviewers, we have revised the manuscript, and we believe it has substantially improved from the previous version. Detailed responses to comments by reviewers are given below, with the reviewer’s comment in black and our response in blue. Quoted text from the revised manuscript is given in *italics*.

The revised version has been uploaded on ArXiv as version 3.

Anonymous Reviewer #1

Section 2.1

1. **Please expand this subsection. It is helpful to introduce the methodology formally, but currently it lacks clarity and rigour. For non-mathematical readers, the description is particularly difficult to follow.**
   A: Thanks to the comments below, we believe the section has substantially improved, where we explicitly state the goal of the optimization, clarify the loss function's competing goals, and define all variables.
2. **At the beginning, one or two sentences reminding the reader of the subsection’s aim (why optimization is needed and what problem it addresses) would help.**
   A: We have added the following sentences to the beginning of Section 2.1: “*Our goal is to find the worst-case physically plausible heatwave trajectory our model can produce. To achieve this, we must find the specific, small perturbations to a known initial state that will evolve into the most extreme event. This search is formulated as an optimization problem, where we define a loss function that the model will automatically minimize in an iterative way to find these optimal initial-state perturbations.*”
   This aim directly motivates the specific form of the loss function, which we justify in our response to comment 4 in section 2.1.
3. **Some notations are undefined (e.g., x^i_0). Please define all variables consistently**.
   A: Thank you for pointing out this inconsistency. The superscript 'i' in $\Delta\vec{x}_0^i$ in the general formulation (Eq 2) was intended to denote a specific perturbation, but it was confusing and inconsistent with its use as a component index in Eq 3. We have removed this superscript from Eq 2 and the surrounding text, and now refer to the

perturbation vector simply as $\Delta \vec{x}_0$. The index 'i' in Eq 3 correctly refers to the different components of the state vector being perturbed (i.e., Temperature, Surface Pressure, etc.), as clarified in our response to comment 8 in section 2.1.

4. **The authors state that Eq. (3) is a good loss function. Could you explain *why* this particular form is appropriate for this case study, and why alternatives were not chosen?**

   A: We have expanded the text to justify this choice. This loss function is formulated to accomplish two competing goals simultaneously: 1) an objective term drives the model toward the desired state (in our case, higher temperatures), and 2) a penalty term keeps the initial perturbation small and physically plausible. In that sense, it is a good loss function as it accomplishes the two requirements. We have clarified in the first paragraph of the section as mentioned above.

5. **Please state explicitly what O(X(t)) and F(O(X(t))) represent for this case study.**

   A: The function O(X(t)) in our case is the temperature at 1000-hPa in some domain and a given period while F(O(X(t))) is the temperature objective term as defined in Eq.3. We have clarified this in the text as follows:

   "In particular, we pick our observable O(X(t)) to be the temperature over a domain D and over a period of time τ at the 1000-hPa pressure level of the model

   $\int_0^\tau \int_D T_{1000}(\phi, \theta, t)\, d\phi\, d\theta\, dt$.

   Multiple functions F(O(X)) can be considered, but our main results use F(X) = c divided by X which gives us the loss."

6. **The term "component" is ambiguous—please clarify what is meant.**

   A: The term "component" in the context of the penalty term (Eq 3) refers to the individual variables of the model's state vector that are being perturbed. As listed in our reply to comment 8 in section 2.1, these are: Temperature, Surface Pressure, Vorticity, Divergence, Specific Humidity, Specific Cloud Ice Water Content, and Specific Cloud Liquid Water Content. We have clarified this in the text by replacing the term "component" by "perturbed variable".

7. **Why was a 5-day averaging window chosen? Is this linked specifically to PN2021, or more generally to temperature autocorrelation?**

   A: The period was chosen to capture the three peak days of the PN2021 and a two-day buffer. Experimentation showed that this two-day buffer made the optimization easier. We have added the following in the text:

   "This 5-day period was chosen to fully encompass the three peak days of the PN2021 event, with a two-day buffer at the end, which we found aided in optimization."

