Boosting Ensembles for Statistics of Tails at Conditionally Optimal Advance Split Times

Justin Finkel[1,2] and Paul A. O'Gorman[2]

[1]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology
[2]Current affiliation: Department of Geophysical Sciences and the Data Science Institute, University of Chicago

Correspondence: Justin Finkel (jfinkel@uchicago.edu)

Abstract.

Climate science needs more efficient ways to study high-impact, low-probability extreme events, which are rare by definition and costly to simulate in large numbers. Rare event sampling (RES) and ensemble boosting offer a novel strategy to extract more information from those occasional simulated events: small perturbations can turn a moderate event into a severe one, which otherwise might not come for many more simulation-years. But the viability of this approach hinges on two open questions: (1) are boosted events representative of the yet-unrealized events? (2) How does this depend on the specific form of perturbation, i.e., timing and structure? Timing in particular is crucial for sudden, transient events like precipitation. In this work, we formulate a concrete optimization problem for the advance split time (AST) hyperparameter, and instantiate it on an idealized but physically informative model system: a quasigeostrophic turbulent channel flow advecting a passive tracer, which captures key elements of midlatitude storm track dynamics. Three major questions guide our investigation: (1) Can RES methods, in particular *ensemble boosting* equipped with a method to estimate probabilities and *trying-early adaptive multilevel splitting*, accurately sample extreme events of return periods longer than the simulation time when given an optimal AST? (2) What is the optimal AST, and how does it depend on the definition of the extreme event, in particular the local flow conditions around the target location? (3) Can the AST be optimized "online" while running RES?

Our answers support RES as a viable method: (1) RES can meaningfully improve tail estimation, using (2) an optimal AST of 1-3 eddy turnover timescales, which varies weakly but detectably with target location. (3) A certain functional that we call the *thresholded entropy* successfully picks out near-optimal ASTs, eliminating the need for arbitrary thresholds that have thus far hindered RES methods. Our work clarifies aspects of the response function of extreme events to perturbations, and can, in our view, guide future research efforts on optimizing and sampling transient extreme events more efficiently in general chaotic systems.

1 Introduction

1.1 Background and motivation

The outsize impact of extreme weather events, and the need to understand the physical processes that cause them, have driven substantial research interest in the tails of climatological probability distributions. The fundamental challenge is scarcity of

data: the historical record is too short to enable robust estimation of extremes rarer than a few times per century, even if the climate were stationary. Different modeling paradigms have developed to confront the issue. The most straightforward is direct numerical simulation (DNS), whereby a climate model is integrated extensively and the extreme events tallied, either as a single long run with stationary forcing (e.g., Huang et al., 2016; O'Gorman and Schneider, 2009) or as an ensemble with non-stationary forcing (e.g., Thompson et al., 2017; John et al., 2022). This increases the sample size of extreme events, and reduces the relative error (mean/standard deviation) of an empirical estimate $\hat{p} = \frac{\#\text{ extremes}}{N = \#\text{ total samples}}$, but at a slow rate of $\frac{\sqrt{\mathbb{V}[\hat{p}]}}{\mathbb{E}[\hat{p}]} = \frac{\sqrt{p(1-p)/N}}{p} \sim (Np)^{-1/2}$ for $p \ll 1$ (Zuev, 2015). For example, estimating the probability of a once-per-century storm ($p = 0.01$ year^{-1}) to within 10% relative error would take roughly $N = \frac{1}{0.01}(0.1)^{-2} = 10^4$ model years. Most of that simulation time is wasted, just waiting for the next event.

Rare event sampling (RES) takes a shortcut by repurposing that time to generate more extremes instead, perturbing simulations in a targeted way to favor extreme behavior—with the tradeoff of having to account for bias properly. RES was first developed for nuclear safety assessment (Kahn and Harris, 1951), and has since been generalized for diverse applications including structural reliability engineering (Au and Beck, 2001), molecular dynamics (Zuckerman and Chong, 2017), and more recently climate and weather (e.g., Ragone et al., 2018; Webber et al., 2019; Baars et al., 2021). RES stands in contrast to many other strategies which, in one way or another, replace the expensive physical model with a cheaper approximation. Extreme value theory gives principles for parametrically estimating distributions tails (Coles, 2001), but its asymptotic assumptions are not always justified by the finite datasets available, and it is best suited to model univariate distributions (e.g., average temperature over a region) rather than full spatiotemporal processes like storms, although spatial extreme value modeling is steadily progressing (Huser and Wadsworth, 2022; Huser et al., 2025). Hybrid statistical/physical models aim to parameterize physical processes rather than the final output statistics, and include linear inverse models (Penland and Magorian, 1993); stochastic weather generators based on analogues or Markov state models (van den Dool, 1989; Ghil et al., 2011; Yiou and Jézéquel, 2020; Finkel et al., 2023; Pons et al., 2024); empirical downscaling (Vandal et al., 2017; Saha and Ravela, 2024; Rampal et al., 2025); statistical (including machine-learned) emulation (Tebaldi et al., 2020; Boulaguiem et al., 2022; Mahesh et al., 2024a, b); and generative modeling (Watt and Mansfield, 2024; Sundar et al., 2024; Giorgini et al., 2024). Machine learning models in particular are proliferating at a dizzying pace, and they can indeed generate new samples at low cost, but their ability to represent physics outside their training data—perhaps the most essential requirement for extreme event modeling—is rightly regarded with suspicion.

In light of these options, modelers have several tools to help deal with the tradeoff between bias (incorrect physics or limited resolution) and variance (erratic statistical estimates due to limited sample size). The methods are not mutually exclusive, with many interesting synergies possible (e.g., as conceptualized in Lucente et al., 2022), but RES in particular is our focus here as an under-utilized and under-developed strategy to reduce variance without incurring extra bias.

1.2 Rare event sampling: promise and pitfalls

The generic RES procedure can be summarized as follows. We denote the full state vector by $\mathbf{x}(t) \in \mathbb{R}^d$, and the measure of *severity* by R^*: some functional of a trajectory $\mathbf{x}$ that is user-defined, e.g., rainfall averaged over any time interval and spatial region of interest.

1. Generate an ensemble of initial conditions to serve as candidate extreme events. Call these "ancestors".
2. Select a subset of ancestors with high propensity to produce extreme events (large R^*), discarding the others. Apply small perturbations to this subset to generate "descendants": new simulations likely to generate large R^* like their parents, but to do so in diverse ways.
3. Adjust the probability weights downward on these selected ancestors, spreading their weight across their descendants to correct for the over-sampling.
4. Repeat steps 2-3 multiple times on the new, extreme-skewed population, until hitting a termination criterion.
5. Estimate any climatological statistics of interest by taking weighted averages of all the simulations.

This template must be specialized for the kind of target event. Diffusion Monte Carlo (DMC), as applied to season-long hot extremes (with a variant called "GKTL" after its inventors; Ragone et al., 2018) and tropical cyclones (with a variant called "QDMC" that applies quantile mapping to intensity values; Webber et al., 2019), performs the split/kill operation at a chronological sequence of time points, extending the timespan of surviving members while aborting discarded members before they can run to completion—thus, before their R^* values can even be measured. This is appropriate when the propensity for a *future* extreme R^* is well-approximated by some property $R(\mathbf{x}(t))$ measurable at the *present*: for example, if R^* is the mean temperature from June to August, $R(\mathbf{x}(t)) =$ (running average temperature from June 1 to t) is a good splitting criterion (Ragone et al., 2018). If R^* is peak wind speed over a tropical cyclone's lifetime, $R(\mathbf{x}(t)) =$ (minimum sea-level pressure in the eye) is a good splitting criterion (Webber et al., 2019).

But suppose that no good predictor exists. In particular, assume that the severity function R^* of a simulation is the maximum over the event's timespan of a user-defined observable $R(\mathbf{x}(t))$, such as the accumulated rainfall over a small region between $t-1$ day and t, which we generically call the *intensity* function. Assume further that no better predictor for R^* is known besides R itself at the present time. In this case, a better choice of RES algorithm might be adaptive multilevel splitting (AMS; Cérou and Guyader, 2007), or more general versions such as "anticipated AMS" (Rolland, 2022) and "trying-early" AMS (TEAMS), which we previously introduced in Finkel and O'Gorman (2024)—itself a special case of subset simulation (Au and Beck, 2001) from engineering—in which every ensemble member runs to completion and produces an actual value of R^*, not some proxy for it. Descendants are then spawned from the ancestor at some *advance split time* (AST) A before R^* is achieved, to give them enough time to diversify and perhaps exceed their ancestor's severity, but not so much time to forget their ancestor's special initial conditions. Fig. 1 illustrates this tradeoff when selecting AST in the context of a simple stochastic system, namely Langevin dynamics (Pavliotis, 2014) with a logarithmic potential which is specified in Appendix A, but the picture

alone conveys the essential phenomenon of an *optimal* AST. The existence of a nontrivial (i.e., strictly positive) optimum is obvious when looking at isolated events, but its precise value is subtle to quantify when our purpose relates to *climatological* statistics, i.e., averages over many events.

There is no general procedure for selecting AST and other hyperparameters, which impedes the application of RES methods to arbitrary target events and models. We have shown empirically in Finkel and O'Gorman (2024) the existence of an optimal AST—in the sense of accuracy of long return period estimates—that is roughly approximated by the time until $\frac{3}{8}$ of error saturation. But this result might be specific to the Lorenz-96 system and a number of choices made in Finkel and O'Gorman (2024), in particular relating to

1. The target variable defining intensity (energy density, x_k^2, with site index $k = 0$, though for Lorenz-96 all sites are statistically equivalent).

2. The spatial and temporal scale for averaging the target variable (we simply studied the instantaneous maximum at a single site, $k = 0$)

3. The stochastic parameterization (smooth in space, white in time)

4. The metric in which to measure distances between ensemble members (Euclidean distance, $D(\mathbf{x}, \mathbf{x}') = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(x_k - x'_k)^2}$)

Practitioners face a vast menu of choices in all four domains, the first two falling under the purview of domain science and the last two falling under algorithm design. If the physical model or the choice of target variable changes, it stands to reason that the choice of metric should also change, and any single prescription of AST (like the $\frac{3}{8}$-saturation time) is unlikely to work for all cases. Indeed in our recent application of TEAMS to extremes of temperature and daily precipitation in a general circulation model, we found that the $\frac{3}{8}$ rule provided some guidance but underestimated the optimal AST for both temperature and precipitation (Finkel and O'Gorman, 2025). Error norms incorporating global information will be less relevant than local norms around the target region, which tend to saturate more slowly (Finkel and O'Gorman, 2025).

Our primary goal in this study is to establish a general principle for optimizing AST. To explore its possible dependencies that don't exist in Lorenz-96, we upgrade to a 2-layer quasigeostrophic (QG) flow with a passive tracer, whose local concentration is our target variable. The 2-layer QG system is paradigmatic minimal model for baroclinic instability in the atmosphere and ocean, which Lorenz-96 resembles loosely via its Hopf bifurcation structure (van Kekem and Sterk, 2018), and the tracer represents one important part of the dynamics governing precipitation, namely advection of water vapor; we leave the extra complexity of condensation and latent heating to future work. This way, our study provides a common jumping-off point for other advection-related extremes such as pollution loading (Neelin et al., 2010) and temperature extremes (Linz et al., 2020). This path up the model hierarchy has been trodden before by Qi and Majda (2016, 2018), who added passive tracers to Lorenz-96 and a QG model respectively and studied extreme fluctuations in the tracer's Fourier modes. Also, Gálfi et al. (2017) quantified extreme value statistics—including local and global statistics—of QG wind fields themselves. All these works have inspired and guided this one, but we focus distinctly on the link between *short-time perturbation dynamics* and *long-term climate statistics*.

The QG model has enough "space" to explore the effects of all four decision axes listed above on optimal AST. In principle, one can do this with an exhaustive suite of experiments: for every target region (location, size) and every version of stochastic input (e.g., perturbation magnitude and spatial scale) of interest, run TEAMS with a wide range of AST parameters, measure the skill of each AST in matching a reference ground truth distribution, and select the optimal AST. In practice, this exhaustive procedure is not feasible, in part because of the huge number of potential targets, but more fundamentally because TEAMS' performance is *highly subject to randomness*. Measuring the effect of any parameter change on the algorithm's performance requires many repetitions—several dozen at least—to average out the variability inherent in Monte Carlo. Moreover, other hyperparameters related to "population management" exist within TEAMS and other rare event algorithms: the number of initial ensemble members, how many of them to kill and clone at every iteration, and the termination criterion, to name a few. Randomness appears not only as physical forcing, but also in selecting which members to clone, thus interacting tightly with the population hyperparameters. One can think of this as confounding due to sampling bias, which further blurs the imprint of AST itself on performance.

So instead of using TEAMS for our investigation, we turn to a related method of ensemble boosting (Gessner et al., 2021; Fischer et al., 2023). The idea of ensemble boosting is simple: identify some extremes from an initial climatic timeseries, and re-simulate them with perturbed antecedent conditions to generate unrealized but physically plausible (and possibly more extreme) scenarios. By focusing on a limited set of ancestor events to boost, we avoid the additional randomness that occurs in TEAMS as the level is raised and additional ensemble members are stochastically added, which simplifies our investigation. In addition, Bloin-Wibe et al. (2025) has developed an approach to estimate probabilities based on the boosted ensembles, and we have also been developing such an estimator that is introduced below. With the addition of an ability to estimate probabilities, ensemble boosting may now be viewed as an RES algorithm.

We suspect that the optimal AST is closely related to a physically intrinsic quantity that is not particular to a given algorithm. Analogously to Lyapunov exponents, which encode the timescale for small perturbations to double, the optimal AST should encode the timescale for *extreme values of some target variable* to *maximize in variability*. This statement is heuristic, and a primary goal here is to propose some quantities that are very close to the optimal AST and that, like Lyapunov exponents, are intrinsic to the system and don't depend on arbitrary algorithmic choices. We propose and evaluate several candidates including metrics based on entropy and expected improvement.

We have three major contributions. First, we develop a new estimator for low probabilities of extreme fluctuations from boosted ensembles, similar to the estimator of Bloin-Wibe et al. (2025) but distinct in the aggregation step. Our approach includes an optional parametric fit of the response function to perturbations (applicable to both estimators), a simple quadratic regression model that imposes regularity on the resulting severity distribution. Second, we use the two estimators to measure the quality of a range of ASTs across a range of target events (tracer concentration at different target locations), finding evidence for an entropy-based optimality principle. Third, and most importantly from a practical perspective, we demonstrate that both estimators successfully approximate low probabilities when the ensembles are launched from a good AST, which the optimality principle can help to select efficiently. Our goal here is not to demonstrate a performant rare event algorithm—only to elucidate

a necessary ingredient (AST) to be optimized in future algorithms—but even when comparing statistical errors at equal cost, we find that our boosted ensembles are already competitive with an equal-cost DNS.

The rest of the paper is organized as follows. Sect. 2 details the procedure of generating samples and estimating tail statistics, at a model-agnostic level, and proposes several candidate indicators of measuring ensemble dispersion that may help select an optimal AST. Sect. 3 specifies the QG system, its numerical simulation, and its extreme value statistics. Sect. 4 specifies the perturbed-ensemble design at a model-specific level. Sect. 5 visualizes some examples of perturbed events, and how the AST selection criteria behave on these examples. Sect. 6 reports the performance of different AST choices, and visualizes the overall "optimization landscape". Sect. 7 concludes with an outlook and proposed roadmap for subsequent research—theoretical, algorithmic, and applied.

2 Sampling and estimation methodology

Our methodology can be separated into three parts, summarized here and expounded in three subsections. For a given target variable and location defining the extreme event, we

1. run a relatively short direct numerical simulation ("short DNS"), identify the extreme events within it, and generate a dataset of boosted ensembles for each event at a range of ASTs;
2. estimate tail distributions, conditional on the event and the AST;
3. combine the conditional tails into an unconditional ("climatological") tail, using the estimators specified below, for a range of ASTs, and select the optimal AST based on the skill of the corresponding tail estimate in reproducing the tail of a "long DNS".

We then display the results of applying this procedure to a range of target locations in the model flow domain.

2.1 Generating the dataset of boosted ensembles

There are many design choices in ensemble boosting (Gessner et al., 2021): how to select extreme events to boost, how many boosts to generate, when to launch them, etc. This subsection details the choices used here.

We run a direct numerical simulation ("short DNS") $\{\mathbf{x}(t) : 0 \leq t \leq T_{\text{short}}\}$, long enough to generate some extremes but not enough to estimate probabilities smaller than $1/(\epsilon^2 T_{\text{short}}) = 100/T_{\text{short}}$ for a relative error tolerance of $\epsilon = 0.1$. The premise of RES, and ensemble boosting, is that the extremes it does generate might have been even worse, perhaps just a butterfly flap away from the more intense extremes one would see with a "long DNS" of duration $T_{\text{long}} \gg T_{\text{short}}$. We generate such a long DNS as well to serve as a ground-truth for validation. Following the ensemble boosting methodology laid out in Gessner et al. (2021); Gessner (2022); Fischer et al. (2023) and Noyelle (2024), we first identify a threshold μ with exceedance probability $q(\mu)$ that is moderate enough to estimate precisely with the short DNS. In other words, μ is the $[1-q(\mu)]$th quantile, or "$q(\mu)$th complementary quantile". Equivalently, $q(\mu)$ is the *complementary cumulative density function* (CCDF) of the random variable

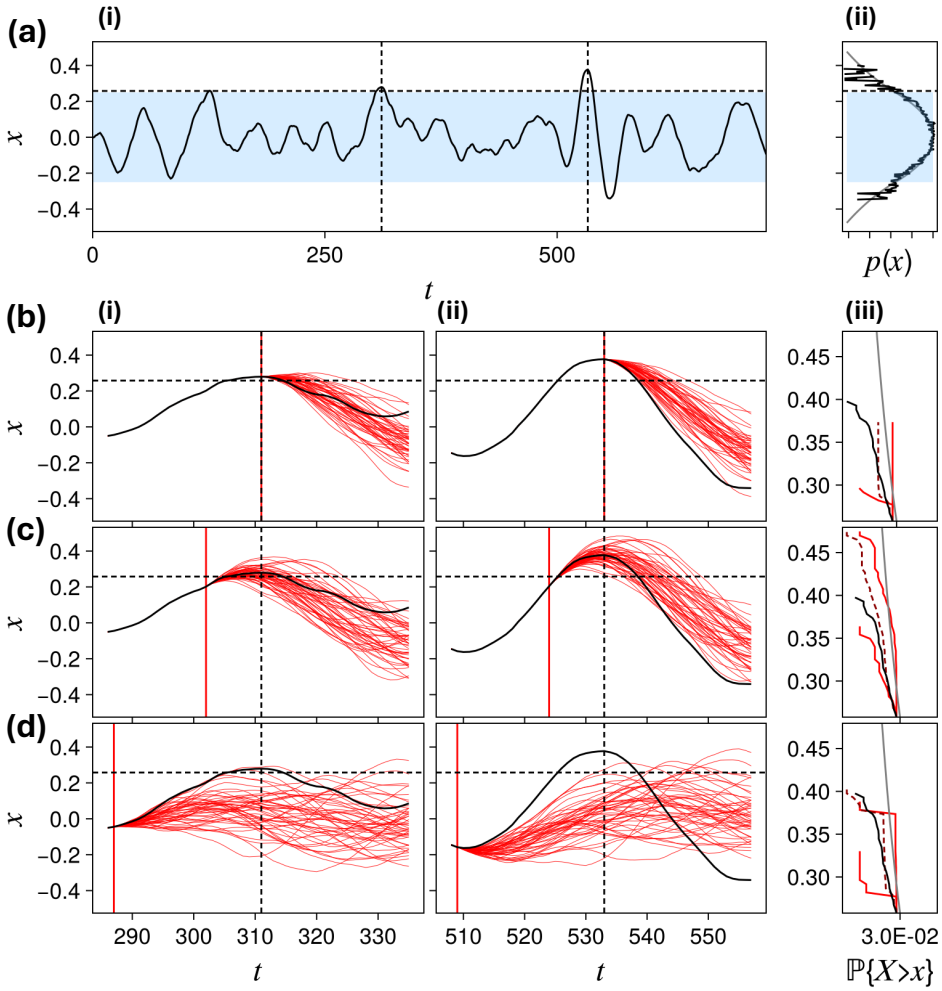

Figure 1. Schematic summarizing the ensemble boosting and tail estimation procedure, using a simple Langevin dynamics with a potential that is quadratic for $x \in (-0.25, 0.25)$—the blue-shaded region—and logarithmic outside this range. Appendix A specifies the system completely. The position variable $X(t)$ exhibits intermittent, transient extremes (a.i) and power law tails $\mathbb{P}\{|X| > |x|\} \sim |x|^{-3.1}$ (a.ii). We set a threshold for severity (horizontal black dashed line) at roughly the minimum probability estimable from a relatively short (duration 1600) timeseries (see the black empirical PDF in a.ii and the black empirical CCDFs in (b,c,d).iii, as compared with the true PDF and CCDF in gray). We then identify the peaks over the threshold (marked by vertical black dashed lines in a.i), and perturb the simulation in advance of these peaks. Three choices of advance split time (AST) are shown in rows b,c,d, marked by vertical red lines, each resulting in "boosted" peak ensembles, shown as red curves in (b,c,d).(i,ii) and summarized by complementary CDFs (CCDFs) shown in light red in (b,c,d).(iii). Combining these conditional CCDFs together using the "MoCTail" estimator introduced later in Eq. (16) gives the dark red dashed line, which is meant to approximate the ground truth (gray line) better than the short DNS alone can do, including by going to higher values of x. The intermediate AST (c) is best among the three for this task, and our goal is to formulate and characterize this optimal AST more generally.

