Rare event sampling for moving targets: extremes of temperature and daily precipitation in a general circulation model

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

[1]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology

Key Points:

- Extreme weather risk is highly uncertain, but can be estimated more accurately by targeted rare event sampling.
- Rare event algorithms are challenged by short time scales of weather events which limit ensemble diversity.
- Optimally timed perturbations enable sped-up probability estimates of precipitation and heat extremes in an aquaplanet climate model.

Corresponding author: Justin Finkel, ju26596@mit.edu

Abstract

Extreme weather events epitomize high cost: to society through their physical impacts, and to computer servers that simulate them to assess risk and advance physical understanding. It costs hundreds of simulation years to sample a few once-per-century events with straightforward model integration, but that cost can be much reduced with *rare event sampling*, which nudges ensembles of simulations to convert moderate events to severe ones, e.g., by steering a cyclone directly through a region of interest. With proper statistical accounting, rare event algorithms can provide quantitative climate risk assessment at reduced cost. But this can only work if ensemble members diverge fast enough. Sudden, transient events characteristic of Earth's midlatitude storm track regions, such as heavy precipitation and heat extremes, pose a particular challenge because they come and go faster than an ensemble can explore the possibilities. Here we extend standard rare event algorithms to handle this challenging case in an idealized atmospheric general circulation model, achieving $\sim 5-10$ times sped-up estimation of long return periods for extremes of surface temperature and daily precipitation (e.g., return periods of 100-150 years from 20 years of simulation). The algorithm, called TEAMS ("trying-early adaptive multilevel splitting"), was developed previously using a toy chaotic system, and relies on a key parameter—the advance split time—which may be estimated based on simple diagnostics of ensemble dispersion rates. The results are promising for accelerated risk assessment across a wide range of physical hazards using more realistic and complex models with acute computational constraints.

Plain Language Summary

Climate hazards are largely felt not through global mean temperature, but through extreme weather events, which are dangerous not only for their physical severity but also for their rarity: by definition, they are very difficult to anticipate and prepare for. The same characteristic makes risk assessment a very hard statistical problem. Numerical simulations can be used to augment small sample sizes, but at great computational cost. Rare event algorithms offer a novel way to "steer" simulations towards the extremes to do targeted risk assessment at reduced cost, but this can be challenging when the events under study are transient in nature, such as passing rainstorms and heat extremes in Earth's midlatitude region. This paper presents a successful application of a rare event algorithm to such transient extremes in an idealized model of Earth's atmospheric circulation, building on previously published results with a simpler toy model of spatial chaos. The core of the method is to select the right time to perturb the simulations, and the fact that the method generalizes is a promising sign that it can scale to even more complex, realistic models.

1 Introduction

The highest-impact extreme weather events are those that occur so seldom as to catch communities—cities, ecologies, and scientists alike—surprised and unprepared (Sillmann et al., 2017). Even with physically accurate numerical models capable of simulating extremes, running them long enough to collect ample statistics can be prohibitive. A key innovation to close this gap is *rare event sampling*, a protocol which steers ensembles of simulations towards the extremes by repeated perturbation, pruning, and cloning steps, all while keeping track of the bias introduced to correct for it in statistical estimation. Originally developed for nuclear physics simulation (Kahn & Harris, 1951), rare event algorithms have been specialized and developed for molecular dynamics (Zuckerman & Chong, 2017), reliability engineering (Huang et al., 2016; Sapsis, 2020; Uribe et al., 2021; Zhang et al., 2022), and climate science (e.g., Ragone et al., 2018; Wouters & Bouchet, 2016; Webber et al., 2019). Rare event algorithms are attractive for being agnostic to the model: importantly, they can operate on models grounded in physics and potentially could also be applied to faster, data driven models with the alluring possibility of generating abundant extreme events at will (Mahesh et al., 2024, 2024).

Yet there remain some methodological roadblocks to the broad deployment of rare event algorithms across different models and different rare events. This paper addresses one such roadblock: a timescale overlap between the event of interest and the ensemble dispersion which is crucially necessary to sample different hypothetical versions of the event. There is no such overlap for long-lasting, spatially extended events such as hot or rainy *seasons*—defined by large *seasonal mean* temperature or precipitation amplitudes. Such events are already a successful application for rare event algorithms (Ragone et al., 2018; Wouters & Bouchet, 2016), as multiple successive rounds of ensemble splitting can fit into a single season and achieve extreme anomalies by essentially chaining together a sequence of moderate anomalies. But transient events of much shorter duration don't yield so easily; naïvely applying the same perturbation protocol simply results in disappointing replication of the same moderate extreme again and again, without meaningful exploration into the far tails (Lestang et al., 2020; Rolland, 2022; Finkel & O'Gorman, 2024). This is a major limitation given that transient cyclones and antiyclones can bring heavy rain and temperature extremes that are among the most impactful extreme events for society.