8. **In Eq. (3), what does the index i represent, and what does it span over? Similarly, please define \gamma and \theta (I assume latitude and longitude).**

A: The index i represents the set of variables being perturbed in the optimization process. In our case, with NeuralGCM, we have i={Temperature, Surface Pressure, Vorticity, Divergence, Specific Humidity, Specific Cloud Ice Water Content, Specific Cloud Liquid Water Content}. We clarified the notation in the text.

"*The terms in the loss function are normalized by their initial means, with T_ref representing a characteristic temperature scale and* Δx_i*,ref denoting a reference perturbation scale for each perturbed variable*
i = {Temperature, Surface Pressure, Vorticity, Divergence, Specific Humidity, Specific Cloud Ice Water Content, Specific Cloud Liquid Water Content}."

9. **I find it misleading to call the first term the "heatwave intensity term" when "heatwave intensity" is formally defined later in Section 2.3 but not being the same object.**

   A: Agreed. This was confusing. We have renamed this term in the text to "Temperature objective term" to avoid confusion with the formal "heatwave intensity" metric defined in Section 2.3.

10. **The "objective function" mentioned after Eq. (3) should be clearly defined.**

    A: The objective function was meant to be "loss function". We have replaced "objective function" with "loss function" in the text.

Section 2.2

1. **The decision to optimize only the initial conditions while keeping all other parameters fixed is understandable for tractability and stability, but could limit representativeness of the extremes. Is this a true limitation for your study? If so, I suggest mentioning it explicitly.**

   A: This is a correct assessment. We do not view this as a limitation for our study's aim, but rather as a deliberate methodological choice. We make the assumption that the already trained NeuralGCM provides a skillful and physically consistent representation of the climate, so we fix its parameters (i.e., the neural network weights). Our goal is to find plausible extreme trajectories within the physics already learned by the model. Varying the model parameters themselves would be a different experiment, more akin to fine-tuning, that would risk creating an unphysical model. We have added a sentence to the discussion (Section 4) to clarify this assumption.

   "Our method focuses on optimizing initial conditions, assuming the underlying model physics (whether learned or explicit) are fixed and skillful. An alternative approach could involve optimizing model parameters themselves (as done for

example by Alet et al. (2025) to generate ensembles), though this would require careful regularization to ensure the resulting model remains physically plausible."

2. **The mention of 1.4° resolution in this section is confusing. It is not clear how or when this configuration is used. The text should clearly distinguish which experiments are at 2.8° and which at 1.4°, and why both are mentioned at this stage.**

   A: The reviewer is right that the 1.4° model was introduced without clear context. The bulk of our analysis (Figs. 2-5) is performed at 2.8° resolution for computational tractability and consistency with the 40-year climatology run already available (unfortunately, a 40-year climatology run is not available for the 1.4° resolution model). The 1.4° model resolution experiment is presented later (Section 3.4, Fig. 6) as a sensitivity test to demonstrate that our method is robust and effective when using higher resolution versions of the model. We have revised the text in Section 2.2 to state this clearly.

   "To simulate the dynamics and evaluate the loss function, we use the NeuralGCM model (Kochkov et al., 2024). Most of the experiments are performed with a horizontal grid spacing of 2.8° (denoted as NeuralGCM2.8) because it is more computationally tractable and a long, 40-year climate simulation is available at this resolution. For sensitivity analysis we also consider simulations performed using a horizontal grid spacing of 1.4° (denoted as NeuralGCM1.4)."

3. **The discussion of grid scales, time steps, and numbers of simulations is unclear. Are multiple optimized runs performed, or only one? Is the 75-member ensemble used for both the stochastic NeuralGCM and the optimized runs?**

   A: We have clarified this in the text. We perform two separate, independent optimization runs, which differ only in their hyperparameters (number of steps and λ values), as detailed in Table 1. To improve clarity, the results are now presented as "EXP50" and "EXP75". The 75-member ensemble is a *completely separate set* of simulations generated using the *stochastic* version of NeuralGCM2.8. It serves as our baseline for comparison, analogous to a traditional large ensemble. The text has been revised as follows:

   *"We conducted two independent optimization experiments, hereafter referred to as 'EXP50' and 'EXP75'. Their configurations—including the learning rate (**α**), loss-function weights (β, λ_i), forecast lead times, initialization dates, and number of gradient descent steps (N) —are detailed in Table 1."*

   As discussed later, we have also improved the parameter selection discussion.