R, evaluated at μ. In line with the *peaks-over-threshold* procedure (Coles, 2001), we take cluster maxima of exceedances above μ as the "ancestral" extreme events. Concretely, a cluster maximum is a state from the DNS, $\mathbf{x}^* = \mathbf{x}(t^*)$, such that

$$R^* = R(\mathbf{x}(t^*)) = \max\Big\{R(\mathbf{x}(t)) : t^* - A_{\text{max}} \le t \le t^* + B\Big\} > \mu. \tag{1}$$

where A_{max} and B are buffer times longer than the mixing timescale of the dynamics (i.e., how long two perturbed simulations need to become independent), ensuring that two consecutive events $\big(\mathbf{x}(t_n^*), \mathbf{x}(t_{n+1}^*)\big)$ are genuinely independent from each other. A_{max} is an upper bound on the ASTs used for boosting.

We collect all such peaks occurring in the short DNS,

$$\{\mathbf{x}_n^* = \mathbf{x}(t_n^*) : n = 1, \ldots, N_{\text{short}}\}, \tag{2}$$

and for a sequence of increasing ASTs $\{A_j : j = 1, \ldots, J\}$ bounded between 0 and A_{max}, launch an ensemble of descendants $\{\mathbf{x}_{n,j,m}^* : m = 1, \ldots, M_{n,j}\}$ by applying $M_{n,j}$ different perturbations to the DNS at time $t_n^* - A_j$, and running each simulation to time $t_n^* + B$. Note that $M_{n,j}$ could in principle vary between ancestors n and lead times j, which is not needed for our exhaustive sweeps in this paper, but certainly would be needed in an "online" rare event sampling procedure that iteratively homes in on a subset of the most extreme-ogenic ancestors $\{n\}$ and ASTs $\{j\}$ to draw more samples from.

A bit more notation helps clarify how the perturbing is done, abstractly at first and concretely in Sect. 3 when we specialize to the QG system. For each (n, j, m), we draw a random sample $\omega_{n,j,m}$ from some sample space Ω. Denoting $\Phi^{\Delta t} : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ be the flow map that integrates the perturbed dynamics forward by a time interval Δt, the (n, j, m)th descendant's trajectory through state space $\mathbb{R}^d$ can be written

$$\mathbf{x}_{n,j,m}(t) = \begin{cases} \mathbf{x}(t) & \text{for } t_n^* - A_{\text{max}} \le t \le t_n^* - A_j \\ \Phi^{t-(t_n^*-A_j)}\Big(\mathbf{x}(t_n^* - A_j), \omega_{n,j,m}\Big) & \text{for } t_n^* - A_j < t \le t_n^* + B. \end{cases} \tag{3}$$

In words, the descendant shares its ancestor's past up until the time of perturbation $t_n^* - A_j$, after which it diverges.

There are two main forms of commonly used perturbation. An *impulsive* perturbation is a kick applied at a single time (which is used in ensemble boosting), in which case $\Omega = \mathbb{R}^k$ or $\mathbb{C}^k$, typically with $k \ll d$, and a sample ω is transformed to spate space via a function $G : \mathbb{R}^k \to \mathbb{R}^d$ (e.g., a low-rank matrix multiplication). Then, the perturbed dynamics can be written $\Phi^{\Delta t}(\mathbf{x}, \omega) = \Phi^{\Delta t}(\mathbf{x} + G(\omega))$, where $\Phi^{\Delta t}$ with only one argument is the unperturbed dynamics. We also use the convention that $G(0) = 0$, i.e., $\omega = 0$ corresponds to no perturbation.

The other common case is where $\mathbf{x}(t)$ is a stochastic process, e.g., an Ito diffusion forced by white noise, as we used in Finkel and O'Gorman (2024) as well as the schematic in Fig. 1. In that case, ω is a white noise process sampled at discrete times, whose dimensionality scales with the number of timesteps. In the QG experiments, we adhere to impulsive perturbations for three reasons: it introduces fewer arbitrary parameters, it is less disruptive to the system's intrinsic dynamics, and it keeps the dimensionality of the random space low. If, as we conjecture, even low-dimensional butterfly flaps are sufficient to excite the more extreme fluctuations, it would make deterministic search methods—which should always be preferred over Monte Carlo—more viable.

Following the perturbation, the descendant drifts away from the parent and achieves its own severity R^* (peak of its intensity function R) at some time $t^*_{n,j,m}$ possibly different from its ancestor's peak time t^*_n:

$$R^*_{n,j,m} = R(\mathbf{x}_{n,j,m}(t^*_{n,j,m})) = R^*_{n,j}(\omega_{n,j,m}) \tag{4}$$

where the latter notation emphasizes dependence on ω, while recognizing that each (n, j) induces a different severity function R^* because perturbations may be felt differently depending on the initial condition.

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. Tersely, $t^* = \text{argmax}_t R(\mathbf{x}(t))$ might be a discontinuous function of ω, and $R^*(\omega)$ a non-differentiable function of ω. This is a nuisance for our goal to optimize over ω, and so 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 $t^*_{n,m,j} = \text{argmax}\{R(\mathbf{x}_{n,j,m}(t)) : t^*_n - \delta t^* \le t \le t^*_n + \delta t^*\}$.
- 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*, rather than *discover totally new events*—which may as well be done by extending the DNS.

2.2 Estimating conditional and climatological probabilities from boosted ensembles

Assume now there is a probability measure $\mathbb{P}^\Omega$ on Ω with associated density function $p^\Omega(\omega)$, which might for example place higher weight on smaller kicks. The Ω superscript will generally relate to statistics over this conditional probability measure, to distinguish it from long-term climatological statistics. A major aim of this paper is to show how they relate to each other. Each ensemble of descendants at each lead time gives rise to its own conditional severity distribution:

$$Q^\Omega_{n,j}(r) = \mathbb{P}^\Omega\{R^*_{n,j} > r\} = \int_\Omega \mathbb{I}\{R^*_{n,j}(\omega) > r\} p^\Omega(\omega)\, d\omega, \tag{5}$$

which can be estimated from the samples $\{R^*_{n,j,m} : m = 1, \ldots, M_{n,j}\}$. Here *conditional* means starting with a perturbation of the nth ancestor's particular initial condition at time $t^*_n - A_j$ and running forward until time $t^*_n + B$. By contrast, we refer to the *climatological* severity distribution as that resulting from a long DNS. Whereas Monte Carlo is the typical strategy in rare event sampling due to an imposed, high-dimensional perturbation space meant to represent extrinsic uncertainty (e.g., wind or waves buffeting an engineered structure; Au and Beck, 2001; Mohamad and Sapsis, 2018), in our setting the perturbation space is an arbitrary design choice aiming at an indirect goal (climate estimation), and nothing stops us here from deliberately choosing low-dimensional perturbations instead of high-dimensional ones as in Ragone et al. (2018); Bloin-Wibe et al. (2025).

This enables numerical quadrature instead of Monte Carlo, and saves on cost by allowing sample re-use across different input distributions. Determining whether this works is one central question this paper aims to answer.

Based on the samples drawn from Ω, we fit a regression model $\widehat{R}^*_{n,j}(\omega;\theta)$ with parameters θ, in our case coefficients for linear and quadratic polynomials. In general $\widehat{R}_{n,j}$ could be a more elaborate parametric model, e.g., a Gaussian process or neural network with learned weights θ, as often used in modern uncertainty quantification (Kabir et al., 2018; Sapsis, 2020; Pickering et al., 2022). Then the integral over Ω can be estimated, either analytically (if p and $\widehat{R}^*$ take simple enough forms) or numerically by densely filling Ω with a grid of points, evaluating $\widehat{R}^*$ and p at each point, and taking their inner product. The result is an estimate $\widehat{Q}^{\Omega}_{n,j}(r)$ for the conditional tail CCDF

$$Q^{\Omega}_{n,j}(r;\mu) = \mathbb{P}^{\Omega}\{R^*_{n,j} > r | R^*_{n,j} > \mu\} = \frac{Q^{\Omega}_{n,j}(r)}{Q^{\Omega}_{n,j}(\mu)}, \tag{6}$$

and can be estimated it by putting hats $\widehat{(\cdot)}$ on every Q. However, this risks dividing by zero, because the fitted function $\widehat{Q}_{n,j}$ may imply zero probability of exceeding the threshold, particularly at long ASTs when descendants have enough time to decorrelate totally with their ancestor. This loss of ancestral "wisdom" is a more fundamental problem than the numerical issue of zero denominator, and we address it by implementing a continuous version of the "accept-reject" step of the TEAMS procedure in Finkel and O'Gorman (2024). Wherever the descendant severity $\widehat{R}^*_{n,j}(\omega)$ falls below μ, we replace it with the ancestor severity, denoted R^*_n (with no second subscript):

$$\widehat{Q}^{\Omega}_{n,j}(r;\mu) := \int_{\Omega} \left\{ \begin{array}{ll} \mathbb{I}\{\widehat{R}^*_{n,j}(\omega) > r\} & \text{if } \widehat{R}^*_{n,j}(\omega) > \mu \\ \mathbb{I}\{R^*_n > r\} & \text{otherwise} \end{array} \right\} p^{\Omega}(\omega)\, d\omega \tag{7}$$

$$= \int_{\{\widehat{R}^*_{n,j}(\omega) > \mu\}} \mathbb{I}\{\widehat{R}^*_{n,j}(\omega) > r\} p^{\Omega}(\omega)\, d\omega + \int_{\{\widehat{R}^*_{n,j}(\omega) \le \mu\}} \mathbb{I}\{R^*_n > r\} p^{\Omega}(\omega)\, d\omega \tag{8}$$

$$= \int_{\Omega} \mathbb{I}\{\widehat{R}^*_{n,j}(\omega) > r\} p^{\Omega}(\omega)\, d\omega + \mathbb{I}\{R^*_n > r\} \int_{\{\widehat{R}^*_{n,j}(\omega) \le \mu\}} p^{\Omega}(\omega)\, d\omega \tag{9}$$

$$= \widehat{Q}^{\Omega}_{n,j}(r) + \mathbb{I}\{R^*_n > r\}\left[1 - \widehat{Q}^{\Omega}_{n,j}(\mu)\right] \tag{10}$$

($\widehat{Q}^{\Omega}_{n,j}(r) = 0$ when $\widehat{Q}^{\Omega}_{n,j}(\mu) = 0$ since $Q^{\Omega}_{n,j}$ is decreasing, hence the two terms in the last expression correspond to the two cases). Another heuristic way to justify this expression is to stipulate that we care about approximating *only the extreme part of the boosting distribution*, i.e., those ωs near enough to 0 that $R^*(\omega) > \mu$, excluding the descendants bound to fall below μ, We re-allocate the probability mass in the "non-extreme" region of the disc (where $R^*(\omega) \le \mu$) to the very center of the disc (the ancestor, where $R^* > \mu$ by construction). This rearrangement ensures that $\widehat{Q}^{\Omega}(\mu)$ is close to 1, justifying a Taylor series

expansion in $1 - \widehat{Q}^{\Omega}(\mu)$

$$Q^{\Omega}_{n,j}(r;\mu) = \frac{Q^{\Omega}_{n,j}(r)}{Q^{\Omega}_{n,j}(\mu)} \tag{11}$$

$$= \frac{Q^{\Omega}_{n,j}(r)}{1 - [1 - Q^{\Omega}_{n,j}(\mu)]} \tag{12}$$

$$\approx Q^{\Omega}_{n,j}(r) + [1 - Q^{\Omega}_{n,j}(\mu)] Q^{\Omega}_{n,j}(r) \tag{13}$$

$$\approx \widehat{Q}^{\Omega}_{n,j}(r) + \left[1 - \widehat{Q}^{\Omega}_{n,j}(\mu)\right] \mathbb{I}\{R^*_n > r\} \tag{14}$$

$$=: \widehat{Q}^{\Omega}_{n,j}(r;\mu) \tag{15}$$

The crux of our hypothesis is that these conditional distributions from boosting can be aggregated across ancestors to approximate the climatological distribution $Q^{\Theta}(r,\mu) = P(R^* > r | R^* > \mu)$, where Θ is used to denote the ground truth that would be obtained from a long DNS. We specifically propose to aggregate the conditional CCDFs as a uniform mixture over ancestors, selecting one representative AST A_{j_n} from each ancestor n to best represent its alternate realities according to some selection rule (different rules will be evaluated thoroughly for the QG system in Sect. 6). We write the mixture as

$$\widehat{Q}^{M}(r;\mu) = \frac{1}{N_{\text{short}}} \sum_{n=1}^{N_{\text{short}}} \widehat{Q}^{\Omega}_{n,j_n}(r;\mu), \tag{16}$$

and call it the "MoCTail" estimator of $Q^{\Theta}(r,\mu)$, for "Mixture of Conditional Tails."

The recent works Noyelle (2024) and Bloin-Wibe et al. (2025) formulate a different estimator, which makes for an interesting comparison. Rather than summing N_{short} tail CCDFs, each approximating a ratio of the form (6), they construct a single ratio by summing N_{short} numerators and N_{short} denominators.Translated into our own notation, this becomes

$$\widehat{Q}^{P}(r;\mu) = \frac{\sum_{n=1}^{N_{\text{short}}} \widehat{Q}^{\Omega}_{n,j_n}(r)}{\sum_{n=1}^{N_{\text{short}}} \widehat{Q}^{\Omega}_{n,j_n}(\mu)}. \tag{17}$$

We call this the "PoPTail" estimator of $Q^{\Theta}(r,\mu)$, for "Pool of Perturbed Tails." Bloin-Wibe et al. (2025) do not model $R^*(\omega)$ parametrically, but instead use a standard Monte Carlo estimate $\widehat{Q}^{\Omega}_{n,j}(r) =$ (fraction of descendants exceeding r), which is probably necessary for their high-dimensional perturbations. However, we can convert the PoPTail estimator to our parametric version just by thinking in terms of CCDFs, hence the formulation in Eq. (17). The more important difference is that PoPTail avoids the potential degeneracy $\widehat{Q}^{\Omega}(\mu) = 0$ by "pooling" non-extreme descendants together with extreme ones in the denominator.

One could argue for either estimator based on the validity of its underlying assumptions which are challenging to rigorously verify. Here we adopt a more openly empirical perspective in testing the skill of both.

An important advantage of both estimators is *extensibility* with respect to the dataset: if the variance is too high, one can always either generate new ancestors by extending the short DNS, or extend the range of ASTs sampled, or enlarge the ensemble at any ASTs deemed promising, without discarding the laborious samples already generated. This is unfortunately not the case with an algorithm like AMS, TEAMS, GKTL, or QDMC: because of the random rules by which ancestors are

selected and new members generated, a completed run cannot be enlarged while retaining its estimation properties unless we are willing to do an entirely new additional run and combine estimates from multiple runs as was done in Ragone et al. (2018); Webber et al. (2019); Finkel and O'Gorman (2024). This results in waste during the fine-tuning process of calibrating TEAMS. For example, one might decide in retrospect that a TEAMS run was too aggressive in killing non-extreme simulations and raising the threshold and we can't easily extend the run with a new set of hyperparameters. With boosting, we can simply go back, perturb those less-extreme simulations, and incorporate them into the dataset, without needing to re-generate everything. To make boosting competitive at sampling the highest levels of severity, we suspect it will be necessary to augment our current scheme with an iterative level-raising schedule, like TEAMS, but with less restriction on the sampling procedure.

2.3 Evaluating performance

We evaluate the MoCTail and PoPTail estimators $\widehat{Q}^M$ and $\widehat{Q}^P$ by comparing to the ground truth Q^Θ as estimated from a long DNS. DNS is in fact a trivial special case of ensemble boosting with $M = 0$ (no descendants), reducing each summand of Eq. (16) and the numerator of Eq. (17) to $\mathbb{I}\{R_n^* > r\}$ and the denominator of Eq. (17) to N_{short}. Both estimators reduce to the same vanilla empirical CCDF in this case, and this is what we use to estimate Q^Θ.

We use χ^2-divergence to measure the disparity of $\widehat{Q}^M$ and $\widehat{Q}^P$ from Q^Θ. This is estimated from a discrete histogram with a sequence of thresholds $\mu = r_0 < r_1 < r_2 < \ldots < r_{K-1} < r_K = \infty$, and define the probability mass function $\Delta Q_k^\Theta = Q_k^\Theta - Q_{k+1}^\Theta$ as the probability contained in the kth bin (note that $Q_K^\Theta = 0$ and so $\Delta Q_{K-1}^\Theta = Q_{K-1}$). As described further in Sec. 3.3, we select the r_ks as quantiles with consecutively halving exceedance probabilities, i.e., $Q_k^\Theta = (\frac{1}{2})^{5+k}$ for $0 \le k < K = 11$. These quantiles change with latitude, as the tail is different for each. Note the same set of r_k's based on the climatological distribution is used also for evaluating estimated distributions. The χ^2-divergence of either estimator $\widehat{Q} \in \{\widehat{Q}^M, \widehat{Q}^p\}$ is then defined as

$$\chi^2(\Delta\widehat{Q} \| \Delta Q^\Theta) = \sum_{k=0}^{K-1} \frac{(\Delta Q_k^\Theta - \Delta\widehat{Q}_k)^2}{\Delta Q_k^\Theta} \tag{18}$$

We will compute both the MoCTail and PoPTail estimates on the same dataset, and find them numerically quite similar, both in terms of skill and in terms of individual bin estimates. It would be interesting to develop test cases where they differ more systematically, to clarify which (if either) is generally superior.

Computational efficiency is another important consideration besides accuracy, as the entire goal of rare event algorithms is to improve efficiency or accuracy (or both) relative to DNS. For a boosting-like rare event algorithm to be useful, its error should decrease faster by perturbing existing ancestors (increasing M) than by extending DNS by generating new ancestors (increasing N and not M), at least in some range of N that samples the attractor broadly but not exhaustively. However, this paper does not present a complete rare event algorithm *per se*, in the sense we don't yet stake our claim on a speedup. Rather, we ask a pre-requisite question: does increasing M decrease the error *at all*? Clearly boosting can increase the maximum severity, but that could happen in ways that don't respect the tail CCDF's shape, e.g., if perturbations tend to maximize the event's severity while bypassing moderate severities that carry significant statistical weight. We will thus make two comparisons between boosting and DNS: accuracy at fixed N, and accuracy at fixed cost (where DNS runs an additional length equal to the cost of simulating

descendants, allocating its full budget to "exploration" rather than "exploitation"). Specifically, we approximate the cost of the boosting approach for a given AST A as

$$\text{Average boosting cost per ancestor} = M(A + \delta t^*) + (\text{mean return period}), \tag{19}$$

where δt^*, the "argmax drift" parameter, accounts for the extra time needed to run after the ancestor's peak to account for changes in peak timing. "Mean return period" is the average time between consecutive independent peaks over the threshold μ, which will be longer than $1/(1-q(\mu))$ because of de-clustering. The dependence on A is a complication, as each AST tried would merit a different-length DNS for cost comparison, and we don't want to penalize boosting too severely by summing over all ASTs because in practice we would not bother simulating the obviously sub-optimal ASTs. Rather, we optimistically estimate the cost if A is already known. On the other hand, our chosen $M(=21)$ is likely more samples than necessary to fit a satisfactory parametric model, as we have deliberately sampled the perturbation space more generously than we would if chasing a speedup. We simplify the comparison by fixing A to $\frac{1}{2}A_{\text{max}}$ in Eq. (19), which is close to or slightly greater than the optimal values that we found empirically.

We will show that boosting is unambiguously more accurate than DNS when fixing the number of ancestors N, and similarly accurate with marginal improvements when fixing cost, though with variation across latitudes and AST criteria. Any fixed-cost performance gains we achieve here (not our main objective) should be viewed as a lower bound for future algorithms, which will benefit from the conceptual insights into AST that we glean presently.

To emphasize the *conditional* nature of the AST—its possible dependence on the ancestor n due to initial condition-dependent predictability—we refer to A_{j_n} as the "conditional advance split time" (CAST), and its optimal value (by χ^2 or other criteria) as the "conditionally optimal advance split time" (COAST). Our goal is to define the COAST, calculate it given extensive sampling from boosted ensembles, and finally to suggest useful criteria to estimate it when sample size is limited, as in a real rare event algorithm deployment.

2.4 AST selection criteria

With a data-generating plan and an estimator in place, we return to our central question of interest: how to select the CASTs $\{A_{j_n}\}$? There are three natural kinds of criteria.

1. Choose a single uniform AST $A_{j_n} = A^{\text{U}}$ for all ancestors (U for "uniform"). In this case, the CAST is not really "conditional" at all. In Finkel and O'Gorman (2024), we found the COAST for TEAMS by systematic grid search through candidate ASTs, and found *post-hoc* an empirical relationship for the COAST: $A^{\text{U}} \approx \overline{t_{3/8}}$, where $t_\epsilon(x_0)$ is the time until an ensemble dispersing from initial condition x_0 (each member forced by a different noise realization) reaches a fraction ϵ of its asymptotic root-mean-squared-error (RMSE), and $\overline{t_\epsilon}$ is the average of $t_\epsilon(x_0)$ over different initial conditions x_0. In Finkel and O'Gorman (2024), we sampled x_0 from the stationary distribution; here, for computational expediency, we will repurpose the boosting ensembles for estimating $\overline{t_\epsilon}$, i.e., sampling x_0 from pre-peak antecedent conditions.