We developed a simple remedy to this problem, drawing inspiration from *ensemble boosting* (Gessner et al., 2021; Gessner, 2022), namely to perturb simulations in advance of the event. Ensemble boosting, as originally formulated, does not assign probabilities but only generates "sto-

rylines" (until more recent developments in Bloin-Wibe et al. (2025) and Finkel and O'Gorman (2025)), but we augmented boosting with an acceptance/rejection step from reliability engineering (Au & Beck, 2001) to retain statistics, and demonstrated this on the Lorenz-96 system in Finkel and O'Gorman (2024). Lorenz-96 is relatively simple model of spatiotemporal chaos, but it captures the essence of baroclinic waves and has been a helpful benchmark for data assimilation, which presents similar challenges as sampling algorithms. The resulting algorithm, TEAMS ("trying-early adaptive multilevel splitting"), introduces a key hyperparameter, the *advance split time*, which determines when to split the simulation relative to the event for an optimal balance of exploration (with high risk of rejection) and exploitation (with low risk of rejection but limited rewards). Our main contribution here is to demonstrate a successful use of TEAMS on an actual climate model, albeit an idealized one, to sample short-timescale events, namely high surface temperatures and daily precipitation rates.

This paper is organized as follows. Section 2 briefly specifies the physical model, a general circulation model (GCM) in an aquaplent configuration, emphasizing two modifications of reduced resolution for computational efficiency and the addition of stochastic parameterization. Section 3 outlines the rare event algorithm TEAMS, emphasizing the most recent modifications of how rejection is handled and the halting criteria. Section 4 shows the results of applying TEAMS: efficiency gains in calculating long return periods (100 years and longer), and the generation of corresponding dynamical samples. Section 5 concludes with a summary and outlook on further avenues of development.

2 The physical model

We use an idealized GCM based on the GFDL spectral model and similar to that developed in Frierson et al. (2006) with slight modifications as in O'Gorman and Schneider (2008). A spectral dynamical core integrates the primitive equations, with a lower boundary condition consisting of a slab ocean (aquaplanet) that is shallow, well-mixed, and energy-conserving (not fixed-temperature). Insolation is fixed to an average distribution, with no seasonal or diurnal cycle. A two-stream gray radiation scheme is used, with a prescribed distribution of longwave optical depth. We turn off convection parameterization, so that condensation of water vapor occurs only at the large scale (grid box size), as was found to be adequate for midlatitudes by Frierson et al. (2006). Turbulent diffusivities are smoothed in time following Anderson et al. (2004).

We make two further modifications for this rare event sampling demonstration. To enable computational efficiency, we reduced the resolution to T21 in the horizontal, with six σ-levels in the vertical (half levels at $\sigma = 0.0, 0.0343, 0.15, 0.4, 0.7, 0.966, 1.0$), and a 40-minute timestep. We also present some limited results at a higher horizontal resolution of T42 and 30 vertical levels. To induce variability between ensemble members, we implemented a stochastic parameterization scheme known as stochastically perturbed parameterized tendencies (SPPT) that was developed in numerical weather prediction to enhance ensemble spread to more likely capture the observed evolution (Palmer et al., 2009; Berner et al., 2009, 2015), and which rare event sampling can use to discover unlikely paths towards extremes. Our implementation of SPPT closely follows the specification in Palmer et al. (2009), which contains further details and background. In brief, SPPT modifies the total parameterized tendencies of horizontal winds, humidity, and temperature by a multiplicative factor of $1+r_{\mathrm{SPPT}}(x, y, z, t)$ at every timestep, where $r_{\mathrm{SPPT}}(x, y, z, t)$ is a random spatiotemporal pattern whose spherical harmonic modes each evolve as an independent red noise process. The total parameterized tendencies include contributions from large-scale condensation, vertical turbulent diffusion, and radiation. Key tunable parameters are the noise amplitude σ_{SPPT} and the characteristic length and time scales L_{SPPT} and τ_{SPPT}. To prevent unrealistically large fluctuations, r_{SPPT} is clipped to the range $(-2\sigma_{\mathrm{SPPT}}, 2\sigma_{\mathrm{SPPT}})$ at every timestep. Sensitivity analysis led us to select $\sigma_{\mathrm{SPPT}} = 0.3$, $L_{\mathrm{SPPT}} = 500$ km, and $\tau_{\mathrm{SPPT}} = 6$ hours for this study, quite similar to the moderate-amplitude experiments in Palmer et al. (2009).

We use this computationally efficient GCM because it accommodates the large ensemble sizes and parameter tuning experiments needed for development and testing of rare-event sampling strategies. Our aim is to demonstrate a novel methodology more than a particular scientific conclusion, and for this purpose a lower rung on the model hierarchy (Held, 2005) take on greater value. The same idealizations (such as zonally symmetric boundary conditions) that make this model attractive for extensive parameter sweeps, as in O'Gorman and Schneider (2008) and O'Gorman and Schneider (2009), also make it well-suited for rare event algorithm development. At the same time, even the coarse model is physically realistic enough that the insights learned here should transfer to more realistic models.

