Author comments on discussion of preprint, ‘Boosting Ensembles for Statistics of Tails at Conditionally Optimal Advance Split Times’

Justin Finkel and Paul A. O’Gorman

Reviewer 1

General comments

This manuscript explores methodological aspects of rare event simulation, which is a promising line of research to improve our understanding of extreme events in a broad range of complex systems. Specifically, the authors study in a systematic manner some important design choices for resampling algorithms applied to deterministic systems such as the timing and structure of the imposed perturbation and their impact on the ability of the algorithm to sample more extreme events. According to me, the major novelties in the manuscript are the following:

- **introducing an estimator for the probability of resampled trajectories in the “ensemble boosting” method, which allows to compute the climatological probability of the rare events under study.**
- **introducing a different method to generate perturbations at resampling times compared to what was done before in the literature, constructed by sampling a low-dimensional space. This allows to study in details the relation between the perturbation and the resampled amplitude of the event (called severity in the manuscript).**
- **introducing optimality criteria to select “advance split times”.**

To investigate these methodological aspects, the manuscript uses as an example the fluctuations of a tracer concentration in a baroclinically unstable two-layer QG model. The work presented in this manuscript is very rigorous and it is presented in a precise manner. The problems under study are an important part of methodological developments which have a great potential for wide applications. For this reason, the manuscript should be very useful for researchers wishing to implement rare event algorithms in practice. In my opinion this is very high-quality work.

We appreciate your careful reading and positive assessment of its potential value.

Perhaps the only general shortcoming I can see is that due to its technical nature, the manuscript is a bit difficult to read, and its conclusions are restricted to methodological aspects. Maybe one way to broaden its potential audience could be to reinforce the physical content, for instance by discussing the potential importance of extreme mixing events in baroclinic turbulence (typical theories diagnose effective diffusivity based only on typical fluctuations). Given the impressive amount of work that already went into the current manuscript I would not make this a condition for recommending acceptance of the manuscript, but only a suggestion that the authors might want to consider to extend the geophysical impact of their work.

We recognize the high mental load for the reader to digest all components of our paper. Though it’s difficult to provide comprehensive extra background without increasing the length too much, your idea is good to emphasize the physical content of the model and its extremes. The revised version describes our thought process for model selection more thoroughly, in Sect. 1.2:

> “Our primary goal in this study is to establish a general principle for optimizing AST for intermittent extreme events in meteorologically relevant dynamical systems. To balance computational

economy with physical realism, we select a system of intermediate complexity between Lorenz-96 and a moist GCM: a 2-layer quasigeostrophic (QG) flow with a passive tracer. Since its original formulation in Phillips (1956), the 2-layer QG model has served as a useful paradigmatic minimal model for baroclinic instability and associated jets, waves, and vortices in the atmosphere and ocean. It has been augmented in many ways to study specific processes, for example by Lapeyre and Held (2004), who coupled in a moisture component and found the resulting precipitation and latent heating to strongly affect the balance between waves and vortices in the underlying flow. However, even the simpler addition of a *passive* tracer—one without feedback through latent heating—is enough to advance the algorithmic questions we pursue here. Passive tracer dynamics is physically interesting in its own right, as seen by many studies of *intermittency* and heavy-tailed tracer statistics in turbulent flows (Castaing et al. 1989; Gollub et al. 1991; Pumir, Shraiman, and Siggia 1991). In climate science too, extremes of pollution concentration and temperature can be captured partially through passive tracer advection (Bourlioux and Majda 2002; Neelin et al. 2010; Linz et al. 2020)."

Technical questions

In addition to this general comment I have a few questions on the technical content of the manuscript.

- **The question of the effect of introducing a random perturbation on the statistics is very interesting. In the example studied in the paper, Fig. 13 shows that some AST selection procedures leads to apparent bias. Presumably this is just a sampling issue: because these AST selection procedures are less efficient at generating larger severities, the tail probability tends to be underestimated. In addition to the empirical evidence, can you obtain analytical insight on the bias of your estimator?**
  - We do not presently have analytical insight on this phenomenon, although theoretical developments are in progress on even simpler models. To acknowledge the importance of understanding it, we add a heuristic idea to the description of Fig. 13: "Our hypothesis for this behavior is that each ancestor has its own predictability timescale, physically linked to the frequency of the wave responsible for that particular event, and that these ancestor-varying timescales cannot all be respected at once by a single, globally imposed threshold criterion like A^{U} and A^{PC}. The optimization-based criteria A^{EI} and A^{TE} are tailored to the ancestor, and might in fact be choosing those predictability timescales implicitly. This is only speculation, however, and must be validated with more detail than fits in our present scope."
- **Could you extend the method to estimate statistics of any observable conditioned on the fact that the severity is above a given threshold within ensemble boosting? In diffusion Monte-Carlo algorithms this is possible because it performs importance sampling in trajectory space. Is it the case here?**
  - This is an excellent point which we were remiss not to point out in the first draft. The new manuscript has an extra paragraph inserted into the derivation of the MoCTail estimator (middle of Sec. 2.2, new Eq. 11), giving a formula for a generic expectation of interest that closely parallels the estimate for $\mathbb{P}\{R^* > r\}$.
- **another interesting aspect of the manuscript is the idea to use construct perturbations from a low-dimensional sample space. Intuitively it seems that it should lead to smaller variance of the estimators. However, in the case considered here I wonder if this would necessarily be the case. If the growth of the perturbation is indeed governed by the linear baroclinic instability mechanism, a random perturbation directly constructed in streamfunction or vorticity space should project onto the most unstable mode (the same one you are forcing by construction in the manuscript), and the evolution should quickly be dominated by this mode. Could you comment on this?**
  - The most important feature of our perturbations is, indeed, the low-dimensionality, which enables quadrature instead of Monte Carlo for MocTail estimation and is also more interpretable. It is probably also true that the estimators have lower variance, though it's a subtle question because a different perturbation space would imply a different optimal advance split time. Regarding our