4. **If only one optimized run is presented, how robust are the results to "luck" in initialization? How should the uncertainty in the optimization outcome be quantified?**

A: In this study, we present two optimization trajectories (EXP50 and EXP75), both yielding more extreme trajectories than the 75-member stochastic ensemble, suggesting that the result is not simply obtained "by luck". In addition, we perform the experiments at two resolutions. However, a full exploration of the optimization's sensitivity to hyperparameters or small variations in initial state is beyond the scope of this proof-of-concept paper. We agree that this is a key area for future work. As suggested in the discussion, one could explore this by running an ensemble of optimizations from slightly different initial states, or by testing the generated perturbations in a fully-physical model, as we mention in the discussion. We have added a sentence to the discussion to acknowledge this limitation
"Furthermore, the consistency of the results across the EXP50 and EXP75 experiments and the simulations at two different resolutions—all of which yield more extreme trajectories than the 75-member stochastic ensemble—suggests that the optimized perturbations are not simply initialization artifacts."
We have also added some labelling to clarify that there are two experiments, one named EXP50 and another EXP75.

5. **Table 1 is useful, but please explain what the listed parameters mean (like all the \lambdas), why there are two different numbers of steps, and recall the definition of \tau.**
   A: We have improved the table caption:
   "Parameters used during the optimization process. Each row corresponds to one experiment. The coefficients λ_T, λ_SP, λ_δ, λ_ζ, λ_SH, λ_SCIWC, and λ_SCLWC *control the relative weight of the temperature term, the surface pressure term, the divergence term, the vorticity term, the specific humidity term, and the ice and liquid cloud water terms in the loss function. The parameter β sets the strength of the temperature objective term. The number of iteration steps differs between the two experiments in order to explore the effect of longer and shorter optimization procedures while all other settings are kept fixed. The quantity τ denotes the forecast lead time used when computing the loss."*

Section 2.3

1. **Please clarify how you treat events separated by only one day: are these counted as two seprate events or merged as one? Otherwise there is a risk of double-counting.**
   A: Thanks for this comment. Our definition relies on consecutive days. Therefore, if the temperature drops below the threshold for even one day, the event is

considered to have ended, and any subsequent exceedance would be counted as a new, separate event. We have clarified this in the text.
"This definition relies on the persistence of temperature extremes (see also heatwave intensity definition); if the temperature drops below the threshold for even a single day, the event is considered terminated, and any subsequent exceedances are treated as distinct, separate events."

2. **This section suggests that heatwave intensity is central to the study, but it seems to be used primarily in Fig. 4e. Consider clarifying that it is one of several diagnostics used.**
A: Yes, that is correct; we mean to only use the heatwave intensity metric as a diagnostic measure to make sure that the optimized temperature time series led to a heatwave. We have clarified this in the text and moved the subsection into section 2.1.

Section 3.1

1. **On page 7, the reference should be to Fig. 4, not Fig. 2 (caption).**
A: Thanks for pointing this out. The text now refers to Fig.4.
2. **The evaluation against ERA5 is valuable, and I appreciate the authors' transparency in acknowledging that NeuralGCM underestimates extreme heat. Could you provide a possible explanation here (e.g., omission of land–atmosphere feedbacks, as later discussed in Section 4)? Even a brief cross-reference would help.**
A: We have expanded the text and include an additional figure comparing the 1.4° resolution model to support the claim. The text now reads:
" This underestimation of the extreme heat, to our knowledge, is due to two factors: *1) there seems to be a dependence on capturing the extreme with the coarseness of the model, when we increase the resolution to the 1.4° model, the prediction quality improves (see Fig.1C)), and 2) other studies have evaluated the ability of the NeuralGCM at simulating extreme heatwave storylines and found that the absence of key processes, such land-surface feedbacks, results in a systematic underestimation of extreme temperatures (Duan et al., 2025a)."*
3. **The distribution in Fig. 1a appears bimodal for NeuralGCM. Is this an artifact, or is there a physical reason?**
A: This was a consequence of including all seasons in the comparison. Following the recommendation of Reviewer #2, we have refined the analysis to focus exclusively on the summer months. The updated Figure 1 now displays the expected unimodal distribution.