Fig. 1 displays some characteristics of the surface temperature and precipitation fields produced by the GCM once it reaches statistical equilibrium after a spinup period. Throughout the paper, surface temperature refers to the surface air temperature evaluated at the lowest model level. Outputs from the GCM are six-hourly; temperature is instantaneous (noting there is no diurnal cycle) and precipitation is averaged over the previous day. Despite the idealized setup and coarse

resolution, the baroclinic waves of Earth's midlatitude storm track and associated precipitation and temperature variability are clearly visible in the model fields (Fig. 1a,b), which grow and decay over synoptic $\sim$ 5-day timescales (indicated by the Hovmöller diagrams in Fig. 1c,d). Our aim is to characterize—using rare event sampling—the extreme, local fluctuations in these fields at the storm track's center. We therefore fix a target latitude of 45°N and a target longitude of 180°E, taking the field value in a single grid cell ($\sim 6°$) as the target variable. The choice of longitude is arbitrary due to the model's zonal homogeneity, but fixing a longitude simplifies the event definition and would be necessary anyway in Earth system models with zonal asymmetries. Still, we take advantage of zonal homogeneity in computing "ground truth" statistics from long simulation by pooling together eleven longitudinal rotations in 30° increments for more stable estimation with twelve times the data. Fig. 1(e,f) displays the long-term climate statistics of precipitation and temperature at the target location, revealing $\sigma_{\text{SPPT}} \approx 0.3$ to be near the upper limit of noise level that still avoids disrupting the deterministic model's statistics too severely. These results are based on a long run of 36,500 days (100 years, or 1200 years including longitudinal rotation) after spinup, which we refer to as a direct numerical simulation (DNS) and which will be used for validation. The data used for initializing TEAMS, on the other hand, is branched from the long DNS after spinup and integrated independently, with a different seed for each run of TEAMS, in order to avoid data leakage (see "ancestor initialization" in the algorithm described in section 3).

3 The TEAMS algorithm

Let us briefly describe the TEAMS algorithm, following Finkel and O'Gorman (2024). Along the way we delineate between generic parameter choices and those made in this study to target local temperature and precipitation extremes in the GCM. Readers interested primarily in the sampling results can skip to Section 4.

1. Ancestor initialization: Sample N initial conditions $\{X_1(0), X_2(0), ..., X_N(0)\}$ from the distribution of interest, denoted ρ_0. For us, ρ_0 is the distribution at statistical steady state, i.e., the limiting distribution of a very long GCM simulation. Other applications might restrict the initial conditions to specific phases of oscillation (e.g., neutral El Niño conditions) or, if a seasonal cycle is present, specific dates (e.g., June 1 conditions). For our study, we can extract the $X_n(0)'s$ as snapshots from a direct numerical simulation (DNS), which is branched from the DNS used for validation by changing the random seed for SPPT after spinup. Consecutive ancestral initial conditions are separated by a gap of $T = 60$