specific choice, we urge the reader not to over-interpret the linear stability analysis, because the (4,0) mode is the leading one *only about the background state*, not about the actual initial conditions which vary by ancestor. Also in response to reviewer #2's comment, we have added some clarifying text before specifying the perturbation: "We stress that our focus here is on optimizing AST, not the perturbation space Ω, so the choice of mode is arbitrary so long as Ω remains low-dimensional. The optimal AST would probably change if Ω changes, e.g., if we perturbed a different mode or multiple modes at once; but the *rule for choosing it* based on entropy may well generalize, which will have to be tested in follow-up research.

- **While the optimality criteria introduced in the manuscript are interesting, for practical use one would have to estimate the AST online, without systematically searching for the optimum. Do you expect this to be a limitation for applications?**
  - You're absolutely correct that successful deployment of will need adaptive online optimization algorithms. Here we only indicated *what to optimize*, which is of course a necessary prerequisite. In ongoing research, we are working on methods to do this, but can't claim to have a decisive answer yet. We feel the manuscript already addresses this, especially in the conclusion, but we now add a sentence for emphasis: "Since exhaustive grid search over ASTs and perturbation spaces is not an option when deploying rare event algorithms in practice, we are actively pursuing efficient optimization strategies, which are important to make use of this research.
- **I understand that it is not the direct goal of the paper to design an algorithm which is more computationally efficient than DNS. Nevertheless, it would be natural to expect that it should be the case even without any specific effort. In previous uses of rare event algorithm with climate models for instance, there was an immediate gain of orders of magnitude. Here, Fig. 13 suggests that ensemble boosting with proper AST selection indeed performs better than equal-cost DNS, but it is difficult to appreciate how much really. Would it be possible to do such a comparison, for instance in terms of the return time of the most extreme events simulated? Unlike algorithms of the interacting particle systems family, ensemble boosting is not iterative: the initial ancestors are used to try to simulate more extreme descendents, but these descendents themselves are never used as ancestors again. Do you think this plays a part for computational efficiency?**
  - We have expanded the discussion of Fig. 13 to quantify the speedup more explicitly: "Another way to measure boosting advantage is by 'speedup': given a prescribed χ^2 error, how much extra simulation is needed with DNS relative to boosting to achieve it. These are the horizontal distances between curves. Across all AST criteria, speedup varies from 1.5× to 3×, and accelerates sharply as the DNS curve flattens around $\chi^2 \sim 10^{-1}$ while the boosting curves continue to decrease linearly. Again, we stress that the computational savings here are only incidental, and substantial improvements should be possible by targeted optimization and, potentially, repeated rounds of boosting." We choose not to quantify speedup at the most extreme event generated, because its return period will be highly uncertain. To keep the speedup estimates trustworthy, we have to refrain from making claims about the very far reaches of the tail, even though these are very physically interesting samples.

Specific comments:

- **some of the questions formulated in the abstract are very general and the manuscript addresses only part of them. I would be in favor of a more focused abstract.**
  - We agree with your comment, and have substantially rearranged and pruned the points in the abstract to make it more streamlined and relevant to the paper's content.
- **Fig. 1: the caption is long and it is not immediately evident when looking at the figure what is plotted on panels b)c)d)iii; perhaps add a legend directly on the figure to make it more clear? Even with the caption it is not completely transparent why there are two solid red lines in those panels.**
  - We have added a legend to explain briefly what the two solid red lines represent ("single-peak boosted CCDFs" as opposed to their combined version, "combined boosted CCDFs" in dark red, which constitutes the MocTail).