4. **In Fig. 1b, please add a legend to indicate lead times, and specify the simulation period in the caption (otherwise "Day of the month" is hard to interpret).**
   A: The lead times are indicated along the x-axis, which is currently labelled as "Days from peak." To make the lead times easier to identify, we have added stars at the start of each simulation.
5. **While you compare against ERA5 temperature, can NeuralGCM also reproduce circulation fields relevant to heatwaves (e.g., Z500)? Showing this would be useful.**
   A: We have added the ERA5 data in the appendix for reference (see Fig. A1,A2). We show T_{1000} and Z_{500} for the event as is presented in Fig.4a)-d). This allows for a qualitative comparison of the large-scale flow and heatwave intensity of the optimized storylines against the reanalysis. We refer to the appendix in the main text: *"The resulting fields from the optimized solution are consistent with what is seen in ERA5 data for the PN2021 event as can be seen in App.A."*
6. **As far as I understand, Section 3.1 uses the 2.8° version of NeuralGCM (worth recalling at the beginning of the section). Since you later show (Section 3.4) that the 1.4° configuration reduces biases and better captures PN2021, it would be valuable to include an ERA5 comparison for the higher resolution as well. Even a supplemental figure would highlight the importance of resolution for heatwave fidelity.**
   A: We have included a new panel in Fig.1 which compares the ERA5 data to the forecast from the 1.4° resolution model simulation.

Section 3.2

1. **The optimized trajectories are compared with the stochastic ensemble, but not directly with ERA5. Could you show whether the optimized Z500 patterns resemble those observed?**
   A: We have included the equivalent of Fig.4 and Fig.5 using ERA5 data in the appendix. We refer to the appendix in the main text: *"The resulting fields from the optimized solution are consistent with what is seen in ERA5 data for the PN2021 event as can be seen in App.A."*
2. **How many optimized trajectories were run? Is it 75, like the stochastic ensemble, or fewer? Please clarify in the text.**
   A: As clarified in our response to Section 2.2, Comment 3, we performed two independent optimization runs. The results shown (e.g., in Fig. 4) are the final

trajectories from these two runs: one with N=50 gradient descent steps, and one with N=75 steps. This clarification has been added to Section 2.2.

3. **You report a 33% reduction in computational cost. How was this calculated? Please provide details.**

   A: The 33% reduction is calculated by comparing the computational cost of the 'Optimized, N=50' run to the 75-member ensemble run. Since one optimization step is approximately equivalent in cost to simulating one ensemble member, our 50-step run is 33% cheaper than generating the 75-member ensemble. The finding, as shown in Fig. 4, is that this cheaper 50-step run *still produces a more extreme event* than any member of the more expensive 75-member ensemble. We have clarified this in the text.

   "… (a 33% reduction in computational cost relative to generating the 75-member ensemble, calculated as (75-50)/75). Notably, this more efficient 50-step optimization run produces a trajectory more extreme than any member of the 75-member ensemble."