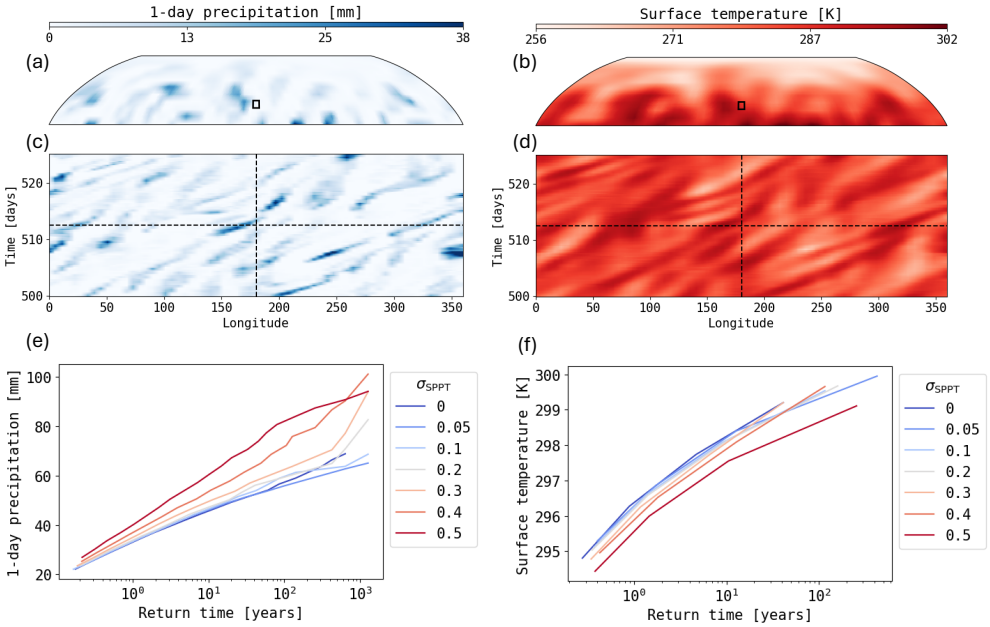

Figure 1. Simulated precipitation and surface temperature fields and their return levels. After a spin-up of 500 days, the aquaplanet GCM produces physically plausible large-scale storm track dynamics: a sequence of extratropical cyclones and anticyclones bringing packets of precipitation (a) and temperature fluctuations (b), propagating eastward with lifetimes of $\sim$ 5 days (Hovmöller diagrams in c and d). We select a target region (one grid cell marked by a black square in (a,b)) to fall at 45°N, near the latitude of maximum mean precipitation, and a longitude of 180°E (which is arbitrary because climatological statistics are zonally uniform). Horizontal and vertical dashed lines in (c,d) indicate the timing of the snapshot and the target longitude. Panels e,f show return level vs. return period plots of both targets, local precipitation and temperature, for a range of values of the SPPT forcing strength σ_{SPPT}. The return levels vary only moderately for $\sigma_{\text{SPPT}} \lesssim 0.3$ and start deviating substantially for larger values, which is why we adhere to $\sigma_{\text{SPPT}} = 0.3$ in panels (a-d) and hereafter.

days; in other words, $X_i(0) = X_{\text{DNS}}(iT)$, where the clock for X_{DNS} starts after spinup and the timestamps are all reset to zero for notational convenience, utilizing the model's autonomous dynamics. The gap helps each successive ancestor lose the memory of the previous one and become more independent, which tends to make the results more stable, i.e. lower-variance, even though the ancestors need not be strictly independent (only identically distributed). Section 4.3 will demonstrate $T = 60$ days is long enough for independence.

2. Ancestor simulation: Run the dynamics forward for a time horizon T from each ancestral initial condition, creating the *ancestral trajectories* $\{X_n(t) : 1 \le n \le N, 0 \le t \le T\}$. For us, this just means extracting segments of the branched-off DNS, and we use the same $T = 60$ days here as the time gap between ancestors. Assign each ancestor a probability weight $W_n = 1$. Furthermore, initialize a set of *active members*

$$\mathcal{A} = \{1, \dots, N\} =: \{a_1, \dots, a_A\} \tag{1}$$

with a size $A = N$, which will be modified by repeated culling and replenishment in following steps. Also initialize an empty list of *severity levels* $\mathcal{S} = []$, which will grow in the following steps.

3. Culling: Rank the active ensemble members $a \in \mathcal{A}$ by their *severity*, $S_a = S(X_a)$ defined as the peak value over time of the *intensity* $R_a(t) = R(X_a(t))$ which defines the target variable of interest. In our case, our outputs are six-hourly and $R(X_a(t))$ is the precipitation (averaged over the preceding day) or surface temperature (measured at a single six-hourly snapshot) at the target grid box indicated in Fig. 1. Choose a number $K < A$ and cull the the K least-extreme active members. We choose $K = \frac{1}{2}N$, but one could also set K as a constant number (commonly $K = 1$, as in Finkel and O'Gorman (2024)) or some other fixed fraction of N (in engineering applications, the related "subset simulation" algorithm commonly culls aggressively with $K \sim 0.9N$ (Au & Beck, 2001)). At this point, by design, the K-th smallest severity s has an estimated exceedance probability of $(N - K)/N$ (for us, $1/2$). Append the list of severity levels, $\mathcal{S} \leftarrow \mathcal{S} \cup [s]$. Remove the culled members from the active set, reducing its size to $A-K$, re-index its members accordingly to $\mathcal{A} = \{a_1, \dots, a_{A-K}\}$, and reset the size A to $A - K$.

4. Cloning: Shuffle the active members in a random order, called the "parent queue". For the first parent a in the queue identify the earliest timestep (in six-hourly outputs) that $R_a(t) > s$ and call this time t_a^s At an *earlier* time $t_a^s - \delta$, spawn a new "child" $\widetilde{X}$ which shares its parent's history up until $t_m^s - \delta$, but then gets perturbed by use of a new seed for random

number generation in the stochastic parameterization scheme—and thereafter diverges from its parent. δ is the key *advance split time* parameter, which we vary systematically in this study from 0 to 20 days. The next step depends on whether the child's severity exceeds s:

(a) If the child's severity exceeds s, we call this "success" and officially admit the child into the active population: $X_{a_{A+1}} = \widetilde{X}$, with the same probability weight as its parent. To maintain a constant total probability weight in the active population, adjust all active weights by the same factor: $W_a \leftarrow \frac{A}{A+1} W_a$ for all $a \in \mathcal{A}$. Finally, increment A to $A + 1$.

(b) Otherwise, in case the child's severity fails to exceed s (which might happen, because the split happens before the parent's first threshold crossing; see Fig. 1 in Finkel and O'Gorman (2024)), discard the child completely (formally, set its weight to zero) and move to the next parent in the queue to clone it in the same way.

Keep cycling through the queue until either the active set is fully replenished to a size $A = N$ (the original population size) with K new successful children, or the total number M of simulations (including ancestors, discarded members, and inactive members) exhausts a pre-determined computational budget, $M = M_{\max}$. For our main experiments, we set $M_{\max} = 150$.