2. Choose the CAST A_n separately for each ancestor n such that that an ensemble launched at $t_n^* - A_n$ disperses to a pre-defined threshold at time t_n^*. One could measure dispersal in different ways like RMSE, but here we opt instead for a *pattern correlation*, defined with respect to spatiotemporal fields F_0 (from the ancestor) and F_m (from the mth ensemble member) as

$$\rho[F_0, F_m] := \frac{\overline{f_0 f_m}}{\sqrt{(\overline{f_0^2})(\overline{f_m^2})}} \text{ where } f := F - \langle F \rangle, \ \langle \cdot \rangle = \text{ time-average (climatology), and } \overline{(\cdot)} = \text{ space-average.} \tag{20}$$

Unless noted otherwise, ρ will refer to the average of $\rho[F_0, F_m]$ over all members $m = 1, \ldots, M$. The reason for subtracting time-averages is to fairly weight spatial regions with smaller background $\langle F \rangle$, e.g., poles if F is temperature. Dividing by spatial standard deviations is simply a useful normalization that restricts ρ to the range $[-1, 1]$ by the Cauchy-Schwarz inequality. ρ tends to decrease over time from 1 to 0 except for occasional negative values when F_0 and F_1 are similar up to translation (but this effect usually disappears when averaging large-enough ensembles). We then choose some threshold $\rho^{\mathrm{U}} \in (0, 1)$, and select the corresponding CAST $A_{j_n} = A_n^{\mathrm{PC}}[\rho^{\mathrm{U}}]$—a function of the threshold—as the smallest sampled AST A_n for which ρ decreases from 1 to ρ^{PC} between the split time $t_n^* - A_n$ and the peak time t_n^*. (PC stands for for "pattern correlation"). Note that the CAST varies with n, but the correlation threshold, denoted ρ^{U}, is uniform. Finding the COASTs A_n^{PC} then boils down to finding the optimal value of ρ^{U}.

The $\frac{3}{8}$ rule from Finkel and O'Gorman (2024), which used Euclidean distance $D^2[F_0, F_m] = \overline{(F_0 - F_m)^2} = \overline{(f_0 - f_m)^2}$ as the dispersion indicator, can be approximately restated in terms of pattern correlation:

$$D^2 = \epsilon^2 \langle D^2 \rangle \qquad \langle D^2 \rangle = \text{saturation value of } D^2 \tag{21}$$

$$\implies \overline{f_0^2} + \overline{f_m^2} - 2\overline{f_0 f_m} = \epsilon^2 (\langle \overline{f_0^2} \rangle + \langle \overline{f_m^2} \rangle) \qquad \text{Using } \langle \overline{f_0 f_m} \rangle = \langle \overline{f_0} \rangle \langle \overline{f_m} \rangle = 0 \tag{22}$$

$$\frac{(\overline{f_0^2} - \epsilon^2 \langle \overline{f_0^2} \rangle) + (\overline{f_m^2} - \epsilon^2 \langle \overline{f_m^2} \rangle)}{\sqrt{(\overline{f_0^2})(\overline{f_m^2})}} = \frac{2\overline{f_0 f_m}}{\sqrt{(\overline{f_0^2})(\overline{f_m^2})}} = 2\rho(F_0, F_m) \tag{23}$$

$$\frac{(1 - \epsilon^2)\langle \overline{f_0^2} \rangle + (1 - \epsilon^2)\langle \overline{f_m^2} \rangle}{\sqrt{\langle \overline{f_0^2} \rangle \langle \overline{f_m^2} \rangle}} \approx 2\rho(F_0, F_m) \qquad \text{Approximating } \overline{f^2} \approx \langle \overline{f^2} \rangle \tag{24}$$

$$1 - \epsilon^2 \approx \rho(F_0, F_m) \qquad \text{Using } \langle \overline{f_0^2} \rangle = \langle \overline{f_m^2} \rangle. \tag{25}$$

(The approximation invoked in the second-to-last step, $\overline{f^2} \approx \langle \overline{f^2} \rangle$, will hold when the spatial region is large enough that global fluctuations in the same direction are unlikely.) This calculation shows that the time until RMSE reaches $\frac{3}{8}$ of its saturation value is roughly equivalent to the time at which pattern correlation drops to $1 - (\frac{3}{8})^2 = 0.86$. We do not assume this threshold is optimal, but include it as a reference for comparison. And we stress that the $\frac{3}{8}$ rule implemented in Finkel and O'Gorman (2024) determines a uniform A^{U}, not a conditional A^{PC}, because their averaging was performed over the attractor, whereas here we will use ρ as an initial condition-specific diagnostic.

3. Define the CAST as the solution to an optimization problem, where we seek to optimize a functional on the boosted severity distribution that favors both a high mean and high variability of the severity. This would implicitly favor intermediate ASTs, as short-AST ensembles have high mean but low variability while long-AST ensembles will have high

variability but low mean (approaching the climatological distribution). We propose and evaluate two such functionals in this paper:

(a) Expected improvement (EI):

$$\mathbb{E}[(\Delta R^*)_+] = \int_\Omega p^\Omega(\omega)[R^*(\omega) - R^*(0)]_+ \, d\omega, \tag{26}$$

where $(\cdot)_+ := \max(\cdot, 0)$ and we recall that $\omega = 0$ means no perturbation (i.e., the ancestor)

(b) Thresholded entropy (TE):

$$S[(R^* - \mu)_+] = -\sum_{k=0}^{K-1} \Delta Q_k \log \Delta Q_k, \tag{27}$$

where the bin boundaries r_k start at μ, and so only the tail part of the conditional CCDF contributes. The thresholded entropy is thus defined based on probability over discrete bins (with the bin boundaries r_k set based on quantiles of the ground-truth distribution) and would change if the bins were changed.

Where it doesn't cause confusion, we will also call the CASTs A^{EI} and A^{TE} themselves COASTs because they optimize something, although it is something different than χ^2. We have conjectured that that these two notions of optimality coincide: if each ancestor separately optimizes EI or TE, the resulting aggregate of distributions (via MoCTail or PoPTail estimators) will minimize χ^2-divergence from the true climatological tail. Our results will approximately confirm the conjecture in the case of TE.

These criteria are each in turn more complex, but also more theoretically appealing. The correlation-based CASTs $\{A_n^{\text{PC}}\}_{n=1}^{N_{\text{short}}}$, unlike the synchronized AST A^{U}, can vary with n to respect differences in predictability between different initial conditions, a well-recognized phenomenon in chaotic systems (Maiocchi et al., 2024), including the atmosphere (Lucarini and Gritsun, 2020). Still, both A^{U} and A_n^{PC} require the user to set some arbitrary global threshold. The open question is whether optimizing A_n^{EI} or A_n^{TE} individually for each n will also optimize the accuracy of the unconditional (MoCTail) climatological CCDF estimator against the ground truth climatological CCDF from a long DNS.

Main result: Climatological tails are estimated more accurately with perturbed ensembles than with un-perturbed ancestors alone (fixed-N comparison between DNS and boosting). This holds with few exceptions for both MoCTail and PoPTail estimators, for all COAST selection rules, and for all target spatial locations. At fixed cost, boosting and DNS are tied overall, but with some variation across latitudes and the value that cost is fixed to, suggesting that substantial speedups are possible with more highly optimized boosting-like algorithms. No single selection rule is superior across the board. The EI and TE criteria, however, have a distinct advantage of needing no arbitrary threshold choices. TE-based estimates strike a reasonable compromise between statistical error and arbitrariness, which is strong enough support that **we recommend TE as a generic AST selection rule**.

Table 1. Three rungs on the model hierarchy. Left: the Lorenz-96 system used in Finkel and O'Gorman (2024) has a one-dimensional spatial domain ("longitude") divided into discrete sites $k = 0, \ldots, 39$, on which generic meteorological variables $\{x_k\}$ evolve in time. Its state space dimension is 40. Right: the aquaplanet model used in Finkel and O'Gorman (2025) has a three-dimensional spatial domain: latitude λ, longitude ϕ, and pressure normalized by its surface value, $\sigma = p/p_s$. It has six prognostic fields: zonal wind u, meridioal wind v, temperature T, and humidity q vary in all three dimensions, whereas surface pressure p_s and precipitation rate R vary only in the horizontal. Center: the 2-layer quasigeostrophic model used in this study has two layers ($z = 1, 2$) of two dimensions each (longitude x, latitude y), and two dynamical fields: streamfunction ψ which is discretized spectrally, and tracer concentration c which is discretized on a grid.

Model	One-tier Lorenz-96	2-layer quasigeostrophic channel	Global aquaplanet
Domain	$k \in \{0, \ldots, 39\}$	$(x, y, z) \in [0, L)^2 \times \{1, 2\}$	$(\lambda, \phi, \sigma) \in [0, 360) \times [-90, 90) \times [0, 1)$
Fields	$\{x_k\}$	$\{\psi_z, c_z\}(x, y)$	$\{u, v, T, q\}(\lambda, \phi, \sigma) \cup \{p_s, R\}(\lambda, \phi)$

The remainder of the paper demonstrates the theoretical framework above on the QG system. Sect. 3 specifies the dynamical model and its numerical simulation, displays some representative output, defines the target intensity functions of interest, and reports on their basic tail statistics. Sect. 4 specifies the perturbation protocol (i.e., the space Ω and probability densities $p^{\Omega}(\omega)$) and visualizes representative examples of the system's response, providing motivation for our choices of AST selection criteria. Sect. 6 compares the performances of all proposed AST selection criteria criteria in matching the climatological tail CCDF. Sect. 7 concludes with a summary and outlook on important future lines of work.

Throughout, we present more in-depth results for one select target latitudes just south of the domain center, and only summarize for the wider range of target latitudes, which reveals large-scale variations in extreme event predictability and representability across space.

3 The quasigeostrophic model

The model setup aims to distill some challenges we have encountered with rare event algorithms across the hierarchy. We first recognized the need for advance splitting (or "trying early") to sample extreme precipitation in an aquaplanet GCM (Finkel and O'Gorman, 2025). A minimal surrogate model replicating this challenge was found in Lorenz-96 Lorenz and Emanuel (1998), which provided a testbed for the first working version of TEAMS and a recognition of an "optimal advance split time" (Finkel and O'Gorman, 2024). There is a huge gap in model complexity between Lorenz-96 and the GCM (see Table 1), and we wish to test our idea in this middle ground where the target spatial location can have an effect. Lorenz-96, with a one-dimensional domain and homogeneous forcing, is too simple. For this reason, and to make closer contact with physics, we selected the two-layer QG model as a suitable intermediate between Lorenz-96 and the GCM.

3.1 Equations of motion and numerical simulation

We implement a version of the QG model combining elements of several classic studies. Our numerical method and friction form follow Haidvogel and Held (1980), but on a smaller domain with weaker friction magnitude as in Panetta (1993) to contain only 1-2 more energetic zonal jets. We furthermore add bottom topography in the lower layer as in Thompson (2010) to fix preferred latitudes for jets while still allowing them to temporarily split, merge, and meander. Thus climate statistics, and hence the COAST itself, can vary with latitude. Further, we augment the system with a passive tracer to represent a key component of precipitation dynamics, following the spirit of Bourlioux and Majda (2002) and Qi and Majda (2016, 2018) who used turbulent advection-diffusion as a paradigm for intermittency.

The model equations are as follows, in non-dimensionalized form using the deformation radius λ as the length scale and a velocity scale $\mathcal{U}$. To make plain the role of the background shear, we define a non-dimensional wind U as the ratio between the imposed upper-level zonal wind and $\mathcal{U}$. All non-dimensional parameter values are listed in Table 2. The horizontal coordinates (x, y) each run from 0 to L. The integer-valued vertical coordinate z is an index for the layer (1 for the top and 2 for the bottom). ψ represents the streamfunction minus a background of $-Uy\delta_{z,1}$. h is the bottom topography which is specified to vary sinusoidally with wavenumber 2 in latitude. q represents potential vorticity minus a background of $\beta y + h\delta_{z,2}$, due to planetary vorticity gradient and topography. c represents the passive tracer field.

$$\Big[\partial_t + (\partial_x\psi)\partial_y + (U\delta_{z,1} - \partial_y\psi)\partial_x\Big](q + h\delta_{z,2} + \beta y) = -\kappa\delta_{z,2}\nabla^2\psi - \nu\nabla^6\psi \tag{28}$$

$$\Big[\partial_t + (\partial_x\psi)\partial_y + (U\delta_{z,1} - \partial_y\psi)\partial_x\Big]c = 0 \tag{29}$$

$$\text{for } (x, y, z) \in [0, L)^2 \times \{1, 2\} \tag{30}$$

$$\text{where} \tag{31}$$

$$q_z = \nabla^2\psi_z + (-1)^z\left(\frac{\psi_1 - \psi_2}{2}\right) \tag{32}$$

$$h(y) = h_0 \sin\left(2 \cdot \frac{2\pi y}{L}\right) \tag{33}$$

For ψ, we impose doubly periodic boundary conditions and timestep with a pseudo-spectral method with 64 Fourier modes in each dimension and standard $\frac{2}{3}$-dealiasing (hence, an effective maximum wavenumber of 20). We time-step linear terms with the trapezoid rule (Crank-Nicolson) and nonlinear and topographic terms with a predictor-corrector (Heun's) method. Meanwhile, boundary conditions on c are periodic in x and Dirichlet in y, with values $(0, 1)$ at $y = (0, L)$. Together with a first-order upwind monotone finite-volume scheme, this setup guarantees that $0 \le c \le 1$ everywhere, making clear that its probability distribution has compact support. Note there is no explicit dissipation for c, but the low-order discretization creates some effective diffusivity.

The number of degrees of freedom, or state space dimension, is

$$d = (2 \text{ layers}) \times (41^2 \text{ Fourier modes for } \psi + 64^2 \text{ grid cells for } c) = 11554, \tag{34}$$

Description	Symbol	Value
Coriolis gradient	β	0.25
Ekman friction coefficient	κ	0.05
Wind shear	U	1
Hyperviscosity	ν	$(0.292)^3$
Topography amplitude	h_0	0.25
Domain size	L	$6 \cdot 2\pi$

Table 2. Non-dimensional physical parameters used for the numerical simulation, similar to those chosen in Panetta (1993).

and we will sometimes refer to the full state vector as $\{\psi, c\}(x, y, z, t) = \mathbf{x}(t) \in \mathbb{R}^d$—not to be confused with the spatial coordinate x. For simplicity, we refer to one time unit as a day, which is $\sim \frac{1}{10}$ of an eddy turnover timescale (see Fig. 3). The common timestep for ψ and c is 0.025 days, and the output frequency is once per day. The spatiotemporal resolution is coarse by modern standards, but we aren't seeking to calculate any real-world physical quantity: we are seeking a general rule that will help make the COAST clear for a wide class of models.

3.2 Baseline simulation and statistics

We run a "short DNS" of length $T_{\text{short}} = 4 \times 10^3$ days ≈ 11 years (after a 500-day spinup) to supply the pool of initially un-perturbed ("ancestral") events. Then, to provide "ground truth" statistics, we run a control simulation, or "long DNS", of duration $T_{\text{long}} = 16 \times 10^3$ days ≈ 44 years, which is $O(1600)$ eddy turnover times and $O(160)$ jet meandering times (see Fig. 3 caption for timescale definitions). However, in estimating climatological statistics from the long DNS, we take advantage of statistical zonal symmetry by concatenating the timeseries of all 64 longitudes, increasing the effective sample size by a factor of $\sim L/$(some typical correlation length). Conceptually, the short and long DNS are analogous to "training" and "validation" datasets in standard machine learning procedures, in the sense that we want to infer properties of the validation set using only information extracted from the training set (for example, by perturbing and re-simulating events seen in training). As we show below, simply counting events from the short DNS gives probability estimates that deterioriate at high levels of severity, which we aim to rectify with boosting.

Fig. 2 shows representative snapshots of three dynamical fields in the upper layer from the long DNS: tracer concentration c, zonal velocity $u = U - \partial_y \psi$, and meridional velocity $v = \partial_x \psi$. Fig. 3 shows Hovmöller diagrams of zonal-mean anomalies of c and u (not v, since zonal-mean meridional velocity is zero), as well as their climatological means and standard deviations plotted alongside the topography. These are statistics of the grid-cell values, not zonal means, but depend only on latitude because so does topography. Two eastward jets are prominent in the snapshots Fig. 2(b) and in the zonal mean profile Fig. 3b.iii, with preferred latitudes of $\sim \frac{1}{4}L$ and $\sim \frac{3}{4}L$. The Hovmöller diagram gives a sense of characteristic timescales: jets tend to remain roughly stationary for stretches of ~ 100 days at a time before shifting, as seen by the group of closed contours of ψ and associated dipole of u centered at time $t = 3400$. and persisting ± 50 days to either side. Within these stretches of quasi-stationarity, there are shorter undulations of duration ~ 10, which we identify as the eddy turnover timescale.

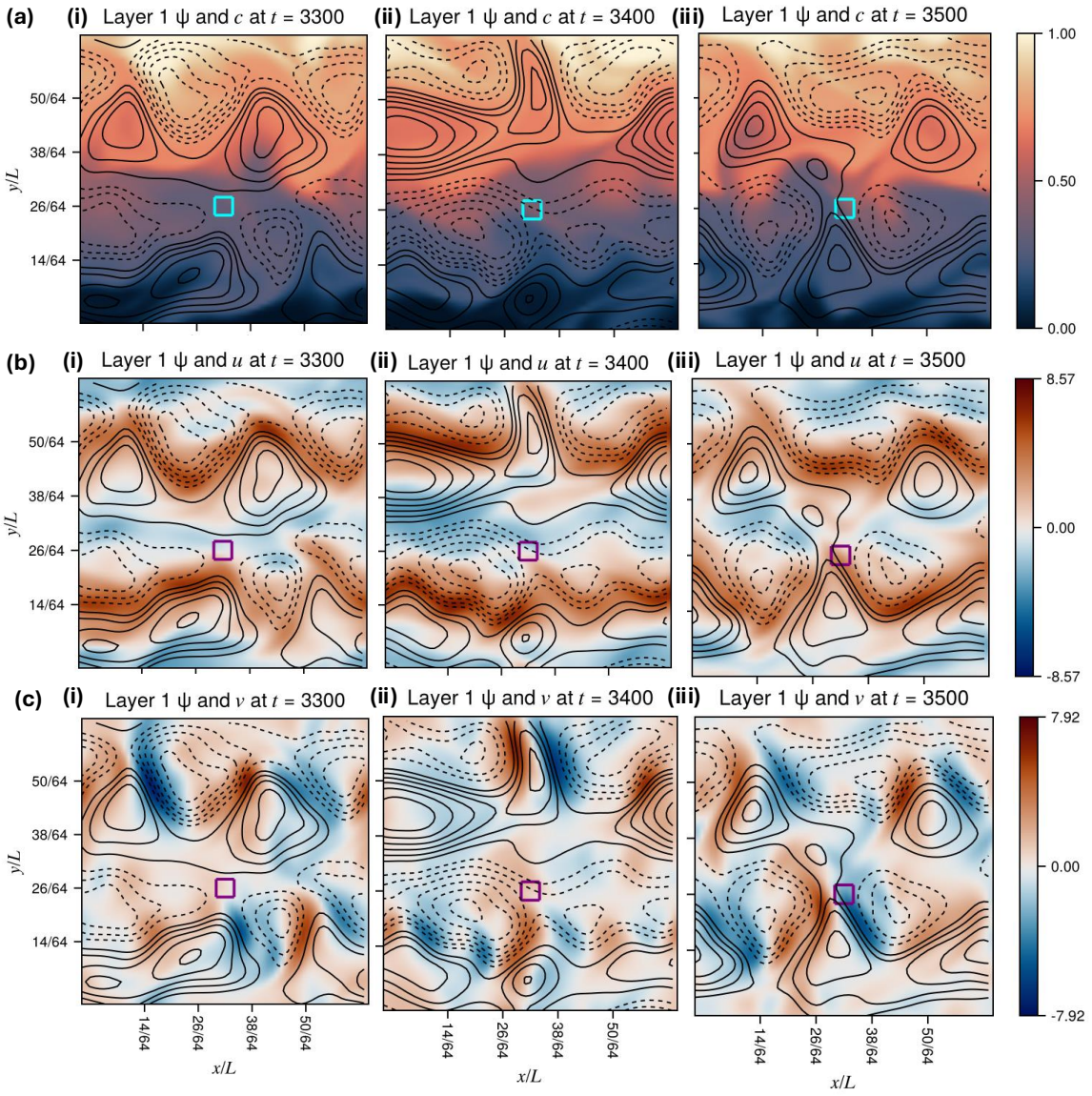

Figure 2. Snapshots of the QG system configuration in the upper layer. Contours indicate the anomaly streamfunction ψ, which varies over a non-dimensional range of approximately ± 18, dashed contours indicating negative anomalies. Colors indicate (a) tracer concentration c, (b) zonal wind velocity $u = U - \partial_y \psi$, where $U = 1$ is the basic background shear, and (c) meridional velocity $v = \partial_x \psi$. The timestamps increase from left to right, and come from the long DNS. The small square represents an example target region in which to sample extremes of the local tracer concentration, in this case centered at $x_0 = \frac{1}{2}L, y_0 = \frac{26}{64}L$ and extending $\pm\ell = \frac{2}{64}L$ in both meridional and zonal directions. This same region is the target used in the following results, and we consistently refer to the domain coordinates in fractions of 64 across all figures.