4. **Is there an optimal way to select the number of optimization steps ?**

   A: In theory, there should be, but in practice it is difficult to find. Selecting these parameters is analogous to hyperparameter tuning in machine learning; while an exhaustive, automated search would be ideal, it is computationally prohibitive in this context. We initially chose N=75 to match the computational cost of a standard 75-member ensemble for a fair baseline comparison. We then tested N=50 to see if similar results could be achieved with reduced computational resources, finding that they could, provided λ was retuned. Regarding sensitivity to other values, while it is likely possible to find functional parameters for a much smaller N (e.g., N=10), this would require significant re-tuning of the regularization (λ) and learning rate (α). We have clarified this rationale and the tuning trade-offs in the methods section:

   "We investigate extreme events by perturbing the initial conditions primarily around PN2021 using data from the ERA5 reanalysis dataset (Hersbach et al., 2020). We conducted two independent optimization experiments, hereafter referred to as "EXP50" and "EXP75". Their configurations—including the learning rate (α), loss-function weights (β, λi), forecast lead times, initialization dates, and number of gradient descent steps (N)—are detailed in Table 1. These parameters were selected via an experimental approach analogous to machine learning hyperparameter tuning, as an exhaustive automated search would be computationally prohibitive. We initially selected N = 75 to establish a baseline comparable in computational cost to a 75-member ensemble. Subsequently, we performed the N = 50 experiment to assess whether similar results could be achieved with fewer resources. This required retuning the λi parameters; generally, a

larger N implies a longer search time, allowing perturbations to grow larger, which in turn necessitates a higher λ to constrain their size. Finally, forecast lead times were chosen to strike a balance: sufficiently close to the event to ensure forecastability, yet distant enough to allow the introduced perturbations adequate time to evolve."

5. **Please ensure consistency in the definition of "intensity" across the manuscript, and cross-reference to the section where it was defined.**
   A: The draft was edited to use the term "intensity" exclusively for the defined heatwave intensity metric in section 2.1.

Section 3.4

1. **Why is the 1.4° experiment presented more briefly than the 2.8°, even though it shows better agreement with ERA5? Presenting the 1.4° case in more detail (with the 2.8° as a supporting comparison) would seem the more logical choice.**
   A: This is a fair question. We chose to focus the main analysis on the 2.8° resolution for two practical reasons: 1) the 40-year climatology (Fig. 1a) was run at this resolution, providing a consistent baseline for statistical comparisons, and 2) the 2.8° model is computationally much faster, allowing for more rapid experimentation (as noted in Sec 2.2, it runs on an A4000 GPU, while the 1.4° model requires an A100). We present the 1.4° experiment as a sensitivity test to show that our method is not limited to the coarse model and that the results hold (and are, in fact, improved) at a higher, more realistic resolution.

Conclusion

1. **The statement "a fraction of the computational cost" is vague. Please quantify—e.g., what fraction compared to a 75-member ensemble?**
   A: We have made this statement more precise. The 'fraction' refers to the fact that our N=50 optimization run (which found a more extreme event than all 75 ensemble members) used 33% less computation than generating the 75-member ensemble. We have clarified this in the conclusion.
   "… a fraction of the computational cost of traditional ensemble methods. For example, our 50-step optimization run produced a more extreme event than any member of a 75-member ensemble, while using 33% less computational resources than it took to generate that ensemble."

Anonymous Reviewer #2

Major comments

1. **Table 2 provides information about the perturbations used in the optimized runs (and the range for the ensemble). Is the max across space? This is not clear. What is the spatial structure of the perturbations? Is it constrained in some way, or emerges directly from the differentiation? Can the authors show the perturbations in a figure, and can we learn from their structure?**
   A: The values in Table 2 report the maximum perturbation across all spatial dimensions, including horizontal and vertical levels. We have clarified this in the caption. The perturbations are fully three-dimensional, which makes direct visualization challenging. To provide some insight, we include the spectrum of the perturbations at selected vertical levels in a figure in the appendix (Fig.B1). A more detailed analysis of the spatial structure of the perturbations could be informative, but it is beyond the scope of this work.
2. **Can the authors comment further on the use of N=50 vs N=75? How were these chosen? If we wanted to reduce compute, would we still have success with e.g. N=10? How different would N=75 be from N=something large? I don't necessarily expect the author to do this experiment, but simply comment on their expectations for the sensitivity of the results to this choice.**
   A: Selecting these parameters is analogous to hyperparameter tuning in machine learning; while an exhaustive, automated search would be ideal, it is computationally prohibitive in this context. We initially chose N=75 to match the computational cost of a standard 75-member ensemble for a fair baseline comparison. We then tested N=50 to see if similar results could be achieved with reduced computational resources, finding that they could, provided λ was retuned. Regarding sensitivity to other values, while it is likely possible to find functional parameters for a much smaller N (e.g., N=10), this would require re-tuning of the regularization (λ) and learning rate (α). We have clarified this rationale and the tuning trade-offs in the methods section:
   “We investigate extreme events by perturbing the initial conditions primarily around PN2021 using data from the ERA5 reanalysis dataset (Hersbach et al., 2020). We conducted two independent optimization experiments, hereafter referred to as “EXP50” and “EXP75”. Their configurations—including the learning rate (α), loss-function weights (β, λi), forecast lead times, initialization dates, and number of gradient descent steps (N)—are detailed in Table 1. These parameters were selected via an experimental approach analogous to machine learning hyperparameter tuning, as an exhaustive automated search would be