- **p10: "can be estimated it by. . . ": remove "it".**
  - Thank you, it's fixed.
- **p18, paragraph 2: the relevant timescales, in particular the eddy turnover time, is estimated empirically here if I understand correctly. Is it compatible with the phenomenology of baroclinic turbulence, which should allow you to estimate it a priori from model parameters?**
  - This is a very interesting question to pursue. Held and Larichev (1996) gives estimates for length- and timescales in baroclinic turbulence, which we can apply to our model as follows. The average background wind (which they call U) is $\frac{1}{2}$; the internal Rossby deformation radius is 1 (it is the length scale λ by which we non-dimensionalize); $\beta = 1/4$; therefore the supercriticality $\xi = U/(\beta\lambda^2) = 2$. The eddy time scales as $(k_0V)^{-1} \sim \lambda U^{-1}$ where k_0 is the eddy wavenumber and V is the rms eddy velocity. The eddy wavenumber k_0 may be estimated as $(\xi\lambda)^{-1} = 1/2$; if n eddies fit inside of the domain $L = 6 \cdot 2\pi$, we have $k_0 = 2\pi/(L/n) \implies n = 3$ which is similar to what is seen in Fig. 2 (two to three eddies across the zonal width of the domain). To evaluate the eddy time scaling as an eddy turnover time $T =$ (length scale)/(time scale), arguably there is an additional factor of 2π to convert the wavenumber k_0 to an inverse wavelength such that $T = 2\pi/(k_0V) = 2\pi\lambda/U = 4\pi$ which is roughly consistent with the value of 10 used in the paper. However, given the ambiguity about converting a scaling to one specific value, and given that it is not central to the purpose of the paper, we chose not to include this discussion in the paper.
- **p18, paragraph 3: the results suggest that mixing is more efficient in eastward jets, can you relate this to known results on the phenomenology of baroclinic transport?**
  - The eddy diffusivity may be larger in the eastward jets than in the westward jet because the rms velocity is larger in the eastward jets, but this could be offset by suppression of the diffusivity by the stronger mean flow in the eastward jets. Embarking on a detailed dynamical explanation would considerably lengthen the paper and so we leave this to future work.
- **Fig. 4: shouldn't the GPD parameters exhibit some symmetry with respect to the middle of the domain? Could the lack of such symmetry be related to problems of statistical convergence for the estimators of these parameters?**
  - Somewhat interestingly, no, because GPD parameters describer *upper* tails, which become *lower* tails when reflecting about the midline. We already alluded to the reasons much later in the paper, in Sect. 5.1, but in light of your comment it's clear we should remark on it immediately. Hence we added a new bullet point to the description of Fig. 4:
    - "The mean appears odd-symmetric and the standard deviation appears even-symmetric about the midline, which is not surprising given the tracer boundary conditions which transform as $c \mapsto 1 - c$ when $y \mapsto L - y$, negating the sign of fluctuations but leaving their absolute value constant (or perhaps disrupted slightly by topography). However, the GPD parameters are not symmetric, because they describe the *upper* tail of the local R^* distribution, and the transformation $c \mapsto 1 - c$ swaps the lower and upper tails. The subsequent figures (5 and 7) demonstrate pronounced skewness, so the upper and lower tails are markedly different. These partial symmetries will imprint upon the COAST's latitudinal variation seen later in Figs. 14 and 15."
- **Fig. 7 caption: "from the long DNS (dashed black curves)": did you mean short DNS?**
  - Yes we did, and it's corrected now – thanks for the catch.
- **p30, "displayed in Fig. 9a": the figure does not have multiple panels.**
  - We actually left out the panel labels by mistake, but also don't need to refer to panel (a) in the sentence you point out. Both mistakes are fixed now.
- **section 5.2 and caption of Fig. 8: the notation R^2 for the coefficient of determination is a bit confusing given that R is the intensity, even if you never need to square it. . .**
  - We have removed all R^2s from axis labels.
- **Fig. 9: for long ASTs and small scale parameter s the conditional severity PDFs have most of their mass outside of the sampled severities. Is this robust from a statistical point of view?**
  - Probably not; we would hesitate to trust these conditional PDFs so far beyond the range of ASTs where the quadratic fit is high-quality. We now make note of this in the caption of Fig. 9:

  “Note that the longest ASTs (40 and 32 days) show a substantial probability mass beyond the most-extreme sample. This is a sign of poor quadratic fit, which is consistent with Fig. 8e, and fortunately does not affect the later analysis since optimal ASTs are well short of 32 days.”

- **Fig. 10: I am not fond of the unusual scale used for the two bottom rows, which at first glance gives the impression that the correlation coefficient depends essentially linearly on the AST.**
  - We chose this scale deliberately to make the correlation curves roughly linear, and we still use this scale for choosing the grid of correlation values, but the plot now uses a standard scale. Agreed, it was more confusing than clarifying as shown on the plot.
- **Fig. 11 caption: “Shaded regions show the areas between truncated upper and lower means.” I am not sure I understand the justification for this.**
  - We’ve re-worded the caption in a hopefully clearer way: “Shaded regions indicate variation across ancestors, which we quantify using *truncated upper- and lower-means*. For example, the upper truncated mean for correlation ρ is the mean of ρ across ancestors with above-average ρ: $\mathbb{E}[\rho|\rho > \mathbb{E}[\rho]]$, separately at each AST. We choose truncated means to avoid the awkward properties of more standard measures of spread: interquartile ranges would be erratic for the relatively small sample size of ancestors, whereas standard deviation envelopes can misleadingly fall outside the bounds $[0, 1]$ to which ρ is constrained.”