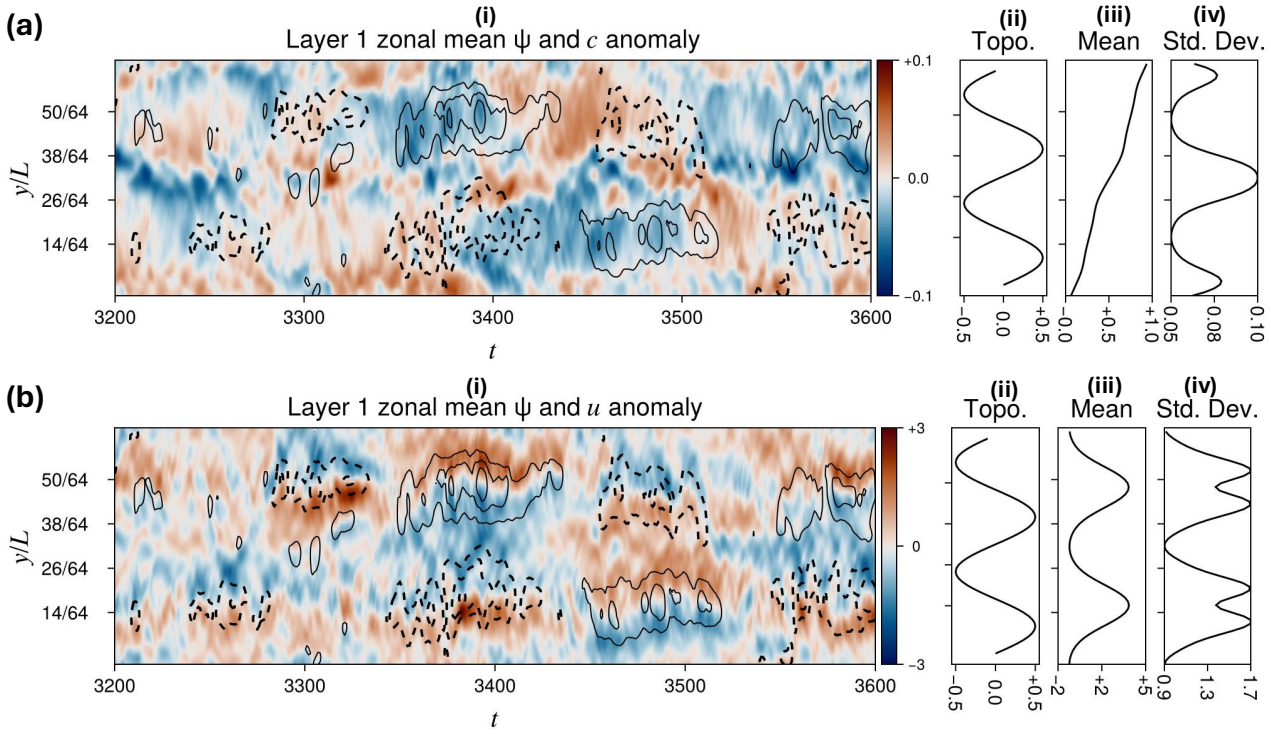

Figure 3. Hovmöller diagrams of anomalies (departures from time-means) of zonal-mean concentration (a.i) and zonal-mean zonal wind (b.i). Contours indicate zonal-mean streamfunction anomaly (range ± 10, negatives values dashed). Column (ii) shows bottom topography, which *directly* affects the lower layer only, but indirectly sets the preferred jet positions in the upper layer as well. For the same quantities, column (iii) shows the zonal and time mean and column (iv) shows the zonal mean of the temporal standard deviation. The Hovmöller diagrams give context to the snapshots of u from Fig. 2b, which come from times (i) 3300, when the upper and lower jets are both shifted south; (ii) 3400, when the jets are unusually far apart; and (iii) 3500, when the jets are unusually close together. These intermittent, discrete shifts in jet location happen every ~ 100 days, which we call the "jet meandering timescale". During a typical 100-day timespan of stationary jet, the fields shown oscillate roughly 10 times; hence we assign the eddy turnover timescale a nominal value of 10 days.

The tracer statistics (Fig. 3a.(iii,iv)) have some easily explainable large-scale patterns and some subtler small-scale patterns. The tracer time-mean $\langle c \rangle(y)$ increases linearly overall as $\frac{y}{L}$, in keeping with its Dirichlet boundary conditions. However, in the central region of the domain (inside the weak westward jet) the tracer mean varies more rapidly with latitude and has a larger standard deviation (see also dashed curves in Fig. 4b,c). In the eastward jets, the tracer mean varies more slowly with latitude and has a smaller standard deviation. Comparison with the Hovmöller diagram (Fig. 3a.i) suggests that the central region owes its high variance to short-lived anomalous pulses, both positive and negative, which are more intense than in surrounding regions. We won't try to explain these patterns from first principles, but simply state that the setup accomplishes our intention to provide a variety of statistical behaviors as a suite of test cases for our approach.

3.3 Target variable

We define the intensity function of interest $R(\mathbf{x})$ as the upper-level concentration, c_1 (henceforth, simply c), averaged over a small square box $[x_0-\ell, x_0+\ell] \times [y_0-\ell, y_0+\ell]$ of half-width $\ell = \frac{2}{64}$, and 23 evenly spaced latitudes $y_0 \in \left\{\frac{10}{64}, \frac{12}{64}, \ldots, \frac{54}{64}\right\} L$, restricted to the central region to avoid boundary effects. The central longitude x_0 is fixed to $L/2$, but by zonal homogeneity any longitude would be statistically equivalent. We also repeated the analysis with double the box length, and found results to be qualitatively similar, so we only show results for the smaller box size. The effect of spatial scale is worth considering in its own right with a wider range, which we postpone to future work.

Fig. 4 displays some summary statistics of $R(\mathbf{x}(t))$ as functions of the target latitude y_0: alongside (a) the topography for reference, we show (b) the meridionally de-trended time-mean $\langle R \rangle(y_0) - \frac{y_0}{L}$ and (c) the standard deviation $\sqrt{\langle R^2 \rangle(y_0) - \langle R \rangle^2(y_0)}$. Note the restricted latitude range. In (a) and (b), dashed lines show the corresponding mean and standard deviation of c itself, as in Fig. 3(c,d), of which R is a regional average: note that R has the same mean as c but a smaller standard deviation, and larger box sizes would reduce it even further.

While the low-order moments capture ordinary behavior of intensities R, the intensity peaks—i.e., severities R^*, defined in Sect. 2—are better viewed from an extreme value theory perspective, and summarized with the peaks-over-threshold procedure (Coles, 2001). We set a threshold μ as the $(\frac{1}{2})^5$th complementary quantile of R, also denoted $\mu[(\frac{1}{2})^5]$, i.e., the level whose exceedance probability is $q(\mu) = (\frac{1}{2})^5$. Severities R^* are extracted as cluster maxima above μ, with buffer times $A_{\max} = 40$ days and $B = 20$ days. All cluster maxima from the long DNS are used as input data points to infer the parameters (scale σ, shape ξ) of a generalized Pareto distribution (GPD), using the maximum-likelihood routine of the `Extremes.jl` package (Jalbert et al., 2024):

$$\mathbb{P}\{R^* > r | R^* > \mu\} \approx G_\mu(r; \sigma, \xi) = \begin{cases} \left[1 + \xi\left(\frac{r-\mu}{\sigma}\right)\right]_+^{-1/\xi} & \xi \neq 0 \\ \exp\left[-\left(\frac{r-\mu}{\sigma}\right)_+\right] & \xi = 0 \end{cases} \tag{35}$$

where $(\cdot)_+ = \max(\cdot, 0)$. Fig. 4(d,e,f) display the threshold (detrended by $\frac{y_0}{L}$), scale parameter σ, and shape parameter ξ. Several characteristics are noteworthy.

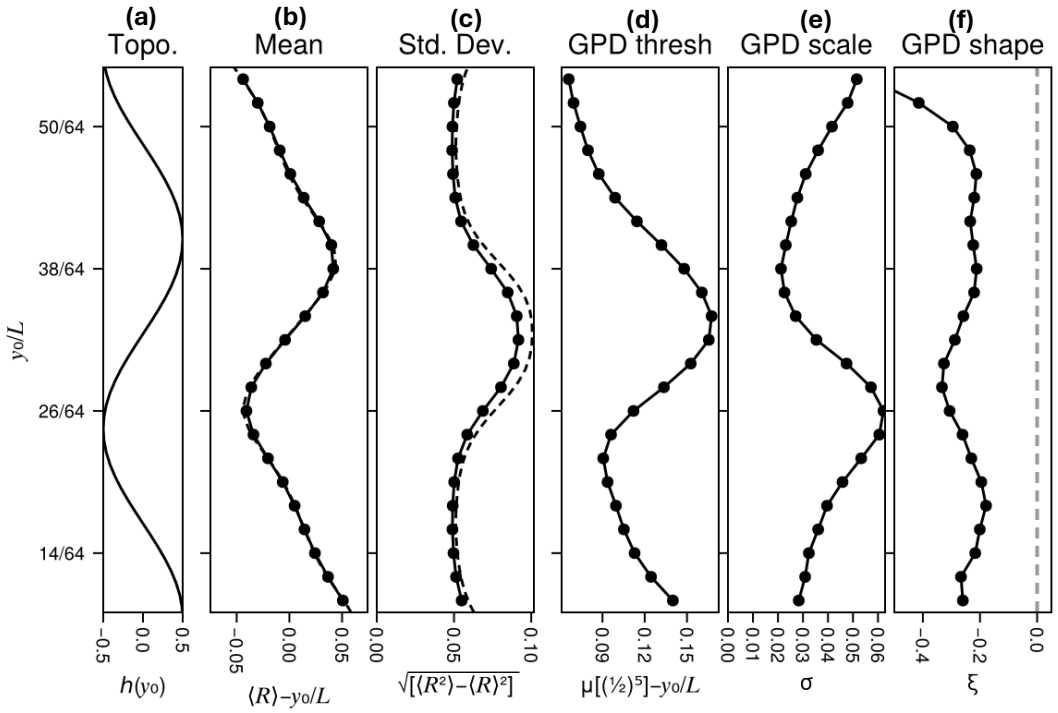

Figure 4. Summary statistics of latitude-dependent climatological tail distributions of local tracer concentrations, also called "intensities", which are denoted R and defined as the average concentration c over a box $(x, y) \in (x_0, y_0) + [-\ell, \ell]^2$. $x_0 = \frac{1}{2}L$ and $\ell = \frac{1}{32}L$ are fixed, while y_0 varies across the midlatitudes from $\frac{10}{64}L$ to $\frac{54}{64}L$. Panel (a) shows the lower-layer topography in this same range of middle latitudes, (b) shows the mean intensity $\langle R \rangle(y_0)$, after subtracting a nominal trend of $\frac{y_0}{L}$ to reveal a finer-scale structure that resembles the underlying topography, and (c) shows the standard deviation of intensity $\sqrt{\langle R^2 \rangle - \langle R \rangle^2}$. Dashed curves in (b) and (c) indicate the mean and standard deviation, respectively, of the concentration field c without box-averaging. Panels (d,e,f) summarize the distribution of intensities R^* via the parameters of the generalized Pareto distributions (GPD), inferred by the peaks-over-threshold fitting procedure (see section 3.3 for details). The threshold is set to the $(\frac{1}{2})^5$-complementary quantile, denoted $\mu[(\frac{1}{2})^5]$ and shown in (d) with linear trend removed. Panels (e, f) display the estimated (scale, shape) parameters (σ, ξ).

- The detrended threshold $\mu - \frac{y_0}{L}$ has a maximum-over-minimum profile similar to the the detrended mean intensity $\langle R \rangle - \frac{y_0}{L}$, but shifted southward. The maximum of $\mu - \frac{y_0}{L}$ is close to the mid-channel maximum in the standard deviation of R, perhaps because extremes depend more on variability than on average behavior.

- As expected for an upper bounded tail, we find $\xi < 0$ (Fig. 4f).

- The GPD scale parameter, σ, is anti-correlated with the (detrended) mean. The constraint $R^* \leq 1$ can explain this, as a higher distribution center leaves less room for an expansive tail. In addition, the threshold μ tracks approximately with the mean, and we can understand the anticorrelation mathematically through the non-uniqueness of GPD parameters: the same tail can be adequately described by two different choices of threshold (μ_1, μ_2), and the two corresponding scale parameters will be related by $\sigma_2 - \sigma_1 = \xi(\mu_2 - \mu_1)$. Only the shape parameter, ξ, is invariant with respect to μ. Seeing that $\xi < 0$ varies only slightly with latitude, σ and μ would vary inversely even if the tail itself were not changing.

We implemented the boosting and estimation procedures for every latitude separately, but for illustration focus the in-depth analysis on $y_0 = \frac{26}{64}L$ (the small boxes in Fig. 2), an interesting location where the (detrended) mean and threshold $\mu[(\frac{1}{2})^5]$ are both low, the GPD scale σ is large, and the GPD shape slightly more negative than in surrounding regions. Fig. 5 displays the underlying probability distributions at $y_0 = \frac{26}{64}L$ to show the nature of the tails of the distributions and also to help clarify the relationship between intensities, severities, and GPD parameters. The full PDF of intensity, in (a), has a positive skew and sub-Gaussian tail. Black and red solid curves are estimates obtained from the long and short DNS, respectively, and 90% confidence intervals are obtained by longitudinal translation. Specifically, the shaded intervals are the 5th-95th percentile ranges of intensities at the same y_0, but with x_0 shifted from its base location of $\frac{1}{2}L$ by $\frac{0}{64}L, \frac{1}{64}L, \frac{2}{64}L, \ldots, \frac{63}{64}L$. The dashed black curve is the mean of all 64 curves, our best available estimate of ground truth. The discrepancy between short and long DNS is most pronounced in the upper tail, which in panel (b) is magnified and integrated from the top, giving the CCDF. Gray lines mark the threshold, $\mu = 0.52$, and its CCDF value $\frac{1}{32} \approx 0.03$. Above this level, the short DNS becomes rapidly more uncertain (error bar widens), and severely underestimates probabilities smaller than ~ 0.005.

Both short and long DNS estimates diverge markedly from the GPD fit shown in gray in panel (b). This is where the distinction between intensity and severity comes into play: the GPD is fitted to *peaks over the threshold* μ—i.e., severities—whose distribution differs (most notably in the upward direction) from that of *all* exceedances over μ, which would include the clusters surrounding the peaks. Panel (c) confirms that the GPD fit is much more appropriate for severities R^* than for intensities R, and thereby clarifies the distinction. If the threshold were raised, the clusters would shrink, the sequence of peaks would form a Poisson process, and the CCDFs of R and R^* would converge. For computational economy and because non-asymptotic extremes are of interest for climate risk, we keep the threshold at $\mu[(\frac{1}{2})^5]$ and formally define our goal with boosting as correcting the distribution of severities—not intensities. Hence, our measure of success will be whether the short-DNS severity CCDF in Fig. 5c, when augmented by boosting, will become closer to the long-DNS severity CCDF.

4 Ensemble design

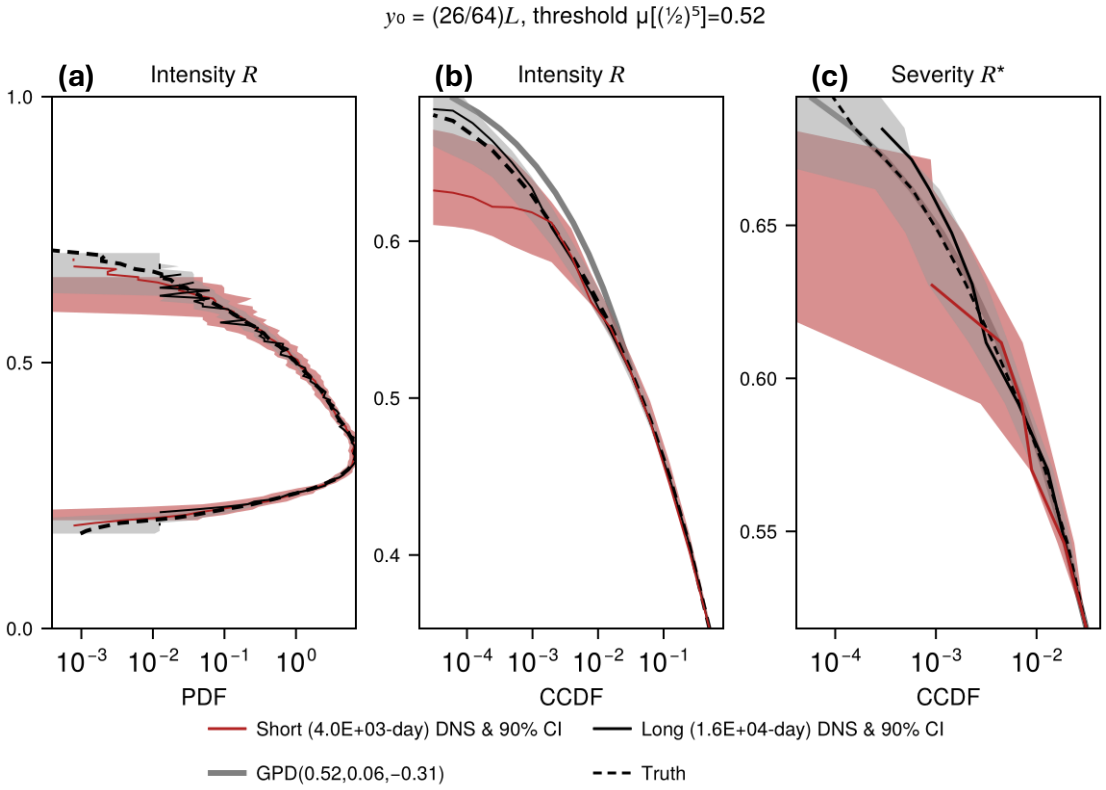

Figure 5. Probability distributions of local tracer concentrations at latitude $y_0 = \frac{26}{64}L$ and averaged over a box of half-width $\ell = \frac{2}{64}L$. (a) The full PDF of intensity R. (b) The CCDF (tail integral) of intensity R, restricted to $R > \mu[\frac{1}{2}]$. (c) Further zoomed-in CCDF of the severity R^* (peaks of R over $\mu[(\frac{1}{2})^5]$). In all three panels, solid black and red lines represent estimates from long and short DNS, respectively, with shaded 90% confidence intervals obtained by repeating the inference 64 times, once for each possible longitudinal rotation of the dataset. Error bars become degenerate at levels experienced by $< 5\%$ of longitudes. Black dashed lines show the mean over all longitudinal rotations, our best estimate of ground truth. The gray line in (b,c) represents the GPD fit to R^* with $\mu = 0.52$, $\sigma = 0.06$, and $\xi = -0.31$, and this is a much better fit to the severities in (c) which makes sense given they are defined in terms of peaks.

4.1 Stochastic inputs

We perturb the QG model with impulsive forcing, which we now specify as a concrete instantiation of the generic form in Sect. 2. The stochastic input ω lives in the complex plane $\mathbb{C}(=\Omega$, the "input space"), and the state-space perturbation $G(\omega)$ consists of a single Fourier mode to be added to ψ. We choose the mode based on linear stability analysis, which is more easily explained as a procedure than as a closed formula:

1. Decompose ψ into a Fourier basis $\psi_z(x,y) = \sum_{k,\ell} \widehat{\psi}_z(k,\ell) e^{i(kx+\ell y)}$, and write the linearized dynamics (about the baroclinically unstable background state with vertical zonal wind shear and $\psi = 0$, and neglecting topography) into the abstract form

$$C(k,\ell)\frac{d}{dt}\begin{bmatrix}\widehat{\psi}_1(k,\ell)\\ \widehat{\psi}_2(k,\ell)\end{bmatrix} = D(k,\ell)\begin{bmatrix}\widehat{\psi}_1(k,\ell)\\ \widehat{\psi}_2(k,\ell)\end{bmatrix} \tag{36}$$

   where $C \in \mathbb{C}^{2\times 2}$ represents the conversion from streamfunction to potential vorticity, and $D \in \mathbb{C}^{2\times 2}$ represents the advection and linear dissipation terms (excluding topography).

2. Calculate the eigenvalues and eigenvectors $\{(\lambda^{(m)}(k,\ell), \widehat{\varphi}^{(m)}(k,\ell)) : m = 1,2\}$ of the Jacobian matrix $C^{-1}(k,\ell)D(k,\ell)$, ordered by stability: $\mathrm{Re}\{\lambda^{(1)}\} \geq \mathrm{Re}\{\lambda^{(2)}\}$, and select $(k^*,\ell^*) = \mathrm{argmax}_{k,\ell}\{\mathrm{Re}\{\lambda^{(1)}(k,\ell)\}$, i.e., the linearly most unstable mode from the background state. Restrict the optimization to (k,ℓ) both nonnegative, and not both zero.

3. For $z = 1,2$, increment $\widehat{\psi}_z(k^*,\ell^*)$ by $\omega\widehat{\varphi}_z^{(1)}(k^*,\ell^*)$, and to maintain the solution's reality add the complex conjugate (c.c.) to $\widehat{\psi}_z(-k^*,-\ell^*)$. The perturbation can be written as a function of space,

$$G(\omega) = \delta\psi_z(x,y) = \omega\widehat{\varphi}_z^{(1)}(k^*,\ell^*)e^{i(k^*x+\ell^*y)} + \text{c.c.}, \tag{37}$$

   which can have pointwise magnitudes up to $2|\omega|$. In the QG model, the mode we identify is $(k^*,\ell^*) = (4,0)$, and $G(\omega)$ is plotted in Fig. 6c for three different example ωs, which correspond to the points labeled 1,2,3 in panel (a). All share the same inter-layer *relative* phase and magnitude, as these are properties of k^*,ℓ^*, and $\widehat{\varphi}_z^{(1)}(k^*,\ell^*)$, but differ in *absolute* phase and magnitude. Note that points 2 and 3 are approximately diametrically opposed, and hence spatially $\sim 180^\circ$ out of phase, whereas point 1 is approximately one-quarter revolution away and spatially $\sim 90^\circ$ out of phase with both 2 and 3. Points (2, 3) are (closest to, farthest from) the center of the circle, and hence have the (smallest, largest)-magnitude spatial perturbations.