computationally prohibitive. We initially selected N = 75 to establish a baseline comparable in computational cost to a 75-member ensemble. Subsequently, we performed the N = 50 experiment to assess whether similar results could be achieved with fewer resources. This required retuning the λi parameters; generally, a larger N implies a longer search time, allowing perturbations to grow larger, which in turn necessitates a higher λ to constrain their size. Finally, forecast lead times were chosen to strike a balance: sufficiently close to the event to ensure forecastability, yet distant enough to allow the introduced perturbations adequate time to evolve."

3. **The authors compare to ERA5 early on, but then drop the comparison. Are the optimized heatwaves comparable to ERA5, perhaps after accounting for any biases in the mean state during the early summer period? Do the other variables shown in Figure 5 follow similar trajectories to ERA5?**

   A: We have added Appendix A (Figure A1) which includes the ERA5 evolution of the 1000-hPa temperature and 500 hPa geopotential height for reference. This allows for a qualitative comparison of the large-scale flow and heatwave intensity of the optimized storylines against the reanalysis. We have also added the equivalent of Figure 5 in the appendix (Figure A2). The specific humidity and U-winds seem to follow the range of the NeuralGCM ensemble; the V-winds appear to have a small negative bias. The surface pressure on the other hand has a positive bias for NeuralGCM due to its coarse representation of the orography which is used in the estimation of the surface pressure.

4. **Figure 5 shows that some plausible drivers of the heatwave are largely within the envelope of the original NeuralGCM ensemble, raising the question of what actually caused the extreme heat. One option is that no single driver is extreme, but they are collectively extreme given correlation between them. It may also be worth looking at shortwave radiation and advection if these are available in NeuralGCM.**

   A: The main driver seems to be the amplification of the 500 hPa geopotential height as shown in Fig.4. Unfortunately, NeuralGCM does not output shortwave radiation fluxes. We have included the advection of temperature in Fig.5 for completion but there is no obvious signal. For a more in-depth analysis, one could use the newly found, optimized, initial conditions in a full physics-based model and analyze the full range of interactions. As discussed in the discussion, we will leave this for future work. We acknowledge in the text that: *"While the variables are within range of the ensemble envelope (i.e., not extreme), there might be a confluence of factors that leads to the extreme."*