Reviewer 2

General comments:

This manuscript studies the estimation of rare event probabilities in chaotic and high-dimensional systems, such as climate models. This has been a long-standing challenge: naive sampling methods can accurately estimate the probability of high-probability (or frequent) events, but they lead to large errors when estimating low-probability (i.e., rare) events.

Among existing methods, this manuscript focuses on rare event sampling (RES) with ensemble boosting. As the authors point out, two immediate questions arise: 1) What type of perturbations should one introduce to create the ensemble, yet avoid unrealistic rare events. 2) At what times should the perturbations be applied (Advanced Split Time, AST)? The authors mostly focus on the second question and propose/investigate three heuristics for choosing an optimal AST. Two of these methods are threshold based, whereas the third one is more systematic and selects AST by optimizing a prescribed functional. The authors propose two such functionals: Eq. (26), Expected Improvement (EI) and Eq. (27), Thresholded Entropy (TE). The proposed methods are investigated exhaustively on a two-layer quasi-geostrophic (2LQG) model.

Overall, this manuscript addresses an important aspect of RES based on ensemble boosting. The proposed criteria are motivated well and seem sensible. Furthermore, the authors are admirably detailed in describing their methods and the related numerical results. However, I have a few comments/questions that will hopefully improve the manuscript.

Thank you for your close reading of the manuscript, and for accurately summarizing the main points. We address your comments below to further strengthen the paper.

Specific comments:

- **In principle, the rare event probabilities can be estimated to an arbitrary accuracy if the DNS simulations are long enough. Of course, this would be computationally expensive for exceedingly rare events. Therefore, the main purpose of RES is to reduce the computational cost for a desired accuracy. With that in mind, I was surprised to find no data (table/figure) reporting the computational cost of different methods. The authors should report these in the manuscript since computational cost is a vital piece of information for this work.**

  - Computational cost is indeed central to the idea of rare event sampling. We actually do report the relative costs of different methods in Fig. 13, in the form of curves between χ^2-divergence and number of ancestors for all the different methods tried. The horizontal difference between black (DNS) and red (MoCTail or PoPTail) estimates is the difference in cost for a fixed accuracy (measured by χ^2 divergence), whereas vertical differences show the improvement in accuracy for a fixed cost. However, since both you and Reviewer #1 asked for information about cost, we take the point that this information is a little bit buried. The revision now takes pains to emphasize and direct the reader's attention to this information early in several places. The end of the introduction now states that "even when comparing statistical errors at equal cost, we find (and report at the end of the analysis, in Fig. 13) that our boosted ensembles are already competitive with an equal-cost DNS." We've added "computational cost" to the section 2.3 heading, and in the same section summarize the speedup result: "we do achieve some speedup, even though it is not (yet) our main objective." We also expand the detailed description of Fig. 13 to quantify the cost savings in a more conventional way: "Another way to measure boosting advantage is by 'speedup': given a prescribed χ^2 error, how much extra simulation is needed with DNS relative to boosting to achieve it. These are the horizontal distances between curves. Across all AST criteria, speedup varies from 1.5× to 3×, and accelerates sharply as the DNS curve flattens around $\chi^2 \sim 10^{-1}$ while the boosting curves continue to decrease linearly. Again, we stress that the computational savings here are only incidental, and substantial improvements should be possible by targeted optimization and, potentially, repeated rounds of boosting."

- **Related to the previous comment: if I understand correctly, the equal-cost curves in Fig. 13 indicate that the ensemble boosted RES with the optimal AST is almost as good as a long DNS simulation. In fact, in this figure the RES results do not capture the most extreme events (severity>.67), which makes me think the long DNS is doing even better than RES. At any rate, if at the same computational cost, the DNS results provide us with the same information (or even more) as RES, why should one bother with RES and finding an optimal AST?**
  - There seems to be some confusion between different versions of DNS. The dashed lines in Fig. 13a(i-vi) are supposed to represent "ground truth", which is estimated by concatenating all longitudes together as explained in Fig. 5. This is understandable confusion, because the label "long DNS" on Fig. 13a(vi) should actually read "ground truth", which we update in the new manuscript. Since it has 64-fold as much data as the "long DNS", it is not fair to compare its cost with any of the other lines: what should be compared are "equal-N DNS", "equal-cost DNS", and "MoCTail/PoPTail". By this criterion, boosting is more efficient than DNS, as explained in the text.

- **I understand that the type of perturbations (spatial structure) is not the focus of this paper, but it would be helpful to comment on its impact on AST. For example, do you expect the optimal AST to change if a different type of perturbation is used? Why or why not? The answer to this question is particularly relevant since the perturbations in section 4.1 arise from the most unstable mode around the baroclinically background state. This mode has no physical relevance to the chaotic trajectories being perturbed (since they are presumably far away from the background state).**
  - Absolutely, the optimal AST would probably change with the perturbation space. The reader should be cautious to over-interpret our specific choice here. We now add new emphasis to this point in section 4.1 right before specifying the perturbations: "We stress that our focus here is on optimizing AST, not the perturbation space Ω, so the choice of mode is arbitrary so long as Ω remains low-dimensional. It is important to remember that optimal AST would probably change if Ω changes, e.g., if we perturbed a different mode or multiple modes at once; but the *rule for choosing it* based on entropy may well generalize, which will have to be tested in follow-up research."