The steps above completely specify $G(\omega)$, a linear map from $\mathbb{C}$ to functions of (x,y,z), which can be easily computed offline before running any ensembles. One could argue for two obvious refinements of this choice: (1) specializing the linearization to the actual initial state, not just the background state, by linearizing the quadratic form $J(q,\psi)$ and including that in the calculation of $D(k,\ell)$; and (2) accounting for finite time horizons by using the leading singular vector of the *linear propagator*, i.e., the initial *infinitesimal* error whose magnitude amplifies the most over a given time horizon (Farrell and Ioannou, 1996a, b).

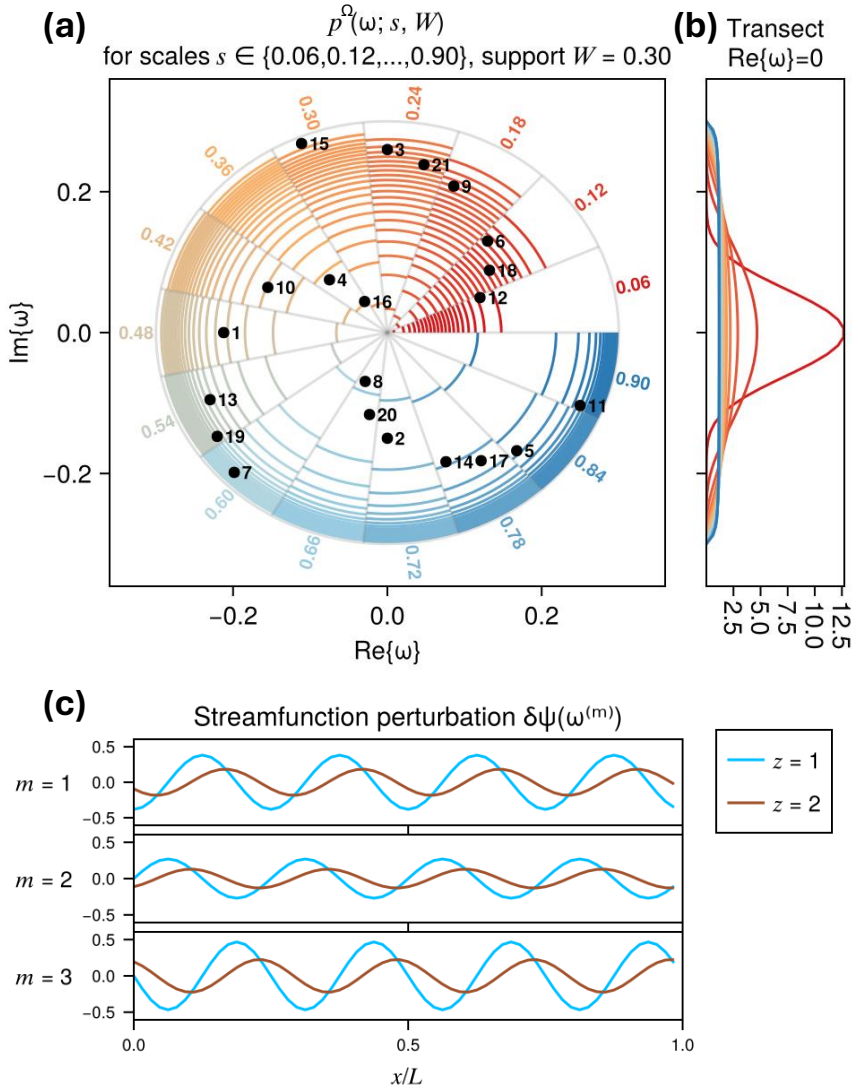

Figure 6. Structure of perturbations and their probability distribution. (a) Level sets of each considered input distribution from scales $s = 0.06$ (red) to $s = 0.9$ (blue), each scale restricted to $\frac{1}{15}$ of the circle each so that all scales may be seen. Labels on the outer edge of the circle indicate the corresponding scale. Dots show the 21 impulses used at each AST before each ancestor, sampled by quasi-Monte Carlo. (b) One-dimensional transects of $p(\omega; s, W)$ at each scale. (c) The shape of perturbations to the streamfunction corresponding to $\omega_1, \omega_2, \omega_3$. Note that the absolute amplitudes and phases vary, sampling the two degrees of freedom in the disc, but the relative amplitudes and phases of the upper and lower layers are fixed.

We stick to the simpler approach of the most unstable modes of the background shear, choosing to focus attention on the less-studied optimization of the advance split time given a fixed perturbation shape. There are several reasons that singular vectors may not be suitable for our goals. First, it is easier to compare different initial conditions, different advance split times, and even different topographies (which we don't do here) when they are all subject to precisely the same perturbation. Second, as our results will demonstrate, the COAST tends to lie beyond the time range where linearized error dynamics are appropriate, which is natural because we aim for finite-amplitude boosts in extreme events. Third, singular vectors are typically designed to optimize global errors, which might not be as relevant for local extremes. Fourth, such highly specialized perturbation shapes might not be accessible in a generic GCMs. Nonetheless, sensitivity analysis with respect to perturbation shape leads the agenda for follow-up work.

Having fixed a subspace $\Omega = \mathbb{C}$ for perturbations ω, we need to specify an input distribution $p^{\Omega}(\omega)$ over that space. We design the PDF for ω as a radially symmetric, smooth, "bump function" which has compact support in order to prevent perturbations so large as to induce oscillatory transients. The PDF is parameterized by two scales: W which is the maximum permissible magnitude of ω, and s which sets the typical perturbation strength:

$$p(\omega; s, W) \propto \exp\left[-\frac{|\omega|^2}{2s^2}\left(1 - \frac{|\omega|^2}{W^2}\right)^{-1}\right] \text{ for } |\omega| < W, \text{ and 0 for } |\omega| \geq W. \tag{38}$$

When $s \ll W$, p is approximately a bivariate Gaussian density with diagonal covariance $s^2 I$. When $s \gtrsim W$, p is approximately uniform over the W-disc $\{\omega : |\omega| \leq W\}$, with rapid (but mathematically smooth) transition to 0 on the boundary. We fix $W = 0.3$, limiting the maximum possible perturbation amplitude to $|\delta\psi| \leq 2W = 0.6$ (see text after Eq. (37)), which is small compared to the characteristic streamfunction amplitude of $|\psi| \sim 10$. We include s as a parameter to vary because there is no established principle to set the magnitude of impulses for the purpose of rare event sampling. In contrast, numerical weather prediction has an established (if heuristic) practice of tuning noise amplitude to match ensemble spread with model error (e.g., Berner et al., 2015). Optimizing for climatological accuracy is a different, murkier goal calling for less prejudice with regard to perturbation magnitude. We therefore vary s widely from 0.06 to 0.9 in increments of 0.06 for 15 total values. s is the impulsive-forcing analogue to the continuous-forcing amplitude that we called F_4 in Finkel and O'Gorman (2024), which strongly influenced the perturbation growth rate and therefore the optimal advance split time.

Fig. 6(a,b) depicts $p(\omega; s, W)$ in two ways: (a) two-dimensional level sets of the unnormalized density (38) logarithmically spaced from e^{-4} to $e^{-0.01}$, each value of s occupying one of 15 sectors of the circle; and (b) one-dimensional transects across $p(\omega; s, W)$ fixing $\mathrm{Re}\{\omega\} = 0$. To save the labor of drawing Monte Carlo samples from $p(\omega; s, W)$ separately and simulating the perturbed children for each value of s, we compute the MoCTail and PoPTail estimators using numerical quadrature over the W-disc using a single set of samples drawn by *quasi*-Monte Carlo (QMC), and displayed as black dots in 6a. QMC is a general strategy which places samples deterministically across the input space in a way that mimics properties of randomness, but with lower *discrepancy* (fewer clumps and patches), thereby aiming to reduce variance in estimated statistics (Leobacher and Pillichshammer, 2014). We specifically use the `LatticeRuleSampler` from the `QuasiMonteCarlo.jl` Julia library (Rackauckas, 2023) to distribute points $\{(U_m, V_m)\}_{m=1}^{M}$ quasi-uniformly on the unit square $[0,1]^2$, and transform them to the

W-disc with the formula

$$\omega_m = W\sqrt{U_m}\exp(2\pi i V_m). \tag{39}$$

Since U_m is a "quasi-random sample" of the uniformly distributed random variable $U \sim \mathcal{U}([0,1])$, we have

$$\mathbb{P}\{r_1 \le |\omega| \le r_2\} = \mathbb{P}\{r_1^2 \le W^2 U \le r_2^2\} = \mathbb{P}\left\{\frac{r_1^2}{W^2} \le U \le \frac{r_2^2}{W^2}\right\} = \frac{r_2^2 - r_1^2}{W^2} \tag{40}$$

which is the fraction of the W-disc between the radii r_1 and r_2. The phase $2\pi V$ is clearly $\mathcal{U}([0,2\pi])$. If U and V were independent random variables, we would immediately conclude ω is uniformly distributed over the W-disc; in QMC they are not independent, but the conclusion still holds true (Leobacher and Pillichshammer, 2014). In all experiments to follow, $M = 21$, corresponding to the 21 points plotted in Fig. 6a. While other sampling rules are possible, the `LatticeRuleSampler` enjoys a distinct advantage of being extensible: sampling 12 points at first and later deciding to add 9 more gives the same result as sampling 21 in one batch.

4.2 Sweeping over ancestors and advance split times

Following the procedure laid out in Sect. 2, we apply each perturbation $\{\omega_m\}_{m=1}^{M}$ to a collection of ancestor events $\{\mathbf{x}(t_n^*)\}_{n=1}^{N}$ at a range of ASTs $\{t_n^* - A_j\}_{j=1}^{J}$. We set the number of ancestors, N to whichever is smaller: the total number of cluster maxima (see Sect. 3) in the short DNS, or 32. Considering all latitudes, the minimum N was 14, the median was 22, and the maximum 32 was found at four latitudes including $y_0 = \frac{26}{64}L$ which we consider in more depth. In the equal-cost comparisons to be shown later, we restrict N to smaller values. The ASTs sampled are $\{A_j\}_{j=1}^{J=20} = \{2, 4, \ldots, 40\}$, with a two-day spacing chosen as roughly half the period of small fluctuations in $R(\mathbf{x}(t))$ (see Fig. 7).

5 Results: conditional severity distributions

In this section we present some case studies of conditional perturbed ensembles (from individual ancestors) and corresponding dispersion measures to be subsequently used in the MoCTail and PoPTail estimation. The results will add context and motivation to the protocols laid out above, and set the stage for the aggregation of results across ancestors.

5.1 Perturbed ensembles: case studies

Fig. 7 displays a small but representative sample of boosted ensembles at two target latitudes at the inner edges of the two eastward jets: (a) $y_0 = \frac{38}{64}L$ and (b) $y_0 = \frac{26}{64}$. The ancestors' intensity (black dashed curves) reach their respective peaks at times $t^* = 4152$ for (a) and 2702 for (b). Note the differences in peak value and peak shape: the upper latitude has long-lasting, flat maxima and the lower latitude has brief, spiky maxima. The statistical properties at these two locations, both in Fig. 7 and in Fig. 3, are approximately equivalent after reflection about $\frac{1}{2}$ ($c \to 1 - c$), meaning the upper tail of one resembles the lower tail of the other. This can be understood by the approximate north-south symmetry of the tracer's dynamics imposed by Dirichlet boundary conditions.

We show the perturbed intensities launched from three ASTs $A \in \{2, 16, 32\}$, colored (red, orange, blue) respectively. Following the split time, the ensemble members spread apart from the parent and from each other, achieving their own peak values (severities) that differ in both amplitude and timing from the ancestor, the discrepancies increasing with A. The red curves ($A = 2$) replicate the ancestral peak very closely; the orange curves ($A = 16$) peak at substantially higher or lower levels, and up to ~ 2 days earlier or later. Still, the orange peaks are clearly dynamically related to the ancestral peaks. This is no longer true for the blue curves ($A = 32$), whose intensity peaks are widely scattered in time and systematically lower than the ancestors' peaks.

Besides these three selected ASTs, each descendant is charted in Fig. 7(a,b).i as a circle color-coded by AST, positioned vertically at its severity value and horizontally at its launch time. A corresponding star is plotted in Fig. 7(a,b).ii, positioned vertically at its severity value (on a zoomed-in scale) and horizontally at its peak timing (constrained by the "argmax drift" parameter $\delta t^* = 5$ days, as explained in Sect. 2.1). We can see the transition of the R^* ensemble from tightly clustered (for short AST) to roughly independent and climatologically distributed (for long AST), and in between there is a golden window of opportunity where severities can be both large and diverse. The optimal AST must balance these two objectives, a task akin to the exploitation-exploration tradeoff in Bayesian optimization and reinforcement learning (e.g., Yang et al., 2022). In this light, the two functionals defined in Eqs. (26) and (27) are candidate *acquisition functions*.

5.2 Relating severities to impulses: case studies

We now construct "severity response functions" $\widehat{R}^*_{n,j}(\omega;\theta)$ mapping impulses $\omega \in \mathbb{C}$ to severities R^*, approximating the action of the flow map using some empirical parameters θ. This will be needed to estimate conditional and unconditional probabilities through the MoCTail and PoPTail estimators (see Eq. (5)), and will also help to understand the joint dependence between impulses $\omega \in \mathbb{C}$ and the times $\{t^*_n - A_j\}$ at which they are applied.

How should the response functions be parameterized? The simplest choice would be a linear model, often used in numerical weather prediction to optimize ensemble spread by perturbing in the most-effective directions, so-called singular vectors (Diaconescu and Laprise, 2012). However, linear models are strictly valid only for infinitesimal perturbations, hence short lead times. Similar logic should apply when optimizing for severity instead of ensemble spread, and indeed we demonstrate below that the COAST tends to lie beyond the range where a linear model $\widehat{R}^*$ is valid. We therefore construct a quadratic model as well, and it turns out that this minor upgrade is sufficient. Future work with more complex dynamics and objectives may call for more elaborate response functions (orthogonal polynomials, Gaussian processes, and neural networks for example), but we adhere to quadratic models in this study as a proof of concept that is easy to construct and interpret, which we do in the following two figures.

The linear and quadratic response functions take the form

$$\widehat{R}^*(\omega;\theta) = \theta_0 + \theta_1 \mathrm{Re}\{\omega\} + \theta_2 \mathrm{Im}\{\omega\} \qquad \theta_0, \theta_1, \theta_2 \text{ fitted for both linear and quadratic models} \tag{41}$$

$$+ \theta_3 \mathrm{Re}\{\omega\}^2 + \theta_4 \mathrm{Re}\{\omega\}\mathrm{Im}\{\omega\} + \theta_5 \mathrm{Im}\{\omega\}^2 \qquad \theta_3, \theta_4, \theta_5 \text{ fitted for quadratic model only.} \tag{42}$$

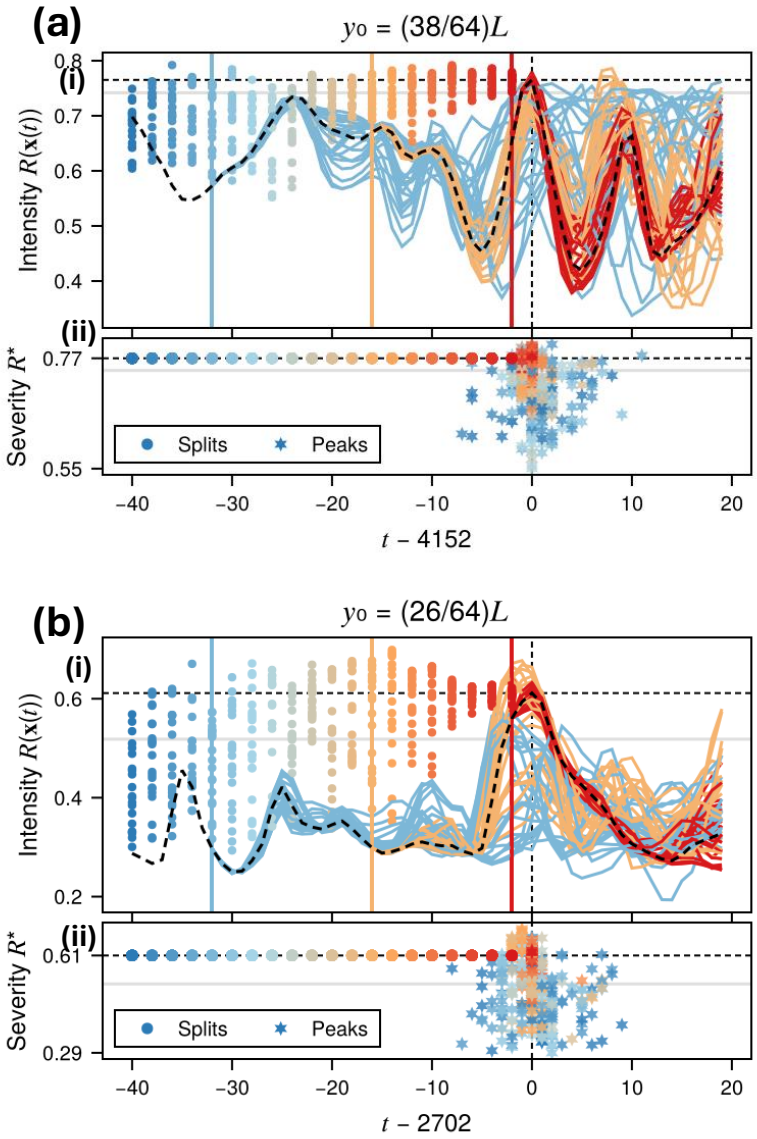

Figure 7. Boosted ensembles of two selected events: (a) time $t^* = 4152$ at latitude $y_0 = \frac{38}{64}L$, and (b) time $t^* = 2702$ at latitude $y_0 = \frac{26}{64}L$. These are times when the intensity function $R(\mathbf{x}(t))$ from the long DNS (dashed black curves) achieved a peak value (horizontal dashed black lines) above the threshold $\mu[(\frac{1}{2})^5]$ (horizontal gray lines). For each AST $A \in \{2, 4, \ldots, 40\}$, an ensemble of perturbed events (descendants) is launched at $t^* - A$, indexed by $m = 1, \ldots, 21$. For three selected ASTs $A = 2, 16, 32$, the full timeseries $\{R_m(t)\}_{m=1}^{21}$ are shown in (a,b).i. The red-to-blue color scale indicates short-to-long ASTs. Each descendant achieves a different severity R_m^* (peak intensity), indicated by circles in (a,b).i at $(-A, R_m^*)$ for all values of A. The peaks also occur at different times t_m^*, indicated in (a,b).ii by stars at $(t_m^* - t^*, R_m^*)$, again for all A and colored accordingly.

We use ordinary least squares regression on the $M = 21$ sampled impulses $\{\omega_m\}_{m=1}^{M}$ and associated severities $\{R^*_{n,j,m}\}$, in addition to the non-perturbed ancestor ($\omega_0 := 0$) with severity $R^*_{n,j,0} = R^*_n$. A different set of coefficients is calculated separately for each ancestor n and AST A_j. The response functions for the same ancestor event as in Figs. 7b are visualized in Fig. 8, using (a) the two-dimensional response surfaces, (b) the true vs. fitted response values, (c) the overall slope, measured by the linear coefficient magnitudes, (d) the overall curvature, measured by the eigenvalues of the Hessian of the quadratic fit, and (e) the overall linear and quadratic skills, measured by the coefficient of determination R^2. The response surface gradually transforms from a linear plane, to a curved hilltop, to a saddle, to a jagged landscape, as AST increases. Accordingly, the linear and then the quadratic model lose their skill. The quadratic model is slightly better than the linear model for this particular event, but substantially better when averaged across all events (see the forthcoming Fig. 9c.i), and so we will use quadratic models only as $\widehat{R}^*$ in the tail estimators.

5.3 Conditional severity PDFs: case studies

Equipped with response functions approximated by quadratic models, we can now construct conditional severity PDFs using Eq. (10), which are displayed in Fig. 9a. For the same ancestor as in Fig. 8 and the same six ASTs, we can see the relationship between actually sampled perturbed severities (red circles and lines), fitted severity PDFs (colored curves, one color for each input scale s) evaluated at the bins with lower boundaries $\{\mu[(\frac{1}{2})^k] : k = 5, \ldots, 14\}$, and the climatological PDF (black curves). As AST increases from right to left, the severity PDFs morph from narrow spikes centered at the ancestor severity to long, extended lumps reaching far beyond the ancestor severity, and then recede below the threshold $\mu[(\frac{1}{2})^5]$. The PDF's motion resembles a wave crashing onto a shallow beach, blanketing the sand, and then retreating, hitting the true COAST somewhere in the middle stages. But this general behavior is strongly modulated by the choice of scale s: red PDFs, representing the smallest scale $s = 0.06$, are narrower and located closer to the ancestral severity (horizontal black line) for all ASTs, whereas blue PDFs, representing the largest scale $s = 0.9$, spread out further as a result of giving more weight to bigger impulses. This underscores our claim that the input distribution, an arbitrary choice, merits sensitivity analysis, and so we carry it through the remaining steps.

5.4 AST selection criteria: case studies

Figure 10 display the criteria proposed in Sect. 2.4 that might help determine in which stage of "wave breaking" the severity PDF finds the COAST. The EI and TE criteria shown in panels Fig. 10(a,b) both exhibit non-monotonic behavior by design, maximizing at COASTs denoted A^{EI} and A^{TE} (see Sect. 2.4). The AST dependence can be heuristically understood in light of the PDFs in Fig. 9:

- At small AST, the narrow PDFs have a relatively high *probability* of improvement over the ancestor ($\sim \frac{1}{2}$), but only by small amounts, hence a small EI. By a similar token, the TE terms in Eq. (27) are almost all positive because the PDF is situated well above μ, but being concentrated in a small number of bins makes its information content low.