5. Iteration: Repeatedly perform step 3 starting with the active population, resulting in a higher level s, followed by step 4 on the sub-ensemble exceeding s.
6. Termination: halt the algorithm once the number of severity levels in $\mathcal{S}$ exceeds a pre-set number (in our case, 20), or the total number M of simulations reaches the aforementioned budget $M_{\max}$.
7. Post-analysis: For any observable of interest expressible as $F(X)$, where X denotes a random variable comprising a whole trajectory $\{X(t) : 0 < t \leq T\}$ with $X(0)$ drawn from ρ_0, and F is a generic functional, estimate its expectation as

$$\widehat{F} = \frac{\sum_{m=1}^{M} W_m F(X_m)}{\sum_{m=1}^{M} W_m}. \tag{2}$$

The denominator is always equal to N. In particular, for any given severity s, an estimate $\widehat{\mathbb{P}}\{S > s\}$ for its exceedance probability is found by defining $F(X) := \mathbb{I}\{S(X) > s\}$ in the formula above, where $\mathbb{I}$ is the indicator function (one if its argument is true, zero otherwise). The corresponding *return period* $\tau(s)$—the average time between consecutive exceedances, using a Poisson process statistical model—is estimated following Lestang

et al. (2018) as

$$\widehat{\tau}(s) = -\frac{T}{\log[1 - \widehat{\mathbb{P}}\{S > s\}]}, \tag{3}$$

where T is the time horizon.

This version of TEAMS mostly follows the version in Finkel and O'Gorman (2024), but differs in two substantial ways. First, in step 4, the previous version of TEAMS would allow parents to stand in for their failed children, and raise the level after K cloning attempts even if they all fail, whereas the new version refuses to raise the level before children alone repopulate the ensemble. Heuristically, the new version is more like mastery-based learning (Winget & Persky, 2022), wherein students only advance after demonstrating mastery even if it takes a longer time with remedial coursework. Even if the levels don't advance as high this way, it ensures that the levels reached are more thoroughly sampled and avoids overextending an "aging" ensemble beyond its means. Of course, this risks stagnation at a single level that is impossible to overcome. To cut our losses, we impose a lean budget of $M_{\max} = 150$ as the second major difference from Finkel and O'Gorman (2024), where the budget was 1024 and in practice was rarely reached because of a second "diversity" criterion that is not used here. We have found this version to give more reliable speedup at shorter return periods with reasonable costs, and to reduce the chance of underestimating return values in a given TEAMS run ("apparent bias"), which was critical for extending this algorithm from a toy model (Lorenz-96) to a GCM.

In the sense of repeatedly spawning descendants until success (or computational budget overrun), our new version resembles "anticipated AMS" (Rolland, 2022). However, in another important sense, anticipated AMS still differs by splitting ancestors when $R_a(t)$ crosses a lower threshold than s, rather than at a fixed advance split time. This would not work on precipitation, which rises from zero to peak values more rapidly than ensemble members can diverge; hence, the TEAMS strategy of splitting a fixed time in advance.

The advance split time (AST), δ, is a crucial hyperparameter underlying TEAMS which must be chosen in a cheap and reliable way in order to scale TEAMS successfully to realistic GCMs. In section 4.3, we estimate the proposed AST from Finkel and O'Gorman (2024), namely the time until a perturbed ensemble disperses to a fraction $3/8$ of its saturation dispersion, using a branching procedure. But first, we will present results from TEAMS across a range of ASTs to demonstrate its ability to sample extreme events in the GCM.

4 Results

4.1 TEAMS performance

We ran TEAMS with a range of advance split times $\delta \in \{0, 4, 6, 8, 10, 12, 14, 16, 20, 24\}$ days. Fig. 2 displays the resulting estimates of return level vs. return period for both targets of local precipitation (left), with $\delta = 10$ days, and temperature (right), with $\delta = 12$ days, which are selected as optimal values based on sensitivity analysis to be presented in Sec. 4.2. The main sequence of experiments used T21 resolution and $N = 16$ ancestors, but as a test of the robustness of δ, we performed two "pivot" experiments about the optimal values: doubling the ancestor pool to $N = 32$, and doubling horizontal resolution to T42 with 30 vertical levels and a time step of 600s.

Our overall assessment of TEAMS is that it speeds up estimation of extreme events relative to DNS by factors of 5-10. Since GCMs are far more expensive than toy models like Lorenz-96, here we focus on the performance of individual runs of TEAMS instead of pooled estimation across many such runs as we did in Finkel and O'Gorman (2024). In Fig. 2, the median return level across TEAMS runs (purple line) is generally very close to the DNS ground truth (black dashed line), indicating that the overall bias is not large. The red bands in Fig. 2 assess reliability by how close to the ground truth one can expect a single TEAMS run to land with 50% probability. Comparing red to gray error bars—the latter coming from DNS, computed with a budget equal to a single TEAMS run—we see a tradeoff between the bulk and the tail. For the default case of $N = 16$ (Fig.2c,d), one run of TEAMS is equivalent to ~ 19 years of DNS in computational cost. TEAMS is less certain in return periods < 19 years than DNS, but provides a good estimate for the range $\sim (19 - 100)$ years (for precipitation) and $\sim (19 - 150)$ years (for temperature), which a 19-year DNS simply cannot estimate. We take the upper range of reliability to be where the error bar starts behaving erratically due to fewer TEAMS runs splitting that many times. TEAMS performs similarly on precipitation and temperature, even though the tails are shaped quite differently: from extreme value theory, precipitation shape parameters often take both positive and negative signs, indicating unbounded or bounded tails (Ragulina & Reitan, 2017), whereas temperature shape parameters tend to be negative (Krakauer, 2024).