Specific comments

1. **Please include line numbers in future drafts.**
   A: Unfortunately, we cannot include line numbers on ArXiv.
2. **Page 1: There are a number of papers about the dynamics of the 2021 PNW heatwave beyond the Mass et al study that should be cited, e.g. White et al, https://www.nature.com/articles/s41467-023-36289-3; Neal et al, https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021GL097699; Duan et al, https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2025EF006216 as a starter package.**
   A: Thanks for these suggestions. We have now included the mentioned references in the text. The introduction now reads:
   "The PN2021 heatwave emerged from persistent atmospheric blocking sustained by large-scale Rossby waves that disrupted zonal flow and stalled a high-pressure system over the region (Mass et al., 2024; White et al., 2023). This large-scale setup was fueled by upstream dynamics. Mo et al. (2022) linked it to anomalous atmospheric river activity, while Neal et al. (2022) identified that diabatic heating within the warm conveyor belt of an upstream cyclone provided the necessary Rossby wave activity to establish the block. Once established, the block suppressed cloud formation and drove prolonged subsidence, adiabatically warming near-surface air masses (Loikith and Kalashnikov, 2023). White et al. (2023) corroborated the importance of these mechanisms and estimated via four-day backward trajectory analysis that diabatic processes accounted for approximately 78%
   of the net temperature change of air parcels entering the region, with the remaining ∼22% attributed to adiabatic warming from subsidence. Locally, dry soil conditions further intensified these temperatures through non-linear land-atmosphere interactions (Bartusek et al., 2022; Conrick and Mass, 2023; Schumacher et al., 2022). By studying a 100-member ensemble of PN2021 with varying initial land surface conditions, Duan et al. (2025b) found that variations in antecedent soil moisture led to a spread of approximately 3◦C in peak temperatures, largely driven by regions shifting into a transitional evaporation regime where latent heat flux becomes highly sensitive to soil moisture"
3. **Page 3: The potential role of quasi-resonant amplification is not fully established within the literature for the 2021 heatwave.**
   A: Agreed, removed.
4. **Page 4: Have you confirmed that the 1000 hPa is above the surface at all points in the domain? I suspect it is not based on the topography. Why use 1000 hPa rather than temperature 2m above the surface, which is the more typical choice of variable for heat?**

A: This is a good point. You are correct; due to the σ-coordinates used in NeuralGCM and the high topography of the Canadian Rockies, the 1000-hPa surface is often below surface. Unfortunately, 2-m temperature is not output by the model. To address the sensitivity to the choice of the temperature level, we have analyzed the resulting extreme events at the 850-hPa temperature (T850), which is reliably above the surface across the entire domain. We have included the results in the Appendix C: *"NeuralGCM utilizes σ- coordinates. Over regions with significant elevation, such as the Canadian Rockies, the 1000-hPa geopotential surface is often below ground level. Using the 1000-hPa temperature (T1000) can therefore yield physically inconsistent values when optimizing for near-surface extreme events. To ensure the optimized initial conditions lead to physically meaningful and surface-relevant extreme temperatures across the entire domain, we analyze the 850-hPa temperature (T850). Fig. C1 presents the (T850) fields for the optimized extreme events. Despite the optimization targeting T1000 the temperature at the 850 hPa level still exhibits a clear increase, exceeding the values observed in the ensemble simulations. We note that the magnitude of the anomaly found at 850 hPa is smaller than the maximum value achieved at the 1000-hPa level."*

We also added a mention of the issue in the discussion.

5. **Page 5 / Table 1: Could the authors provide some intuition about the choice of the two sets of parameters for the optimization process?**

   A: The parameters are mostly determined experimentally. We have added some details to the selection as answered in Major comments Q2.

6. **Table 2: The difference between the first and second columns is not clear, and the title of the third column could be improved.**

   A: Column 1 and 2 are two different experiments. One is run with 50 iterations and the other with 75 iterations and their respective parameters in table 1. We have improved the table by clarifying that there are two simulations, one named EXP50, and the other EXP75. The third column represents the ensemble run.

7. **Figure 1a: Given that the heatwave happened in summer, suggest subsetting the data to a relevant summer period (e.g. June-July) and then comparing histograms.**

   A: Agreed, the figure has been updated as per request.

8. **Figure 2: Please reduce the thickness and/or number of contours in the middle panel, since it is hard to see the shading.**

   A: Agreed, the figure has now been updated.

9. **Page 13: Could the authors say more about the dual-initial-condition requirement of purely data driven models?**

A: We only mention these purely data driven models for context and do not run them in this work. As such, we do not have first-hand insight into why they require two consecutive initial states. Our discussion simply notes that this design choice is present in the published implementations. Our own method requires only a single initial condition because it explicitly integrates the governing equations. It is unclear to the authors why the data driven models require that.

10. **There are typos and citation errors with respect to use of in-text vs parenthetical citations that should be corrected.**

    A: Thank you for pointing this out. We have corrected the typos.