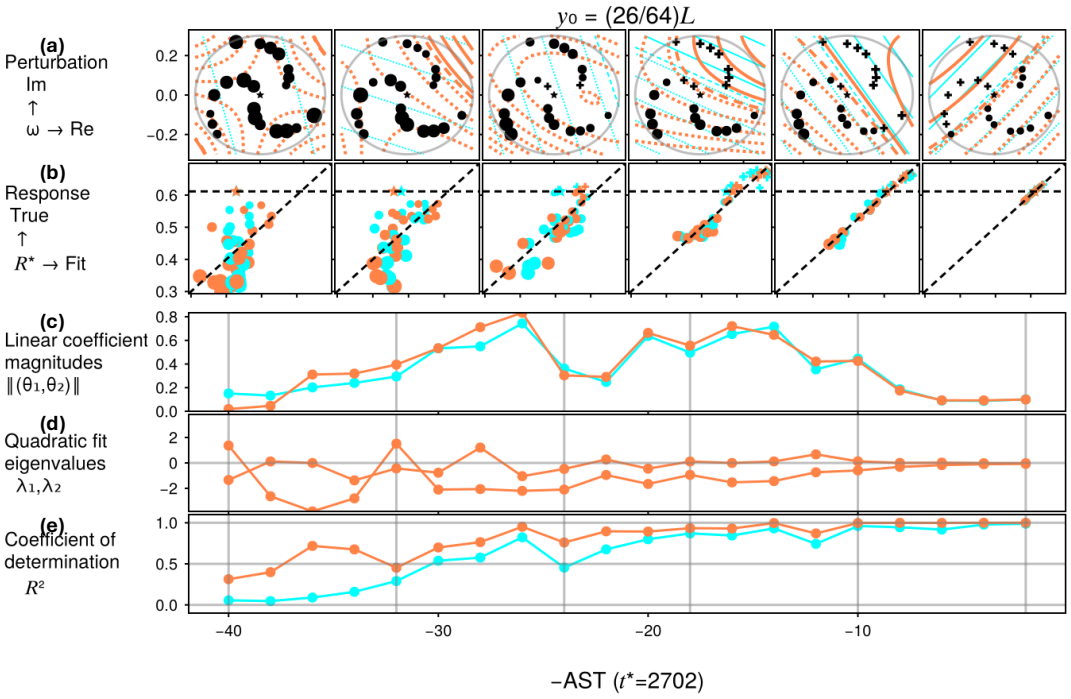

Figure 8. The response of an extreme event to perturbations: magnitude, phase, and timing. The event is the same as in Fig. 7b. Row (a) represents impulses as in Fig. 6, but additionally shows the responses to them separately at six sampled ASTs (2, 10, 18, 24, 32, and 40 days, marked with vertical gray lines in c-e), which increase from right to left (launch time $t^* - A$ increases left to right). Horizontal and vertical scales are equal. At the shortest AST shown, $A = 2$, the response function is clearly linear: the impulses above and left of center are marked by $+$, representing an increased severity, and those below and right of center are marked by $\bullet$, representing decreased severity, with marker sizes representing the magnitude of the change. Colored curves represent level sets of the fitted linear (cyan) and quadratic (orange) models, with (solid, dashed, dotted) contours to differentiate (positive, zero, negative) changes to R^*. Row (b) displays the quality of these models by plotting true vs. fit responses (again, horizontal and vertical scales are equal). As AST increases, the impulses causing higher and lower severities become more intertwined and less linearly separable, as the orange contours progressively bend and separate from the cyan contours. Accordingly, the modeled linear response ceases to correlate with the true response. The modeled quadratic response has a slightly longer range of good quality, but also fails for AST $\gtrsim 26$ days. Row (c) shows that the linear components θ_1, θ_2 are estimated similarly (at least in magnitude) regardless of whether quadratic terms are also included. Row (d) shows that the quadratic model implies a local maximum (both eigenvalues nonpositive) for most of the range $A < 26$, beyond which the landscape starts looking less like a hilltop and more like a saddle. Row (e) displays the coefficients of determination, R^2 (not to be confused with intensity R or severity R^*, which fortunately we never need to square).

– At intermediate ASTs of 10-20 days, the PDFs remain roughly centered at the ancestor's severity, meaning that improvements remain highly probable, but are larger when they happen thanks to the long upper tails, contributing to a large EI. Meanwhile, both upper and lower tails contribute to a large TE, which does not directly favor exceptionally high severities but rather *diverse* severities that are *high enough* to exceed μ.

– At large AST past ~ 25 days, the PDFs have diminishing mass above μ, let alone above the ancestor severity R_n^*, which zeros out most of the contributions to both EI and TE.

The COAST can change with the scale s: even though the overall shapes of TE and EI don't change very much, the location of their maxima might. Fortunately, we will find changes in scale for $s \gtrsim 0.24$ to have negligible impact.

Fig. 10(c,d) display two versions of pattern correlation ρ, defined in Sect. 2.4 for an arbitrary field F: the "global correlation" $\rho[c]$ uses the whole two-dimensional upper-layer concentration field $F(x,y) = c_1(x,y)$, and the "local correlation" $\rho[c(\cdot, y_0)]$ uses only the single-latitude transect $F(x) = c_1(x, y_0)$ at the target latitude y_0. Both drop off steadily with AST, although local correlation fluctuates more due to averaging a smaller spatial region. The influence of perturbation scale s enters at the ensemble-averaging step, where the mth member's pattern correlation $\rho[F_0, F_m]$ is weighted by $p(\omega_m, s, W)$. Since smaller perturbations take longer to grow, smaller input scales lead to slower dropoff of ρ with A—but only at short lead times, where errors are still tiny. Beyond $A \approx 6$ and 10 days for global and local correlations respectively, decorrelation proceeds at a similar rate with respect to increasing AST for all scales. The nominal threshold $\rho = \sqrt{1 - (\frac{3}{8})^2}$ is marked in both, and gives a similar AST for local and global correlations but generally longer than implied by EI or TE.

5.5 AST selection criteria: aggregate results

Fig. 11 goes beyond the case study to show dispersion indicators averaged across all ancestors. The coefficients of determination for linear and quadratic models (Fig. 11a) are farther apart on average than they are for the case study (see Fig. 8e), the quadratic model enjoying much higher skill especially during the pivotal 10-20 day range when EI and TE tend to maximize (Fig. 11b,c). This validates our choice to use the quadratic model. Overall, the EI, TE, global and local correlations (Fig. 11 b-e) are similar on average to the case study, but smoother.

Note, however, that these averaged dispersion indicators are never used directly in AST selection: the COASTs are chosen separately for each ancestor as the maximizer of its own EI or TE, or at the longest AST such that global or local correlation is above ρ^{U}. This nuance is further illustrated in Fig. 12(a,b), where (EI, TE) are plotted as joint functions of AST and input scale. Whereas the heatmaps are averages over ancestors of EI and TE just like Fig. 9c.(ii,iii), the red circles indicate the fraction of ancestors whose EI or TE is maximized at a particular AST for each particular scale. We call the red circle sizes "COAST frequencies". For example, at $s = 0.24$, the mean EI maximizes at $A = 14$ days, and that same AST is the most frequent COAST. However, the second-largest circle indicates that $A = 20$ days is a close second-most frequent COAST according to EI. At the same scale, the most frequent COASTs according to TE are $A = 18$ and 20. In general, we gather two patterns from Fig. 12(a,b): the average EI and TE values (i) are well-correlated with their corresponding COAST frequencies, and (ii) both change rapidly at small scales but stabilize above $s \approx 0.24$, at which point the input distributions are close enough to uniform

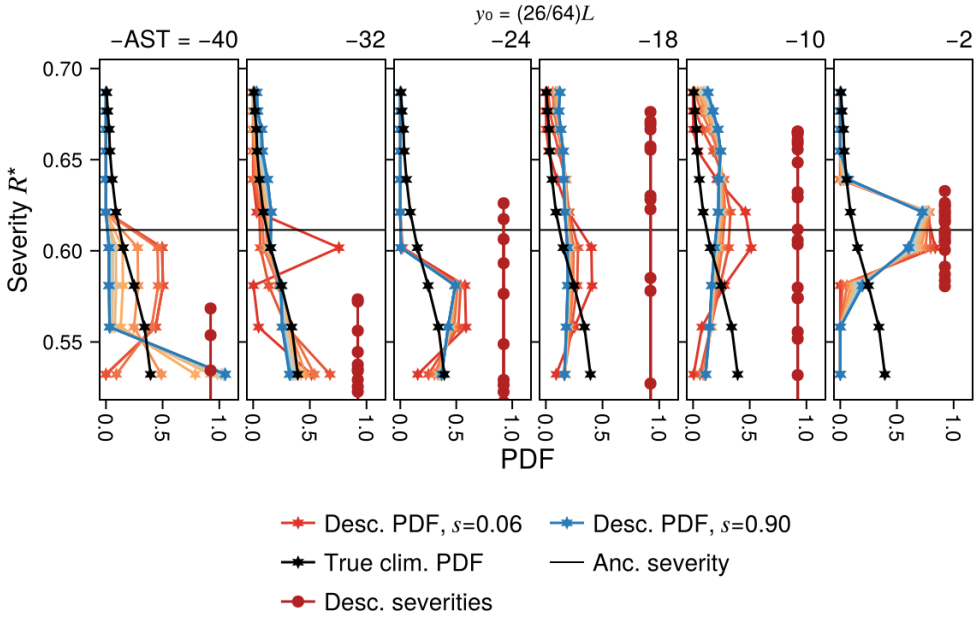

Figure 9. Severities and their conditional distributions for the same case study as Fig. 7b. For six ASTs (same as Fig. 8, decreasing from left to right), perturbed severities are displayed as dark red circles along a vertical line, and the unperturbed (ancestral) severity is marked with a horizontal black line. Colored curves and stars show the severity PDFs above $\mu = 0.52$ as inferred from the quadratic regression, for a range of scales s from 0.06 (red) to 0.9 (blue). Black curves with stars represent the climatological tail PDF, as inferred from the long DNS, which we will seek to estimate by combining conditional distributions over many ancestors (not just the single ancestor considered here).

over the W-disc. This relative stability is reassuring, but we generally prefer smaller noise which disturbs the model dynamics less. To balance these considerations, we select $s = 0.24$ as the nominal scale to examine more closely going forward.

6 Results: Climatological severity distributions

Having explained the construction of conditional distributions, we now aggregate across ancestors using MoCTail and PoPTail estimators to obtain our estimates of the climatological severity distribution from the boosted ensembles. We evaluate the skill of each AST selection rule by the χ^2 divergence of the resulting climatological distribution from ground truth as obtained from the long DNS. We first restrict attention to extremes at $y_0 = \frac{26}{64}L$ and then assess a broader swath of latitudes.

First, consider the simplest AST selection rule $A = A^{\mathrm{U}}$, a uniform AST over all ancestors. We have no *a priori* principle for A^{U}, so we search through all possible values from 2 to 40 days. Fig. 12c displays the resulting χ^2 divergence between the MoCTail and ground truth, as a function of A^{U} and input scale. A clear optimum emerges at $A^{\mathrm{U}} = 14$ days and persists for all scales $s \gtrsim 0.24$, after rapid changes across smaller scales. Red contours also indicate the local correlation, averaged across

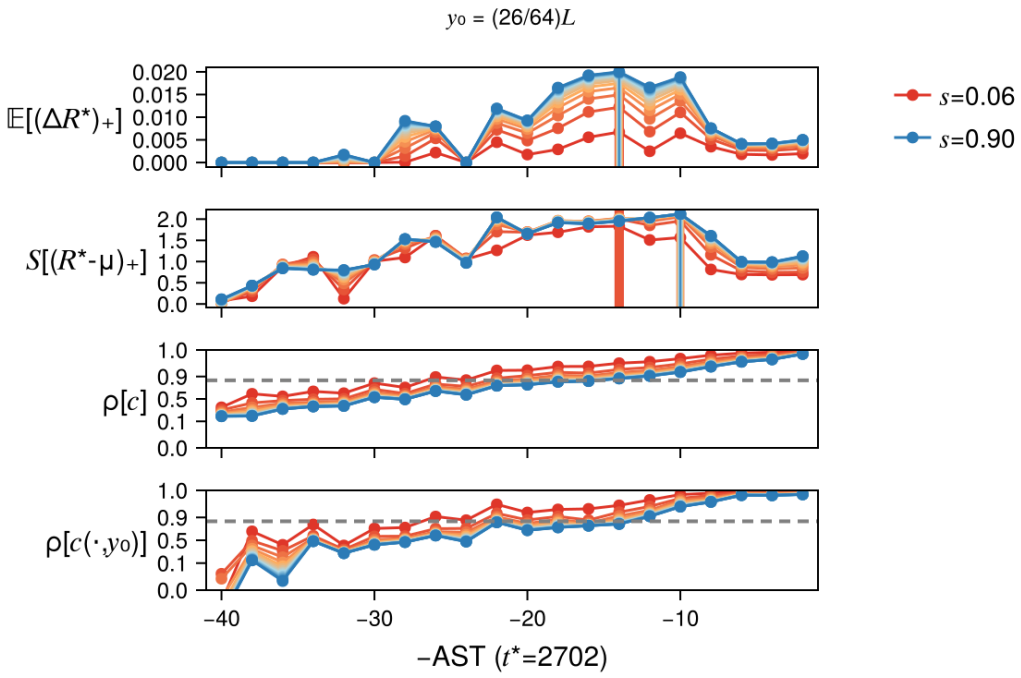

Figure 10. Ensemble dispersion indicators as a function of AST, again for the same case study as Fig. 7b: (a) expected improvement EI, (b) thresholded entropy TE, (c) local and (d) global correlations. Colors indicate input scales s, from small (red: $s = 0.06$) to large (blue: $s = 0.9$). In (a,b), vertical bars mark the respective optimal ASTs, which may depend on the scale. In (c,d), horizontal dashed lines are positioned at $1-(\frac{3}{8})^2$, corresponding to the rule of thumb from Finkel and O'Gorman (2024), and vertical axes are stretched with a modified sigmoid to magnify numbers close to one and zero.

ancestors to give a smooth and monotonic function of AST. In terms of correlation, the COAST $A^{\mathrm{U}} = 14$ days corresponds to $\rho^{\mathrm{U}} \approx 0.92$ depending on the scale, which is slightly above the nominal value $1-(\frac{3}{8})^2 = 0.86$, meaning one should split a little bit closer to the event than the rule of thumb implies.

Overall, the χ^2 landscape (inverted) roughly aligns with the EI and TE landscapes, as do their respective optima. This is remarkable and encouraging: allowing each ancestor to determine its own COAST independently, with no knowledge of the ground truth or even other ancestors' COASTs, leads to a similar solution as the policy of synchronizing them all. Boosting based on EI and TE, therefore, is more parallelizable (optimizations are decoupled across ancestors), extensible (new ancestors can be added without changing the optimal split times for pre-existing ancestors), and interpretable (one can see the optimum clearly based on a case study, without complicated averaging procedures across initial conditions).

Fig. 13 makes a tail-to-tail comparison between all the AST selection rules (a.i-v: A^{U}, A^{PC} local and global, A^{EI}, A^{TE}), fixing the scale to $s = 0.24$ and (in the case of A^{U} and A^{PC}) selecting *post-hoc* the best-performing threshold to set the COASTs. We used subsets of only 11 of the 32 ancestors, resampling such subsets 64 times to obtain medians (solid) and interquartile ranges

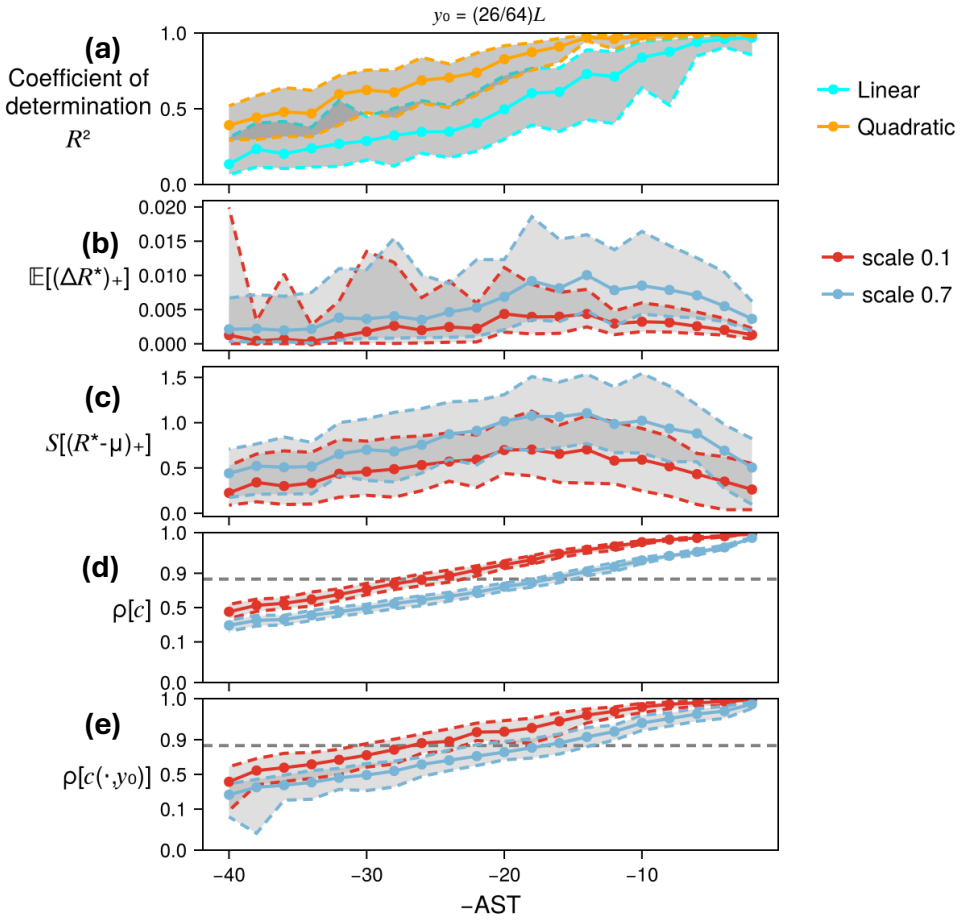

Figure 11. Ensemble dispersion metrics averaged across ancestors at $y_0 = 26/64L$. (a) Coefficients of determination for linear (cyan) and quadratic (orange) regressions, averaged across ancestors. (b-e) same quantities as in in Fig. 10(a-d) but averaged across ancestors, with only the largest and smallest scales shown (red: $s = 0.06$, blue: $s = 0.9$). Shaded regions show the areas between truncated upper and lower means. E.g., for the correlation ρ, the truncated upper mean 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 as a compromise between quantiles (which are erratic for the relatively small sample size of ancestors) and standard deviation envelopes (which can misleadingly fall outside the bounds $[0, 1]$ to which ρ is constrained).

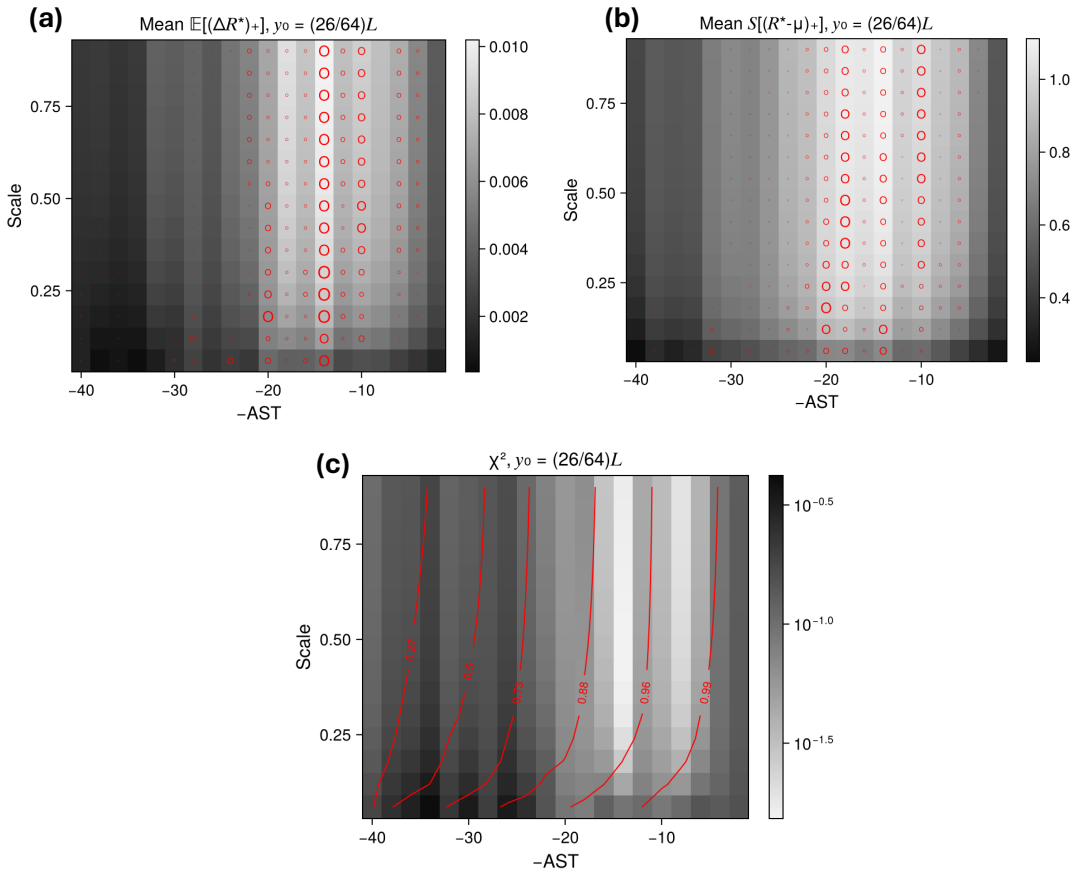

Figure 12. Three optimization landscapes as joint functions of AST and input scale for $y_0 = (26/64)L$: (a) expected improvement (EI), (b) thresholded entropy (TE), and (c) χ^2 divergence between the MoCTail and ground truth. Lighter gray indicates better performance—smaller χ^2 divergence or larger EI and TE—and the corresponding "best" ASTs consistently fall in the *interior* of the domain, across all scales. Contours of local correlation $\rho[c(y_0, \cdot)]$ are overlaid in (c), giving a rough map of correspondence between correlation levels and AST. The size of red circles in (a,b) indicate the "COAST frequency": the fraction of ancestors whose (EI, TE) is maximized at the corresponding AST while holding the scale fixed. Note the multiple local maxima in mean EI and TE (as indicated by the lightness of the gray color in (a,b)), each of which is the global maximum for some significant set of ancestors.