Doubling the ancestor pool from $N = 16$ to 32 (Fig. 2a,b) noticeably improves TEAMS' reliability, narrowing the error bars and giving a larger increase in the longest return period. In this case, one TEAMS run is equivalent to just under 40 years of DNS. We find that one run of TEAMS is less certain than DNS for return periods less than 40 years, but provides a good es-

timate for return periods from 40-300 years (for precipitation) and 40-500 years (for temperature), which a 40 year DNS could not estimate. Extreme value theory could be applied to the DNS to extrapolate return values, but this would not generate dynamical samples of events in the same way that TEAMS does.

Doubling the resolution somewhat degrades the extent of the speedup, especially for precipitation, but keeps some of the advantage (Fig. 2e,f). The higher-resolution runs are significantly more expensive: besides doubling horizontal resolution, we also increased vertical levels from 6 to 30 and reduced the timestep from 2400 to 600 seconds, resulting in $\sim (2\times2\times5\times 4 = 80)$-times more expensive simulations. We expect that with further experimentation with population control parameters (such as N, K), it should be possible to improve performance at this and much higher resolutions. The generalizability to a higher resolution shown here, though modest, is enough to draw cautious optimism for the algorithm's scalability.

We can better understand the mechanism for TEAMS' success by examining a few case studies, or "storylines", of events which are mutated from moderate ancestors into extreme descendants. Fig. 3 displays one case study for each target variable (precipitation and temperature), with the same advance split times as used in Fig. 2 (10 and 12 days, respectively). Boosting happens either by amplifying an existing spike, or by materializing a new spike where none existed before. In Fig. 3a, the first cloning (green) mutated the ancestral spike into a smaller spike, but still cleared the threshold ($\sim$ 20 mm/day), whereas the second cloning (yellow) first produced an even smaller spike at $t \approx 25$ but then discovered a new spike at $t \approx 48$. The two subsequent descendants (orange and brown) built further on this second spike, ultimately rising above the ancestor's original score. In Fig. 3c, descendants build on the original spike leading to higher and higher severities. This is a desirable behavior for TEAMS. Metaphorically, "the apple shouldn't fall too far from the tree", or equivalently, subsequent generations should "stand on the shoulders of their predecessors". Shortening the time horizon T might help ensure this behavior, but T must also be long enough for later generations to distinguish themselves. How to quantify the dynamical relationships between parents and children in terms of advance split time is an ongoing research agenda, which might fruitfully be attacked by deterministic optimization strategies, like Newton's Method, in the space of perturbations.