- **In Section 3.3 the target variable is introduced as upper-level concentration averaged over a small spatial box. Does this target variable have some physical significance? If so, please explain it more clearly. If not, I'm wondering what the motivation was for choosing this particular "intensity function of interest".**

  - We agree this is useful to justify a bit more. We have added an explanatory passage to Sect. 3.3: “This function is designed to capture the real-world considerations and algorithmic difficulties that originally motivated the AST: it describes *localized* conditions, similar to concentrated pollution, high heat, or heavy rainfall over a region on Earth, and it is mediated by traveling baroclinic waves, and as a result it displays intermittency, with extreme spikes that come and go quickly. The choice of upper- instead of lower-level concentration is simply to weaken the impact of arbitrary aspects of the model setup like the surface topography. Real-world applications would of course refine this choice in many ways, but our choice is suitable for the QG level of model idealization.”
- **Although the authors do a good job reviewing the relevant literature, there are two prominent methods that are missing. One is a rare event estimation method based on LDT, see e.g.,**
  - **(a) G. Dematteis,T. Grafke, & E. Vanden-Eijnden, Rogue waves and large deviations in deep sea, Proc. Natl. Acad. Sci. U.S.A. 115 (5) 855-860 (2018).**
  - **(b) Tong, Shanyin, Eric Vanden-Eijnden, and Georg Stadler. “Extreme event probability estimation using PDE-constrained optimization and large deviation theory, with application to tsunamis.” Communications in Applied Mathematics and Computational Science 16.2 181-225 (2021).**
  - Thank you for pointing out these references. The subtle distinction between methods has prompted us to add a paragraph at the beginning of Sect. 2.2, which hopefully clarifies: “Integrals of the form (5) arise in many diverse risk analysis tasks, such as reliability engineering, where Ω often represents wind, waves, or tremors buffetting a built structure (Au and Beck 2001; Mohamad and Sapsis 2018), and is therefore *high-dimensional*. The default strategy to sample high-dimensional spaces is Monte Carlo, whose famously slow convergence has motivated tremendous methodological innovation to estimate such integrals more efficiently. A particular class of methods based on large deviation theory (Dematteis, Grafke, and Vanden-Eijnden 2019; Tong, Vanden-Eijnden, and Stadler 2021) and first- and second-order reliability methods (Breitung 2021) exploits the shrinking space of possible pathways for increasingly rare events, to estimate integrals with optimization, importance sampling, and Laplace asymptotics. We could certainly use those methods here, but there is a crucial distinction: in our setting, the perturbation space is an arbitrary design choice aiming at an indirect goal (climate estimation), rather than some externally imposed distribution (e.g., a Gaussian process model for ocean bathymetry in Dematteis, Grafke, and Vanden-Eijnden (2019) and Tong, Vanden-Eijnden, and Stadler (2021)). Therefore, nothing stops us here from deliberately choosing low-dimensional perturbations instead of high-dimensional ones as in Francesco Ragone, Wouters, and Bouchet (2018) and Bloin-Wibe et al. (2025). This enables numerical quadrature instead of Monte Carlo or elaborate large-deviation approaches, and saves on cost by allowing sample re-use across different input distributions. It is possible that higher-dimensional spaces are more effective for exciting extreme fluctuations, which would make the above-cited methodologies very useful for our purpose in future research. They can also be useful when conditional risk estimation (for near-term weather forecasting) is the end goal. But our first goal is to determine whether our chosen low-dimensional kicks can suffice for climatological estimation.”
- **Another related method is the variational approach introduced in**
  - **(a) Mohammad Farazmand, Themistoklis P. Sapsis ,A variational approach to probing extreme events in turbulent dynamical systems.Sci. Adv.3,e1701533 (2017)**
  - **(b) Blonigan, Patrick J., Mohammad Farazmand, and Themistoklis P. Sapsis. “Are extreme dissipation events predictable in turbulent fluid flows?.” Physical Review Fluids 4.4 044606 (2019).**
- **Both above methods seek to identify the onset of rare events in a systematic way and therefore closely relevant to the present manuscript. I don’t mean for the authors to compare their RES against these methods, but some discussion seems appropriate.**
  - We now acknowledge these works in the introduction, to help delineate rare event sampling from the closely related goal of optimization: “The need to track probabilities makes rare event *sampling* distinct from *optimization*, i.e., finding the most extreme event possible (or plausible) given physical constraints. That problem that has been attacked successfully with constrained optimization algorithms by Farazmand and Sapsis (2017) and Blonigan, Farazmand, and Sapsis

(2019) for extreme dissipation events in turbulence, and even in AI-based weather forecasting by Whittaker and Luca (2025) for extreme heat waves. RES can benefit from these techniques, but aims to represent the entire tail *distribution* of extremes with statistical fidelity and not just the maximum.”