(shading) on CCDFs. The numerical values of optimal AST and ρ reported above a.(i-iii), with PoPTail optima parenthesized, are the optima obtained from $N = 32$, i.e., the best estimates of the true optima; they don't necessarily correspond to the values used for plotting with $N = 11$, which are optimized separately for each resampling. The brown CCDF in panel (a.vi) is the estimate from the unboosted acestors alone ("equal-N"), and the black is the estimate from a larger number of ancestors to equal the cost of boosting. The curves underneath in panel (b) show the rate of improvement of χ^2 with N.

In terms of quantitative improvements in χ^2, 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 with moderate N, 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)).

We selected $N = 11$ to display the full CCDFs in Fig. 13(a) as the middle range of values tried, and where enough equal-size ancestor subsets are available for uncertainty quantification by bootstrapping. When comparing with DNS CCDFs, all five rules successfully extend the short, equal-N DNS tail into a longer tail that tracks closer to the ground truth farther into the extreme severity range. They also all find a larger maximum than even the equal-cost DNS found. However, the threshold-based rules exhibit apparent bias, systematically underestimating probabilities for $R^* \gtrsim 0.64$ with asymmetric variabilities, whereas the optimization-based rules are both more accurate and more confident.

The COASTs identified by all rules lie strictly between the shortest and longest ASTs considered. For example, $A^{\mathrm{U}} = 14$ according to the MoCTail estimator (using all $N = 32$ ancestors). By comparing with Fig. 12c, we recognize 14 as the minimum of the χ^2 landscape for $s = 0.24$ (and larger scales), with an approximate local-correlation equivalent of 0.98.

Similar patterns hold across target latitudes, but with some notable caveats. The χ^2 divergences of each selection rule are plotted in Fig. 14, of which Fig. 13c is one slice. The most obvious and important point holds: perturbed ensembles improve upon the DNS equal-N estimate, for almost all latitudes and AST selection rules, and they also improve on the equal cost estimate in many cases. But A^{EI} is less reliable; its favorable performance noted above in Fig. 13 is peculiar to the latitude $y_0 = \frac{26}{64}L$. At some other latitudes, it is similar or worse in skill than equal-N and even equal-cost DNS. Even so, it tends to fail by *overestimating* severities, which we have confirmed by examining the corresponding CCDFs (not shown), and thus it may serve as a useful upper bound. The MoCTail and PoPTail estimators are similar in quality across latitudes, but as observed in Fig. 13, PoPTail has an advantage with threshold-based rules ($A^{\mathrm{U}}, A^{\mathrm{PC}}$ local and global) whereas MoCTail performs better with optimization-based rules ($A^{\mathrm{EI}}, A^{\mathrm{TE}}$).

The various estimators and AST selection rules have differences in skill, but a more important commonality: all of them indicate that *an optimal advance split time exists* that is strictly positive, which is not a foregone conclusion in light of standard rare event algorithms like adaptive multilevel splitting (AMS; Lestang et al., 2018) without "trying early". Fig. 12 shows clear intermediate optima when targeting the single latitude $y_0 = \frac{26}{64}L$, and Fig. 15 extends this result to all latitudes by stacking together cross-sections of the per-latitude counterparts of Fig. 12 at $s = 0.24$. The COAST frequency and mean-TE landscapes

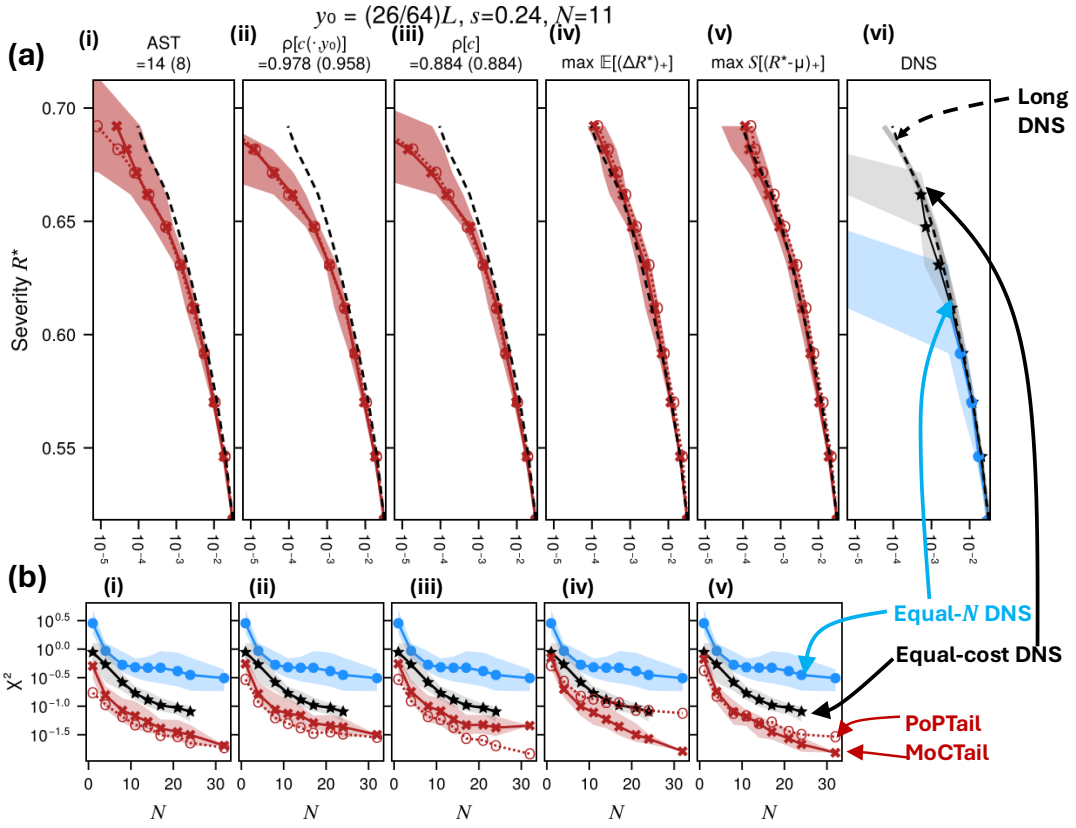

Figure 13. CCDF approximations by various mixing criteria and associated errors, at the latitude $y_0 = \frac{26}{64}L$ and input scale choice $s = 0.24$. (a.i-v) Tail CCDFs by various estimates using only $N = 11$ ancestors, with lines showing medians and bands showing interquartile ranges across many size-11 subsamples of the total set of 32 ancestors. Lines are medians, and bands are interquartile ranges. Dotted lines with open circles are PoPTails, while solid lines with crosses are MoCTails. Dashed black lines show the ground truth estimate. Panel a(i) shows the tail approximation using a single uniform AST indicated at the top: 14 days for MoCTail and 8 days (parenthesized) for PoPTail. Panels a.(ii,iii) show the tail approximations using thresholds of (local, global) correlations as AST selection criteria. Panels a.(iv,v) show the tail approximations obtained by maximizing (EI, TE), which unlike the other criteria do not rely on knowing the ground truth to select ancestor-wise ASTs, either directly or through threshold choice. (a.vi) also shows estimates from DNS with equal cost to boosting on 11 ancestors (black stars, gray envelope) and DNS from only $N = 11$ peaks (brown circles and envelope), in both cases estimating uncertainty by longitudinal rotation. The GPD fit to ground truth is shown as a gray curve. In a.(i-iii), the thresholds shown at the top (PoPTail thresholds parenthesized) are obtained by using all 32 ancestors, but the CCDFs displayed each choose an AST to minimize χ^2 divergence from ground truth, separately for each subsample. Because this requires ground truth knowledge, the χ^2 divergences must be interpreted as practical lower bounds. The 90% error bar applies to the MoCTail estimator only, and comes from bootstrapping on entire "families" or in other words mixture components (not individual descendants) and choosing the best AST (by χ^2 divergence) for each particular subsample. The error bar widths, too, must then represent lower bounds. (b) χ^2 values for the estimator directly above in each case as a function of N, and compared with DNS at equal cost and equal N. DNS does not run long enough to equal the total cost accrued by boosting 32 ancestors, so the black curve stops before the others.

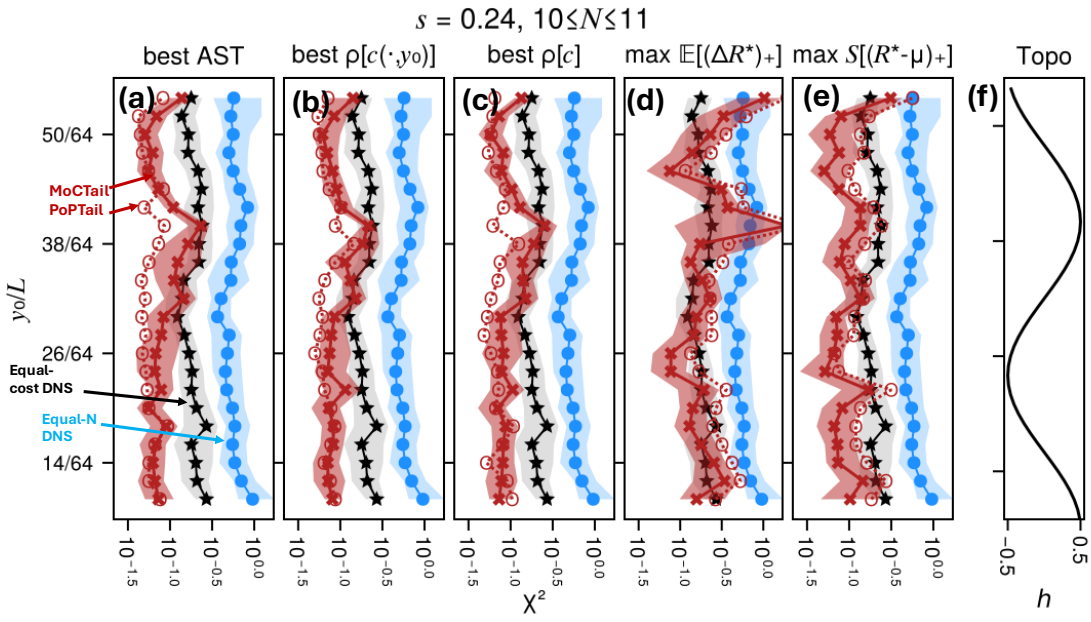

Figure 14. Performance of all AST selection criteria, measured by χ^2 divergence, across all latitudes for $s = 0.24$ and $N = 10$ or 11, whichever is nearest to 1/3 the number of ancestors found for the latitude in question (sometimes less than 32). Black line and gray envelope represent the error from the short DNS and its 90% error bar according to quantiles across longitudes. Panels a-e parallel Fig. 13a.(ii-vi). Solid lines and crosses represent the MoCTail estimator, while dotted lines with open circles represent the PoPTail estimator.

have broad ridges that meander slowly in AST space with latitude, approximately in phase with topography: smaller ASTs are favored at $y_0 \approx \frac{26}{64}L$, where topography is minimized and meridional wind shear is negative, and larger ASTs are favored at $y_0 \approx \frac{38}{64}L$, where topography is maximized and meridional wind shear is positive. A similar pattern, but with bigger swings, is seen in the χ^2 landscape. All these patterns are a bit noisy, especially for the COAST frequencies and χ^2-COAST locations, since both come from an inherently unstable "argmax" function. Nonetheless, the detailed latitude dependence is only a secondary effect on top of the main point, which is clearly demonstrated: splitting is most effective at intermediate ASTs rather than very short or long ASTs.

We can also now evaluate the $\frac{3}{8}$ rule from Finkel and O'Gorman (2024) in this broader multi-latitude context, though here we simplify the procedure by first averaging ρ across ancestors and then calculating A^{U} as a threshold-crossing time of that average, which we call $A^{\mathrm{U}}_{3/8}$, rather than averaging times $A^{\mathrm{PC}}_n[\rho^{\mathrm{U}} = 1 - (\frac{3}{8})^2]$ across ancestors. The same conclusion holds either way. The AST values $A^{\mathrm{U}}_{3/8}$ are overlaid on the χ^2 heatmap (Fig. 15d) as blue curves. The solid curve, representing a level set of ancestor-averaged global correlation, should be constant with latitude and varies only due to sampling errors. Likewise, the dashed curve, representing a level set of ancestor-averaged local correlation, should be approximately symmetric

with respect to latitude because of the symmetric tracer boundary conditions and approximate mirror symmetry in velocities, as should all the level sets in panel c. Since the A^{U} varies differently with latitude, exhibiting roughly odd symmetry about the midline, the $\frac{3}{8}$ rule cannot possibly be optimal for all latitudes simultaneously. More fundamentally, the COAST depends on more than just a generic metric for ensemble dispersion: it must also depend on the features of the tail being sampled, which in this case is the only possible source of broken symmetry (see Fig. 4).

However, both versions of $A^{\mathrm{U}}_{3/8}$ run right through the mean position of the meandering χ^2 valley and associated COASTs, performing about as well as any such highly-constrained synchronized A^{U} could do. Thus, the $\frac{3}{8}$ rule retains its relevance as a starting point for more refined optimization more tailored to the event, at least for this QG system. Whether the $\frac{3}{8}$ rule generalizes to more heterogeneous systems as the "optimal synchronized AST" requires further investigation. We found it provides some guidance for temperature and precipitation extremes in an idealized general circulation model, but overestimated the optimal AST in both cases (Finkel and O'Gorman, 2025).

7 Conclusion

Rare event sampling is a promising strategy to study extreme weather more efficiently with computer models by repeatedly cloning, perturbing, and re-simulating the most extreme events in an ensemble while tracking statistical weights. However, sudden and transient events such as mid-latitude precipitation present a particular challenge for rare event algorithms, leaving ensembles little time to diversify before the event passes by. Ensemble boosting (Gessner et al., 2021; Gessner, 2022; Fischer et al., 2023; Bloin-Wibe et al., 2025) and "trying-early adaptive multilevel splitting" (TEAMS; Finkel and O'Gorman, 2024) get around this problem by perturbing events farther in advance by some *advance split time* (AST) to allow ensembles to spread, but this opens a pivotal question: how should we choose the AST for maximal accuracy and efficiency? If AST is too short, perturbations can't grow enough to give useful samples, and if it is too long, they regress to climatology. To deploy advance-splitting methods at scale, we need more reliable ways to set the AST as well as other hyperparameters.

In this paper, we have established the *conditionally optimal advance split time* (COAST) as a quantity more intrinsic to the dynamical system than to the whimsies of a particular rare event algorithm by removing the confounding effect of randomly selecting ensemble members to split. The COAST also depends on the target observable of interest, the imposed distribution over perturbations, and the initial conditions which may vary in their predictability. We formulate COAST mathematically as the solution an optimization problem, and through a systematic boosting-based sampling and estimation procedure we discern the optimization landscape in the context of an idealized physical model: a baroclinically unstable quasi-geostrophic flow, with local passive tracer fluctuations as our extreme event of interest. To faciliatate more efficient rare event sampling applications, we have further proposed various parsimonious rules for finding the COAST, and evaluated these rules empirically in the QG model.

We have four conclusions to report:

1. A boosting procedure, generated with a suitable AST, can well-approximate a probability distribution's tail using MoC-Tail or PoPTail estimators.

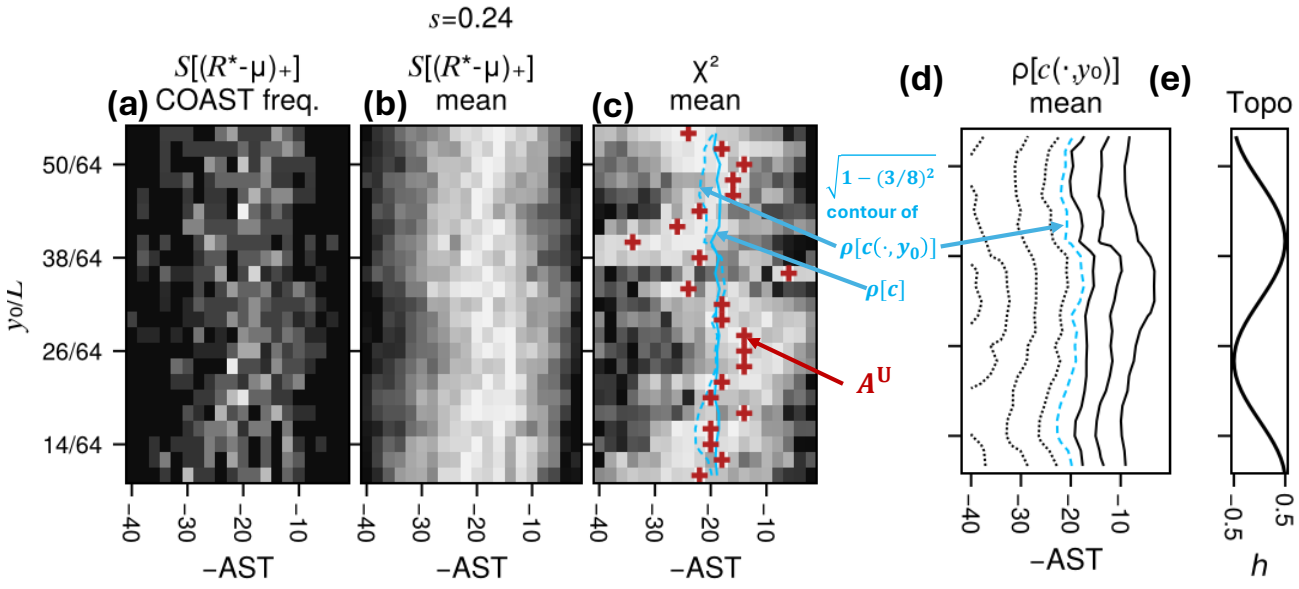

Figure 15. Optimization landscapes and optimal ASTs across latitudes, again fixing the input scale to $s = 0.24$. (a) Frequencies of *conditionally* optimal ASTs (COASTs), in the maximum-thresholded entropy sense, at each latitude, with whiter shading indicating higher frequency. E.g., at $y_0/L = 26/64$, the two adjacent bright pixels at AST $= 18, 20$ indicate that for a large fraction of ancestors, the highest-entropy descendant ensemble is the one launched 18 or 20 days in advance of the peak. (b) Thresholded entropy as a function of AST, normalized to the range 0-1 (black-white, so brighter is better) separately at each latitude. This landscape is smoother than χ^2 and varies less dramatically with latitude, but exhibits directionally similar trends. (c) χ^2 divergence as a function of AST and latitude, normalized to the range 0-1 (white-black, so brighter is better) separately at each latitude so that different latitudes are visually comparable. Red crosses mark the optimal AST at each latitude. Cyan (solid, dashed) curves mark the AST at which the (global, local) correlations, averaged across ancestors, reach $1-(\frac{3}{8})^2$. This nominal choice is based on Finkel and O'Gorman (2024), and falls squarely in the middle of the latitude-dependent ASTs. (d) Contour map of local correlation, averaged over ancestors, as a function of AST and latitude. The levels range from 0.22 (left-most dotted black curve, fragmented by boundary) to 0.99 (rightmost solid black curve), evenly spaced in a stretched sigmoid scale (levels are shown and only for qualitative purposes). The reference level $1-(\frac{3}{8})^2$ appears dashed in cyan. (g) Bottom topography for reference.

2. The optimal AST is strictly greater than zero and varies slowly with latitude, appearing smaller in regions of negative meridional wind shear (e.g., the northern edges of westerly jets) and larger in regions of positive meridional wind shear (e.g., the southern edges of westerly jets).

3. Several different rules for selecting the COAST are equally effective. Beyond the simplest option of setting a single fixed AST (called A^{U}), one can set a conditional AST (called A^{PC}) by thresholding on ensemble dispersion. Both A^{U} and A^{PC} perform similarly at tail reconstruction, but both unfortunately require a threshold choice, which there is no established method for selecting. Here we selected thresholds *post hoc* with knowledge of the ground truth. The rule proposed in Finkel and O'Gorman (2024)—that $A^{\mathrm{U}} \approx$ the time until ensembles disperse to $\frac{3}{8}$ their saturation value—appears to be a good single choice, but further improvement is possible by tailoring AST to the target location and the initial condition.

4. An attractive alternative to thresholding is *optimizing* some functional of the ensemble severity distribution designed to favor both high extremes and wide spread. We have found a suitable functional in *thresholded entropy* (TE), the expected information contained in that part of the ensemble's severity distribution exceeding the pre-selected threshold. Optimization-based AST rules open the door to using Bayesian optimization strategies to home in on the COASTs adaptively during an actual rare event sampling algorithm, avoiding the exhaustive grid searches we have performed here.