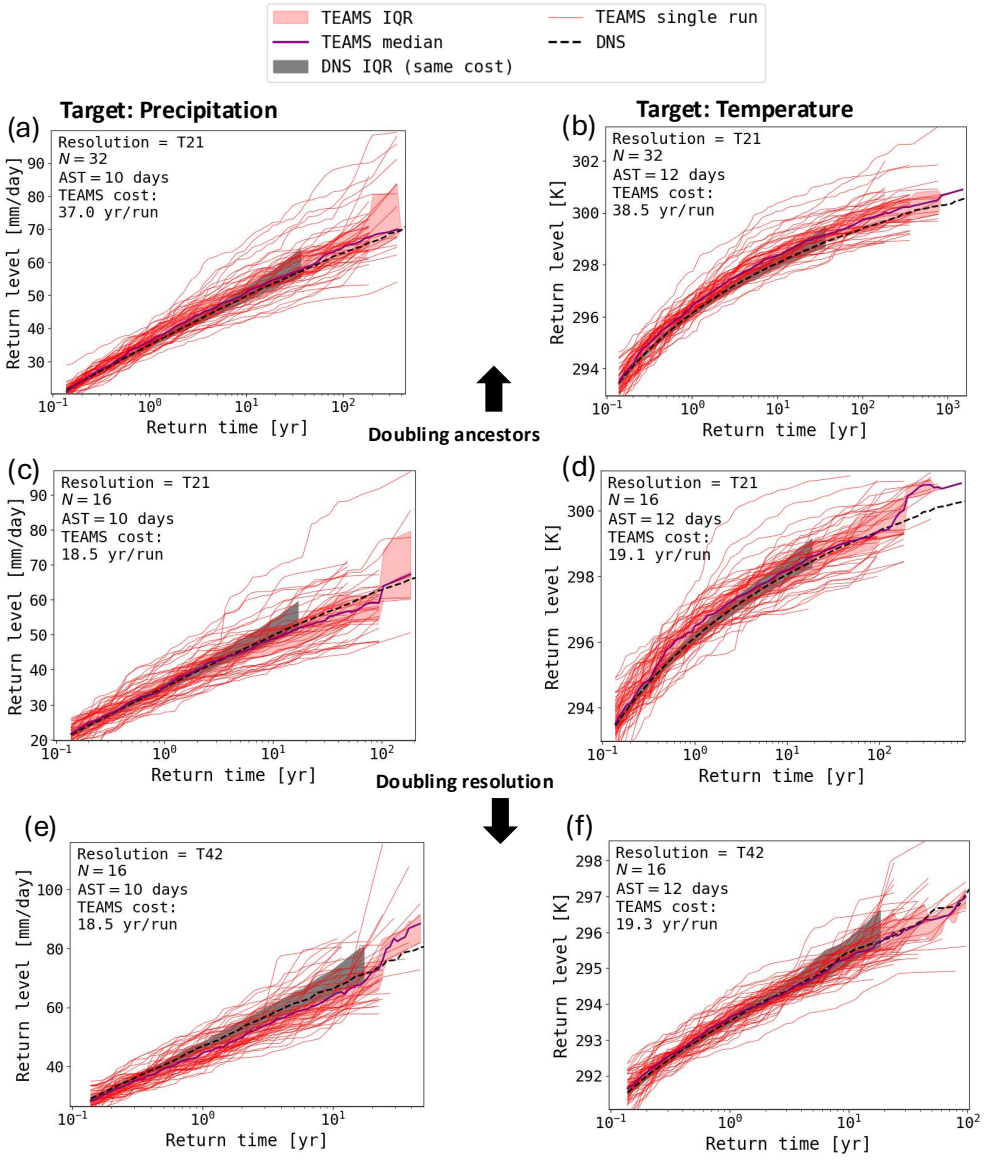

Figure 2. Performance of the rare event algorithm (TEAMS) against the benchmark direct numerical simulation (DNS), as measured by accuracy and uncertainty of return levels for a given computational cost. Target variables are precipitation (left) and surface temperature (right). About a baseline setting of T21 resolution with $N = 16$ ancestors (middle row), we perform two "pivot" experiments: doubling the number of ancestors (top row) and doubling the resolution to T42 (bottom row). All curves are estimates of return level (severity of an event) as a function of return period (the averaged elapsed time between two consecutive events) calculated by different methods: black dashed lines come from a long benchmark DNS, the best estimate of ground truth, and each thin red line comes from a run of TEAMS with a different random seed (48 in total). The purple line and light red band indicate the median and inter-quartile range (25th-75th percentile) of these 48 runs, or somewhat fewer in the far tail to include only those runs that split enough times to estimate the smallest probabilities. For a fair performance comparison, gray error bars show the inter-quartile range of estimates derived from random subsets of the long DNS, with each subset having the same cost as a single TEAMS run, as measured by total duration. Each panel contains a table of corresponding parameters, including the advance split time δ, which is selected differently for the two targets based on Fig. 4.

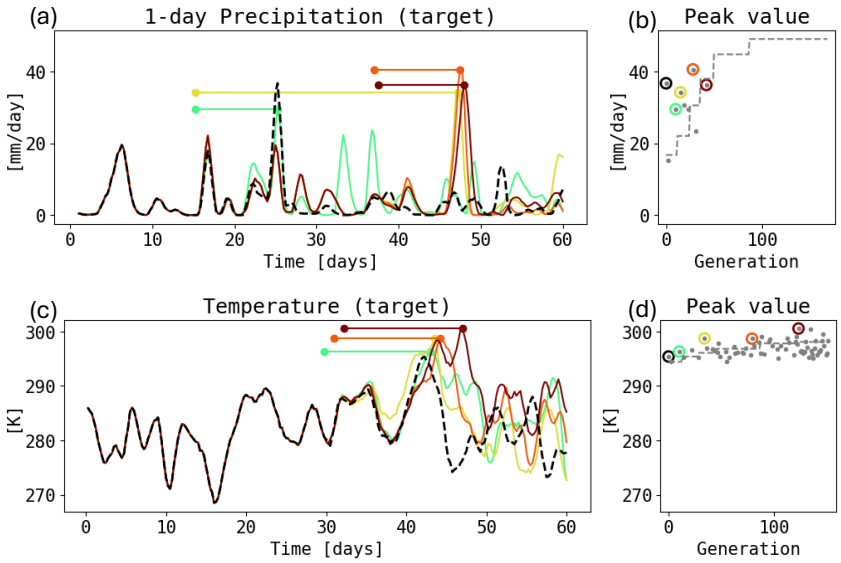

Figure 3. Examples of boosted simulations produced by TEAMS. Results are shown for (a,b) precipitation with advance split time 10 days, and (c,d) temperature with advance split times 12 days—the values found to be optimal. In panels (a,c), black dashed curves are the ancestor and colored curves are descendants (only those in the same lineage as the most-extreme descendant—the "most-extreme lineage"). Each descendant's split time and peak time are marked by circles connected by a horizontal line (note that orange and yellow lines in 3c overlap). In panels (b,d), the full sequence of descendant severities is shown as gray dots, and those in the most-extreme lineage are also circled in color. Their horizontal position indicates the generation of splitting at which they were spawned, and the dashed gray staircase indicates the algorithm's level s at that same generation. Dots falling below the staircase represent rejections, while those rising above are accepted.