Technical corrections:

- **Page 14, Eqs (26-27): It is mentioned that the EI and TE are “optimized”. Please clarify whether they are minimized or maximized.**
  - We change “we seek to optimize a functional” to “we seek to maximize a functional” when introducing EI and TE.
- **Page 15: `We have conjectured that ...'' I believe the authors are conjecturing in this manuscript for the first time, but the wordhave”` makes it sound like it was conjectured in a previous study. If so, please add a reference.**
  - It’s now changed to “we conjecture that”.
- **Page 16: “The model setup aims to distill some challenges we have encountered with rare event algorithms across the hierarchy.” It is unclear what you mean by “across the hierarchy”.**
  - We mean the climate model hierarchy, i.e., Lorenz-96 and an aquaplanet GCM, but we’ve deleted “across the hierarchy”, since the following sentences clarify what we mean.
- **Eqs (28-33): A 2LQG model should have a stream function (and passive scalar density) for each layer. But this model seems to have only one stream function. Is there a subscript missing from the dependent variables?**
  - Yes indeed! Thank you for catching this omission. We’ve added the missing z subscripts to ψ and c.
- **On a more general note, I encourage the authors to edit the manuscript thoroughly. There are numerous sentences/paragraphs whose writing can be improved significantly and therefore improve the readability of the paper.**
  - We did our best to write clearly in the first draft, and to improve the clarity further in the second draft.

Community comment 1

General comment

- **The paper is interesting and addresses a timely problem: the scarcity of extreme-event data in climate systems and the need for more efficient rare-event sampling. With the increasing trend and societal impact of extreme events, methods that can better explore tails of the distribution are of clear importance for future research.**

- **The authors aim to identify an optimal Advance Split Time (AST) at which perturbations should be introduced so that rare-event algorithms produce more realistic, diverse, and physically relevant extremes. Instead of relying on traditional threshold-based methods, they develop system-intrinsic indicators that diagnose when perturbations have grown sufficiently to diversify extremes without losing dynamical connection to the original event. They demonstrate this principle first on a simple system and then on a physically meaningful 2-layer quasigeostrophic (QG) model with a passive tracer, illustrating how optimal AST varies with spatial structure, target region, and underlying dynamics.**

  - Thank you for this positive reception of our work. The manuscript is admittedly long and dense (which we feel is necessary to convey our points) but your accurate summary is a positive signal that our overall message does come through. Below we address your comments to improve it further.

Specific comments

The study is thoughtfully executed and provides a promising conceptual foundation. I have some comments that may strengthen the manuscript:

1. **Computational cost. The manuscript does not quantify the computational cost of evaluating multiple AST values or generating boosted ensembles. Since computational efficiency is central to the motivation for rare-event sampling, it would be helpful for the authors to comment on the relative cost of their procedure compared with established splitting algorithms such as AMS or TEAMS. Even approximate scaling behavior (e.g., with ensemble size, model resolution) would be informative.**

The other reviewers raised similar questions. We have fleshed out the discussion of cost and speedup regarding Fig. 13, while stressing the caveat that speedup is not our goal at this stage; rather, that we seek to characterize the nature of optimal AST:

"In terms of quantitative improvements in χ^2 for a fixed cost (vertical differences between curves), all the rules considered ($A^{\mathrm{U}}, A^{\mathrm{PC}}, A^{\mathrm{EI}}, A^{\mathrm{TE}}$) improve substantially upon an equal-N DNS and modestly upon an equal-cost DNS. The size of the advantage varies with N in the way that we expect from boosting: substantial improvements in χ^2 with moderate N, (~5-10) when the DNS has sampled the attractor broadly but sparsely and extremes are within reach by perturbation. The advantage might diminish if N increases enough for DNS to see those extremes without perturbation, but we haven't reached that regime yet. MoCTail and PoPTail performances are similar, but not identical: PoPTail seems more suited for threshold-based rules ($A^{\mathrm{U}}, A^{\mathrm{PC}}$ local and global in b.(i-iii)), whereas MoCTail seems more suited for optimization-based rules ($A^{\mathrm{EI}}, A^{\mathrm{TE}}$ in b.(iv,v)).

Another way to measure boosting advantage is by "speedup": given a prescribed χ^2 error, how much extra simulation is needed with DNS relative to boosting to achieve it. These are the horizontal distances between curves. Across all AST criteria, speedup varies from 1.5× to 3×, and accelerates sharply as the DNS curve flattens around $\chi^2 \sim 10^{-1}$ while the boosting curves continue to decrease linearly. These are modest speedups compared to other published rare event algorithms, which report between one and four orders of magnitude speedup depending on the event definition and the algorithm (e.g., F. Ragone and Bouchet (2021), Finkel and O'Gorman (2026)), but again, we stress that the computational savings here are only incidental to our main goal of characterizing the COAST. Substantial improvements should be possible by targeted optimization and, potentially, repeated rounds of boosting."

2. **Chaotic divergence and event identity.**

Because climate dynamics are chaotic, boosted descendants launched too early may drift toward unrelated extreme configurations. The manuscript discusses decorrelation qualitatively but does not describe a mechanism to ensure that boosted samples still represent intensifications of the same physical event as the ancestor. Could the authors clarify whether additional constraints are needed to maintain physical relevance in boosted ensembles?