There are many important avenues of research indicated by the present study, both methodology-oriented and science-oriented. On the algorithmic front, it remains to be seen whether thresholded entropy succeeds at matching tail statistics in general systems, but the consistency across different targets within the QG model is encouraging. We suspect that *some* objective function over distributions is broadly applicable. Furthermore, the *shape* of perturbations is a possibly very important lever on the potency of perturbations, acting in concert with their timing. While we limited our present study to a two-dimensional perturbation space based on linearized dynamics about a baroclinically unstable background flow, a natural extension would be to use flow-dependent singular vectors as in operational weather forecasting. By design, they effect faster ensemble spread in the small-perturbation regime; however, it must be checked if their advantages carry into the finite-amplitude regime needed for effective rare event sampling. Computational tools such as adjoints, especially in novel machine learning models, invite the use of gradient-based optimization (Wang et al., 2020; Vonich and Hakim, 2024).

Intriguing dynamical questions also arise from the latitude dependence of the COAST, which can be seen as a predictability index tailored to extremes: how do the physical parameters such as topography, rotation rate, and the spatial domain affect COAST? Is the effect entirely explainable through the extreme value statistics, as we have speculated, or can two similarly shaped tails belie extremely different COAST behavior? These questions merit further parameter exploration, both within and beyond the quasigeostrophic framework. We expect to draw insight from recent theoretical advances relating extreme value theory to the geometry of chaotic attractors (Lucarini et al., 2016).

In summary, our work makes empirical progress on important theoretical and algorithmic questions regarding the probabilities of the most extreme weather events. Perturbed ensemble forecasts of individual weather events are distinct from the climatological distribution, but here we have given quantitative evidence for a relationship between the two—so long as the

perturbations are well-timed – that can be exploited for efficient risk analysis via judicious perturbed simulations. Our work has elucidated what it means to be "well-timed", and furthermore provided quantitative optimization criteria for perturbation timing. Only with this basic pre-requisite information on what to optimize, should we proceed to invest effort into optimizing efficiently.

Code availability. The code to generate all results, including simulation, statistical analysis, and plotting, is available at the Zenodo repository COAST (justinfocus12, 2025). J.F. is happy to provide guidance on use and extension of the code upon request.

Appendix A: Langevin Model

The schematic in Fig.1 comes from Langevin dynamics, consisting of a single particle moving in one dimension with position $X(t)$ and momentum $Y(t)$ subject to a potential gradient force, friction, and stochastic Gaussian white-noise forcing $W(t)$:

$$dX(t) = \frac{1}{m} Y(t)\, dt \tag{A1}$$

$$dY(t) = \Big[-V'(X(t)) - \gamma Y(t) \Big]\, dt + \sigma\, dW(t) \tag{A2}$$

where the potential function $V(x)$ has a quadratic core and logarithmic wings, (A3)

$$V(x) = \begin{cases} \frac{\alpha+1}{\beta}\left(\log(\epsilon) + \frac{(x/\epsilon)^2 - 1}{2} \right) & |x| \le \epsilon \\ \frac{\alpha+1}{\beta} \log|x| & |x| > \epsilon, \end{cases} \tag{A4}$$

which leads to a heavy-tailed (in x) steady-state probability density $p(x,y) \propto \exp\big[-\beta(V(x) + \frac{y^2}{2m}) \big] \sim |x|^{-(\alpha+1)}$ for large $|x|$. Constant parameters are $\gamma = 0.05$ for friction, $m = 1.2$ for mass, $\sigma = 0.005$ for stochastic forcing strength, $\epsilon = 0.25$ for the extent of the quadratic core of the potential, $\alpha = 3.1$ which sets the tail weight, and $\beta = 2m\gamma/\sigma^2$ which is the inverse temperature.

Author contributions. Justin Finkel formulated the initial study, carried out numerical computations, and wrote the initial draft. Paul O'Gorman and Justin Finkel both contributed to refining the methodology and substantially revising the manuscript.

Competing interests. The authors declare no competing interests relevant to this study.

Acknowledgements. We thank Glenn Flierl, Andre Souza, and Talia Tamarin-Brodsky for helpful discussions and advice on theoretical and computational aspects of this work. This research is part of the MIT Climate Grand Challenge on Weather and Climate Extremes. Support was provided by Schmidt Sciences. Computations were performed on the MIT Engaging cluster.

References

Au, S.-K. and Beck, J. L.: Estimation of small failure probabilities in high dimensions by subset simulation, Probabilistic Engineering Mechanics, 16, 263–277, https://doi.org/https://doi.org/10.1016/S0266-8920(01)00019-4, 2001.

Baars, S., Castellana, D., Wubs, F., and Dijkstra, H.: Application of adaptive multilevel splitting to high-dimensional dynamical systems, Journal of Computational Physics, 424, 109 876, https://doi.org/https://doi.org/10.1016/j.jcp.2020.109876, 2021.

Berner, J., Fossell, K. R., Ha, S.-Y., Hacker, J. P., and Snyder, C.: Increasing the Skill of Probabilistic Forecasts: Understanding Performance Improvements from Model-Error Representations, Monthly Weather Review, 143, 1295 – 1320, https://doi.org/10.1175/MWR-D-14-00091.1, 2015.

Bloin-Wibe, L., Noyelle, R., Humphrey, V., Beyerle, U., Knutti, R., and Fischer, E.: Estimating return periods for extreme events in climate models through Ensemble Boosting, EGUsphere, 2025, 1–40, https://doi.org/10.5194/egusphere-2025-525, 2025.

Boulaguiem, Y., Zscheischler, J., Vignotto, E., van der Wiel, K., and Engelke, S.: Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks, Environmental Data Science, 1, e5, https://doi.org/10.1017/eds.2022.4, 2022.

Bourlioux, A. and Majda, A. J.: Elementary models with probability distribution function intermittency for passive scalars with a mean gradient, Physics of Fluids, 14, 881–897, https://doi.org/10.1063/1.1430736, 2002.

Coles, S.: An introduction to statistical modeling of extreme values, Springer Series in Statistics, Springer, 1 edn., ISBN 978-1-85233-459-8, https://doi.org/10.1007/978-1-4471-3675-0, 2001.

Cérou, F. and Guyader, A.: Adaptive Multilevel Splitting for Rare Event Analysis, Stochastic Analysis and Applications, 25, 417–443, https://doi.org/10.1080/07362990601139628, 2007.

Diaconescu, E. P. and Laprise, R.: Singular vectors in atmospheric sciences: A review, Earth-Science Reviews, 113, 161–175, https://doi.org/https://doi.org/10.1016/j.earscirev.2012.05.005, 2012.

Farrell, B. F. and Ioannou, P. J.: Generalized Stability Theory. Part I: Autonomous Operators, Journal of Atmospheric Sciences, 53, 2025 – 2040, https://doi.org/10.1175/1520-0469(1996)053<2025:GSTPIA>2.0.CO;2, 1996a.

Farrell, B. F. and Ioannou, P. J.: Generalized Stability Theory. Part II: Nonautonomous Operators, Journal of Atmospheric Sciences, 53, 2041 – 2053, https://doi.org/10.1175/1520-0469(1996)053<2041:GSTPIN>2.0.CO;2, 1996b.

Finkel, J. and O'Gorman, P. A.: Rare event sampling for moving targets: extremes of temperature and daily precipitation in a general circulation model, https://arxiv.org/abs/2508.13120, 2025.

Finkel, J. and O'Gorman, P. A.: Bringing Statistics to Storylines: Rare Event Sampling for Sudden, Transient Extreme Events, Journal of Advances in Modeling Earth Systems, 16, e2024MS004 264, https://doi.org/https://doi.org/10.1029/2024MS004264, e2024MS004264 2024MS004264, 2024.

Finkel, J., Gerber, E. P., Abbot, D. S., and Weare, J.: Revealing the Statistics of Extreme Events Hidden in Short Weather Forecast Data, AGU Advances, 4, e2023AV000 881, https://doi.org/https://doi.org/10.1029/2023AV000881, e2023AV000881 2023AV000881, 2023.

Fischer, E. M., Beyerle, U., Bloin-Wibe, L., Gessner, C., Humphrey, V., Lehner, F., Pendergrass, A. G., Sippel, S., Zeder, J., and Knutti, R.: Storylines for unprecedented heatwaves based on ensemble boosting, Nature Communications, 14, 4643, https://doi.org/10.1038/s41467-023-40112-4, 2023.

Gálfi, V. M., Bódai, T., and Lucarini, V.: Convergence of Extreme Value Statistics in a Two-Layer Quasi-Geostrophic Atmospheric Model, Complexity, 2017, 5340 858, https://doi.org/10.1155/2017/5340858, 2017.

Gessner, C.: Physical storylines for very rare climate extremes, Ph.D. thesis, ETH Zurich, 2022.

Gessner, C., Fischer, E. M., Beyerle, U., and Knutti, R.: Very Rare Heat Extremes: Quantifying and Understanding Using Ensemble Reinitialization, Journal of Climate, 34, 6619 – 6634, https://doi.org/10.1175/JCLI-D-20-0916.1, 2021.

Ghil, M., Yiou, P., Hallegatte, S., Malamud, B. D., Naveau, P., Soloviev, A., Friederichs, P., Keilis-Borok, V., Kondrashov, D., Kossobokov, V., Mestre, O., Nicolis, C., Rust, H. W., Shebalin, P., Vrac, M., Witt, A., and Zaliapin, I.: Extreme events: dynamics, statistics and prediction, Nonlinear Processes in Geophysics, 18, 295–350, https://doi.org/10.5194/npg-18-295-2011, 2011.

Giorgini, L. T., Deck, K., Bischoff, T., and Souza, A.: Response Theory via Generative Score Modeling, Phys. Rev. Lett., 133, 267 302, https://doi.org/10.1103/PhysRevLett.133.267302, 2024.

Haidvogel, D. B. and Held, I. M.: Homogeneous Quasi-Geostrophic Turbulence Driven by a Uniform Temperature Gradient, Journal of Atmospheric Sciences, 37, 2644 – 2660, https://doi.org/10.1175/1520-0469(1980)037<2644:HQGTDB>2.0.CO;2, 1980.

Huang, W. K., Stein, M. L., McInerney, D. J., Sun, S., and Moyer, E. J.: Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (GEV) distributions, Advances in Statistical Climatology, Meteorology and Oceanography, 2, 79–103, https://doi.org/10.5194/ascmo-2-79-2016, 2016.

Huser, R. and Wadsworth, J. L.: Advances in statistical modeling of spatial extremes, WIREs Computational Statistics, 14, e1537, https://doi.org/https://doi.org/10.1002/wics.1537, 2022.

Huser, R., Opitz, T., and Wadsworth, J. L.: Modeling of spatial extremes in environmental data science: time to move away from max-stable processes, Environmental Data Science, 4, e3, https://doi.org/10.1017/eds.2024.54, 2025.

Jalbert, J., Farmer, M., Gobeil, G., and Roy, P.: Extremes.jl: Extreme Value Analysis in Julia, Journal of Statistical Software, 109, 1–35, https://doi.org/10.18637/jss.v109.i06, 2024.

John, A., Douville, H., Ribes, A., and Yiou, P.: Quantifying CMIP6 model uncertainties in extreme precipitation projections, Weather and Climate Extremes, 36, 100 435, https://doi.org/https://doi.org/10.1016/j.wace.2022.100435, 2022.

justinfocus12: justinfocus12/COAST: Initial release for submission of BEST COAST paper to NPG, https://doi.org/10.5281/zenodo.17355215, 2025.

Kabir, H. M. D., Khosravi, A., Hosen, M. A., and Nahavandi, S.: Neural Network-Based Uncertainty Quantification: A Survey of Methodologies and Applications, IEEE Access, 6, 36 218–36 234, https://doi.org/10.1109/ACCESS.2018.2836917, 2018.

Kahn, H. and Harris, T. E.: Estimation of particle transmission by random sampling, National Bureau of Standards applied mathematics series, 12, 27–30, 1951.

Leobacher, G. and Pillichshammer, F.: Introduction to quasi-Monte Carlo integration and applications, Springer, 2014.

Lestang, T., Ragone, F., Bréhier, C.-E., Herbert, C., and Bouchet, F.: Computing return times or return periods with rare event algorithms, Journal of Statistical Mechanics: Theory and Experiment, 2018, 043 213, https://doi.org/10.1088/1742-5468/aab856, 2018.

Linz, M., Chen, G., Zhang, B., and Zhang, P.: A Framework for Understanding How Dynamics Shape Temperature Distributions, Geophysical Research Letters, 47, e2019GL085 684, https://doi.org/https://doi.org/10.1029/2019GL085684, e2019GL085684 10.1029/2019GL085684, 2020.

Lorenz, E. N. and Emanuel, K. A.: Optimal Sites for Supplementary Weather Observations: Simulation with a Small Model, Journal of the Atmospheric Sciences, 55, 399 – 414, https://doi.org/10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2, 1998.

Lucarini, V. and Gritsun, A.: A new mathematical framework for atmospheric blocking events, Climate Dynamics, 54, 575–598, https://doi.org/10.1007/s00382-019-05018-2, 2020.

Lucarini, V., Faranda, D., de Freitas, J. M. M., Holland, M., Kuna, T., Nicol, M., Todd, M., Vaienti, S., et al.: Extremes and recurrence in dynamical systems, John Wiley & Sons, 2016.

Lucente, D., Rolland, J., Herbert, C., and Bouchet, F.: Coupling rare event algorithms with data-based learned committor functions using the analogue Markov chain, Journal of Statistical Mechanics: Theory and Experiment, 2022, 083 201, https://doi.org/10.1088/1742-5468/ac7aa7, 2022.

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Elms, J., Harrington, P., Kashinath, K., Kurth, T., North, J., OBrien, T., Pritchard, M., Pruitt, D., Risser, M., Subramanian, S., and Willard, J.: Huge Ensembles Part I: Design of Ensemble Weather Forecasts using Spherical Fourier Neural Operators, https://arxiv.org/abs/2408.03100, 2024a.

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Harrington, P., Kashinath, K., Kurth, T., North, J., OBrien, T., Pritchard, M., Pruitt, D., Risser, M., Subramanian, S., and Willard, J.: Huge Ensembles Part II: Properties of a Huge Ensemble of Hindcasts Generated with Spherical Fourier Neural Operators, https://arxiv.org/abs/2408.01581, 2024b.

Maiocchi, C. C., Lucarini, V., Gritsun, A., and Sato, Y.: Heterogeneity of the attractor of the Lorenz '96 model: Lyapunov analysis, unstable periodic orbits, and shadowing properties, Physica D: Nonlinear Phenomena, 457, 133 970, https://doi.org/https://doi.org/10.1016/j.physd.2023.133970, 2024.

Mohamad, M. A. and Sapsis, T. P.: Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems, Proceedings of the National Academy of Sciences, 115, 11 138–11 143, https://doi.org/10.1073/pnas.1813263115, 2018.

Neelin, J. D., Lintner, B. R., Tian, B., Li, Q., Zhang, L., Patra, P. K., Chahine, M. T., and Stechmann, S. N.: Long tails in deep columns of natural and anthropogenic tropospheric tracers, Geophysical Research Letters, 37, https://doi.org/https://doi.org/10.1029/2009GL041726, 2010.

Noyelle, R.: Statistical and dynamical aspects of extreme heatwaves in the mid-latitudes, Theses, Université Paris-Saclay, https://hal.science/tel-04632646, 2024.

O'Gorman, P. A. and Schneider, T.: Scaling of Precipitation Extremes over a Wide Range of Climates Simulated with an Idealized GCM, Journal of Climate, 22, 5676 – 5685, https://doi.org/10.1175/2009JCLI2701.1, 2009.

Panetta, R. L.: Zonal Jets in Wide Baroclinically Unstable Regions: Persistence and Scale Selection, Journal of Atmospheric Sciences, 50, 2073 – 2106, https://doi.org/10.1175/1520-0469(1993)050<2073:ZJIWBU>2.0.CO;2, 1993.

Pavliotis, G. A.: Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations, vol. 60, Springer, 2014.

Penland, C. and Magorian, T.: Prediction of Niño 3 Sea Surface Temperatures Using Linear Inverse Modeling, Journal of Climate, 6, 1067 – 1076, https://doi.org/10.1175/1520-0442(1993)006<1067:PONSST>2.0.CO;2, 1993.

Pickering, E., Guth, S., Karniadakis, G. E., and Sapsis, T. P.: Discovering and forecasting extreme events via active learning in neural operators, Nature Computational Science, 2, 823–833, https://doi.org/10.1038/s43588-022-00376-0, 2022.

Pons, F. M. E., Yiou, P., Jézéquel, A., and Messori, G.: Simulating the Western North America heatwave of 2021 with analogue importance sampling, Weather and Climate Extremes, 43, 100 651, https://doi.org/https://doi.org/10.1016/j.wace.2024.100651, 2024.

Qi, D. and Majda, A. J.: Predicting fat-tailed intermittent probability distributions in passive scalar turbulence with imperfect models through empirical information theory, Communications in Mathematical Sciences, 14, 1687–1722, 2016.

Qi, D. and Majda, A. J.: Predicting extreme events for passive scalar turbulence in two-layer baroclinic flows through reduced-order stochastic models, Communications in Mathematical Sciences, 16, 17–51, 2018.

Rackauckas, C.: QuasiMonteCarlo.jl, https://github.com/SciML/QuasiMonteCarlo.jl, accessed: 2025-05-09, 2023.

Ragone, F., Wouters, J., and Bouchet, F.: Computation of extreme heat waves in climate models using a large deviation algorithm, Proceedings of the National Academy of Sciences, 115, 24–29, https://doi.org/10.1073/pnas.1712645115, 2018.

Rampal, N., Gibson, P. B., Sherwood, S., Abramowitz, G., and Hobeichi, S.: A Reliable Generative Adversarial Network Approach for Climate Downscaling and Weather Generation, Journal of Advances in Modeling Earth Systems, 17, e2024MS004 668, https://doi.org/https://doi.org/10.1029/2024MS004668, e2024MS004668 2024MS004668, 2025.

Rolland, J.: Collapse of transitional wall turbulence captured using a rare events algorithm, Journal of Fluid Mechanics, 931, A22, https://doi.org/10.1017/jfm.2021.957, 2022.

Saha, A. and Ravela, S.: Statistical-Physical Adversarial Learning From Data and Models for Downscaling Rainfall Extremes, Journal of Advances in Modeling Earth Systems, 16, e2023MS003 860, https://doi.org/https://doi.org/10.1029/2023MS003860, e2023MS003860 2023MS003860, 2024.

Sapsis, T. P.: Output-weighted optimal sampling for Bayesian regression and rare event statistics using few samples, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 476, 20190 834, https://doi.org/10.1098/rspa.2019.0834, 2020.

Sundar, R., Parashar, N., Blanchard, A., and Dodov, B.: TAUDiff: Improving statistical downscaling for extreme weather events using generative diffusion models, https://arxiv.org/abs/2412.13627, 2024.

Tebaldi, C., Armbruster, A., Engler, H. P., and Link, R.: Emulating climate extreme indices, Environmental Research Letters, 15, 074 006, https://doi.org/10.1088/1748-9326/ab8332, 2020.

Thompson, A. F.: Jet Formation and Evolution in Baroclinic Turbulence with Simple Topography, Journal of Physical Oceanography, 40, 257 – 278, https://doi.org/10.1175/2009JPO4218.1, 2010.

Thompson, V., Dunstone, N. J., Scaife, A. A., Smith, D. M., Slingo, J. M., Brown, S., and Belcher, S. E.: High risk of unprecedented UK rainfall in the current climate, Nature Communications, 8, 107, https://doi.org/10.1038/s41467-017-00275-3, 2017.

van den Dool, H. M.: A New Look at Weather Forecasting through Analogues, Monthly Weather Review, 117, 2230 – 2247, https://doi.org/10.1175/1520-0493(1989)117<2230:ANLAWF>2.0.CO;2, 1989.

van Kekem, D. L. and Sterk, A. E.: Wave propagation in the Lorenz-96 model, Nonlinear Processes in Geophysics, 25, 301–314, https://doi.org/10.5194/npg-25-301-2018, 2018.

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., and Ganguly, A. R.: DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, p. 1663–1672, Association for Computing Machinery, New York, NY, USA, ISBN 9781450348874, https://doi.org/10.1145/3097983.3098004, 2017.

Vonich, P. T. and Hakim, G. J.: Predictability Limit of the 2021 Pacific Northwest Heatwave From Deep-Learning Sensitivity Analysis, Geophysical Research Letters, 51, e2024GL110 651, https://doi.org/https://doi.org/10.1029/2024GL110651, e2024GL110651 2024GL110651, 2024.

Wang, Q., Mu, M., and Sun, G.: A useful approach to sensitivity and predictability studies in geophysical fluid dynamics: conditional nonlinear optimal perturbation, National Science Review, 7, 214–223, https://doi.org/10.1093/nsr/nwz039, 2020.

Watt, R. A. and Mansfield, L. A.: Generative Diffusion-based Downscaling for Climate, https://arxiv.org/abs/2404.17752, 2024.

Webber, R. J., Plotkin, D. A., O'Neill, M. E., Abbot, D. S., and Weare, J.: Practical rare event sampling for extreme mesoscale weather, Chaos: An Interdisciplinary Journal of Nonlinear Science, 29, 053 109, https://doi.org/10.1063/1.5081461, 2019.

Yang, Y., Blanchard, A., Sapsis, T., and Perdikaris, P.: Output-weighted sampling for multi-armed bandits with extreme payoffs, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 478, 20210 781, https://doi.org/10.1098/rspa.2021.0781, 2022.

Yiou, P. and Jézéquel, A.: Simulation of extreme heat waves with empirical importance sampling, Geoscientific Model Development, 13, 763–781, https://doi.org/10.5194/gmd-13-763-2020, 2020.

Zuckerman, D. M. and Chong, L. T.: Weighted Ensemble Simulation: Review of Methodology, Applications, and Software, Annual Review of Biophysics, 46, 43–57, https://doi.org/10.1146/annurev-biophys-070816-033834, pMID: 28301772, 2017.

Zuev, K.: Subset Simulation Method for Rare Event Estimation: An Introduction, 2015.