Actually we do take care to prevent this loss of event identity. The "argmax drift" parameter δt^*, described at the end of Sect. 2.1, intends to do just that. But our intention could have been worded more clearly, and we have augmented the description in 2.1 accordingly. Hopefully this conveys the intention: "If the perturbation is small, the descendant's peak time $t^*_{n,j,m}$ will be close to the ancestor's peak time t^*_n. However, if the intensity function $R(\mathbf{x}(t))$ tends to oscillate, e.g., with each passing Rossby wave crest, a large-enough perturbation might cause the next wave crest after t^*_n to outgrow the original peak, misappropriating the imposed perturbation to fuel a different event than the original target. Tersely, $t^* = \mathrm{argmax}_t R(\mathbf{x}(t))$ might be a discontinuous function of ω, and $R^*(\omega)$ a non-differentiable function of ω, which is a nuisance for our goal to optimize over ω and, more importantly, complicates the causal chain between perturbation and response. We explicitly prohibit this behavior by restricting the range of $t^*_{n,j,m}$ as follows.

- Set an "argmax drift" parameter δt^* based on physical timescales, e.g., half an oscillation period. Initially set $t^*_{n,m,j} = \mathrm{argmax}\{R(\mathbf{x}_{n,j,m}(t)) : t^*_n - \delta t^* \leq t \leq t^*_n + \delta t^*\}$.
- If $t^*_{n,j,m}$ is a local maximum in R, then don't change it.

- Otherwise, shift $t^*_{n,j,m}$ backward (if at the beginning of the interval) or forward (if at the end of the interval) until it is at a local maximum.

Although it is ad-hoc, this adjustment aims to uphold the core idea of ensemble boosting to *augment existing events*, while preserving their basic identity, rather than *discover totally new events*—which may as well be done by extending the DNS. In general this is a nontrivial condition to impose, as multiple spikes in a sequence may be dynamically correlated to each other, but we use only this simple strategy as demonstration."

3. **Applicability to full climate models.**

The framework is compelling in the idealized QG setting. However, applying entropy-based AST selection and ensemble boosting to operational climate or weather models introduces substantial challenges, including high dimensionality, model biases, observation uncertainty, and the difficulty of maintaining event identity in chaotic flows. Could the authors comment on the main obstacles to such an extension? In particular, do they envision a role for machine learning methods for latent-space reductions or event-type classifiers, which makes the approach computationally feasible in high-dimensional systems?

This is a very important point to consider, and very true that finding the COAST becomes much harder in complex coupled models. Without speculating too much and extending the paper, we have added some emphasis to this point by citing a few key studies in the conclusion, including machine learning-based approaches: "Computational tools such as adjoints, especially in novel machine learning models, invite the use of gradient-based optimization (Wang, Mu, and Sun 2020; Vonich and Hakim 2024; Whittaker and Luca 2025). Since exhaustive grid search over ASTs and perturbation spaces is not an option when deploying rare event algorithms in practice, we are actively pursuing efficient optimization strategies, which are important to make use of this research."

Overall, the paper provides a valuable contribution and opens an important line of inquiry. Addressing these points would clarify the method's practical scope and future potential. - Thank you again for the feedback, which we think has helped solidify and clarify our message.

References

Au, Siu-Kui, and James L. Beck. 2001. "Estimation of Small Failure Probabilities in High Dimensions by Subset Simulation." *Probabilistic Engineering Mechanics* 16 (4): 263–77. https://doi.org/https://doi.org/10.1016/S0266-8920(01)00019-4.

Bloin-Wibe, L., R. Noyelle, V. Humphrey, U. Beyerle, R. Knutti, and E. Fischer. 2025. "Estimating Return Periods for Extreme Events in Climate Models Through Ensemble Boosting." *EGUsphere* 2025: 1–40. https://doi.org/10.5194/egusphere-2025-525.

Blonigan, Patrick J., Mohammad Farazmand, and Themistoklis P. Sapsis. 2019. "Are Extreme Dissipation Events Predictable in Turbulent Fluid Flows?" *Phys. Rev. Fluids* 4 (April): 044606. https://doi.org/10.1103/PhysRevFluids.4.044606.

Bourlioux, A., and A. J. Majda. 2002. "Elementary Models with Probability Distribution Function Intermittency for Passive Scalars with a Mean Gradient." *Physics of Fluids* 14 (2): 881–97. https://doi.org/10.1063/1.1430736.

Breitung, Karl. 2021. "SORM, Design Points, Subset Simulation, and Markov Chain Monte Carlo." *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering* 7 (4): 04021052. https://doi.org/10.1061/AJRUA6.0001166.

Castaing, Bernard, Gemunu Gunaratne, François Heslot, Leo Kadanoff, Albert Libchaber, Stefan Thomae, Xiao-Zhong Wu, Stéphane Zaleski, and Gianluigi Zanetti. 1989. "Scaling of Hard Thermal Turbulence in Rayleigh-Bénard Convection." *Journal of Fluid Mechanics* 204: 1–30. https://doi.org/10.1017/S0022112089001643.

Dematteis, Giovanni, Tobias Grafke, and Eric Vanden-Eijnden. 2019. "Extreme Event Quantification in Dynamical Systems with Random Components." *SIAM/ASA Journal on Uncertainty Quantification* 7 (3): 1029–59. https://doi.org/10.1137/18M1211003.

Farazmand, Mohammad, and Themistoklis P. Sapsis. 2017. "A Variational Approach to Probing Extreme Events in Turbulent Dynamical Systems." *Science Advances* 3 (9): e1701533. https://doi.org/10.1126/sciadv.1701533.

Finkel, Justin, and Paul A. O'Gorman. 2026. "Rare Event Sampling for Moving Targets: Extremes of Temperature and Daily Precipitation in a General Circulation Model." *Journal of Advances in Modeling Earth Systems* 18 (3): e2025MS005456. https://doi.org/https://doi.org/10.1029/2025MS005456.

Gollub, J. P., J. Clarke, M. Gharib, B. Lane, and O. N. Mesquita. 1991. "Fluctuations and Transport in a Stirred Fluid with a Mean Gradient." *Phys. Rev. Lett.* 67 (December): 3507–10. https://doi.org/10.1103/PhysRevLett.67.3507.

Held, Isaac M., and Vitaly D. Larichev. 1996. "A Scaling Theory for Horizontally Homogeneous, Baroclinically Unstable Flow on a Beta Plane." *Journal of Atmospheric Sciences* 53 (7): 946–52. https://doi.org/10.1175/1520-0469(1996)053%3C0946:ASTFHH%3E2.0.CO;2.

Lapeyre, G., and I. M. Held. 2004. "The Role of Moisture in the Dynamics and Energetics of Turbulent Baroclinic Eddies." *Journal of the Atmospheric Sciences* 61 (14): 1693–1710. https://doi.org/10.1175/1520-0469(2004)061%3C1693:TROMIT%3E2.0.CO;2.

Linz, Marianna, Gang Chen, Boer Zhang, and Pengfei Zhang. 2020. "A Framework for Understanding How Dynamics Shape Temperature Distributions." *Geophysical Research Letters* 47 (4): e2019GL085684. https://doi.org/https://doi.org/10.1029/2019GL085684.

Mohamad, Mustafa A., and Themistoklis P. Sapsis. 2018. "Sequential Sampling Strategy for Extreme Event Statistics in Nonlinear Dynamical Systems." *Proceedings of the National Academy of Sciences* 115 (44): 11138–43. https://doi.org/10.1073/pnas.1813263115.

Neelin, J. David, Benjamin R. Lintner, Baijun Tian, Qinbin Li, Li Zhang, Prabir K. Patra, Moustafa T. Chahine, and Samuel N. Stechmann. 2010. "Long Tails in Deep Columns of Natural and Anthropogenic Tropospheric Tracers." *Geophysical Research Letters* 37 (5). https://doi.org/https://doi.org/10.1029/2009GL041726.

Phillips, Norman A. 1956. "The General Circulation of the Atmosphere: A Numerical Experiment." *Quarterly Journal of the Royal Meteorological Society* 82 (352): 123–64. https://doi.org/https://doi.org/10.1002/qj.49708235202.

Pumir, Alain, Boris I. Shraiman, and Eric D. Siggia. 1991. "Exponential Tails and Random Advection." *Phys. Rev. Lett.* 66 (June): 2984–87. https://doi.org/10.1103/PhysRevLett.66.2984.

Ragone, F., and F. Bouchet. 2021. "Rare Event Algorithm Study of Extreme Warm Summers and Heatwaves over Europe." *Geophysical Research Letters* 48 (12): e2020GL091197. https://doi.org/https://doi.org/10.1029/2020GL091197.

Ragone, Francesco, Jeroen Wouters, and Freddy Bouchet. 2018. "Computation of Extreme Heat Waves in Climate Models Using a Large Deviation Algorithm." *Proceedings of the National Academy of Sciences* 115 (1): 24–29. https://doi.org/10.1073/pnas.1712645115.

Tong, Shanyin, Eric Vanden-Eijnden, and Georg Stadler. 2021. "Extreme Event Probability Estimation Using PDE-Constrained Optimization and Large Deviation Theory, with Application to Tsunamis." *Communications in Applied Mathematics and Computational Science* 16 (2): 181–225.

Vonich, P. Trent, and Gregory J. Hakim. 2024. "Predictability Limit of the 2021 Pacific Northwest Heatwave from Deep-Learning Sensitivity Analysis." *Geophysical Research Letters* 51 (19): e2024GL110651. https://doi.org/https://doi.org/10.1029/2024GL110651.

Wang, Qiang, Mu Mu, and Guodong Sun. 2020. "A Useful Approach to Sensitivity and Predictability Studies in Geophysical Fluid Dynamics: Conditional Non-Linear Optimal Perturbation." *National Science Review* 7 (1): 214–23. https://doi.org/10.1093/nsr/nwz039.

Whittaker, Tim, and Alejandro Di Luca. 2025. "Constructing Extreme Heatwave Storylines with Differentiable Climate Models." https://arxiv.org/abs/2506.10660.