4.2 Sensitivity analysis of advance split time

Fig. 4 quantifies the variation in performance with δ using two simple performance indicators. The first measures *statistical* accuracy in high return levels:

$$L^2 \text{ error} = \left(\frac{1}{\log(\tau_{\max}/\tau_{\min})} \int_{\tau_{\min}}^{\tau_{\max}} \left[\widehat{s}_{\text{DNS}}(\tau) - \widehat{s}_{\text{TEAMS}}(\tau) \right]^2 d\left[\log \tau \right] \right)^{1/2} \tag{4}$$

where τ is a return period running from $\tau_{\min} = 50$ days to $\tau_{\max} = 1.6{\times}10^4$ years, and $\widehat{s}_{(\text{DNS,TEAMS})}(\tau)$ represents the corresponding severity return level estimated by (DNS, TEAMS) by inverting the estimator $\widehat{\tau}(s)$ in Eq. (3) with linear (in $\log \tau$ space) interpolation. The integral is approximated by numerical quadrature. Because the DNS is longer than the longest return time estimable by TEAMS (and beyond the range shown in Fig. 2), we extrapolate $\widehat{s}_{\text{TEAMS}}$ to longer return periods using constant extrapolation, which penalizes runs that get stuck at small boosts and abort at shorter return periods. The second indicator measures the efficacy in boosting to larger extremes:

$$\text{Boost} = \frac{1}{M} \sum_{m=1}^{M} \max\{\max(S_\ell - S_m, 0) \,:\, X_\ell \text{ is a descendant of } X_m\} \tag{5}$$

where M is the total number of ensemble members, including all ancestors and all accepted descendants (but not rejects). Fig. 4 shows both performance indicators' δ-dependance, and confirms that *an optimal δ does exist*, in both senses of minimizing L^2 (which has a broad valley) and maximizing Boost (which has a relatively narrow peak). Happily, the same δ is approximately optimal for both, and L^2 is not very sensitive to changes in the value by $\lesssim 2$ days. However, the two targets of precipitation and temperature have slightly different optimal δs of 10 and 12 days respectively, which we will show is consistent with slower ensemble dispersion of temperature in Fig. 5. Thus it appears that the appropriate target time is not universal but rather depends, at least weakly, on the choice of target variable.

4.3 Ensemble spreading rate

Finkel and O'Gorman (2024) found that the optimal δ was well estimated as the time $t_{3/8}$ when a perturbed ensemble disperses to a fraction $3/8$ of its saturation dispersion. Having measured the optimal AST by grid search in the previous section, we now compare it with $t_{3/8}$, which is computed by the following branching procedure (same as in Finkel and O'Gorman (2024)):

1. Draw an initial condition $X(0) \sim \rho_0$, in our case a snapshot from the long DNS run plus some additional spinup of 60 days for good measure.
2. Split $X(0)$ into B branches (each with its own random seed for SPPT) and let them evolve independently for T_B days. Here we set $B = 12$ to balance cost with statistical confidence

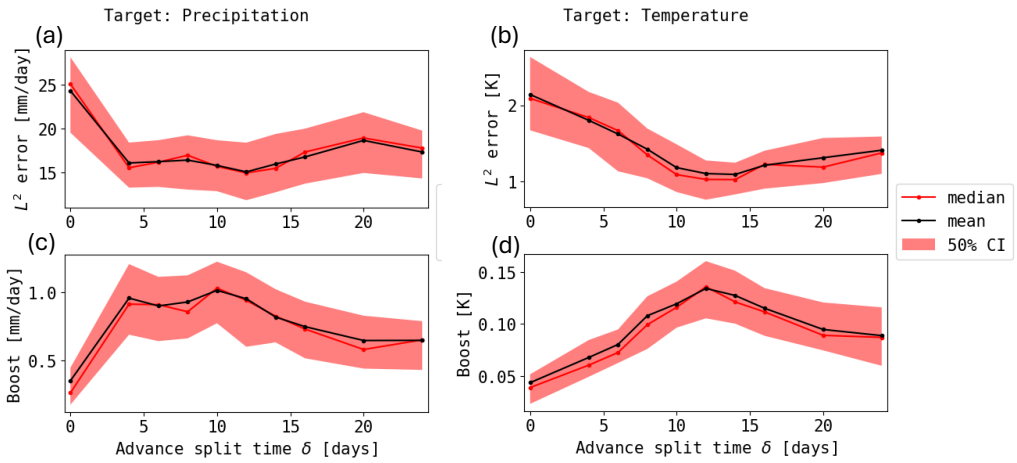

Figure 4. TEAMS performance diagnostics as functions of advance split time. We deployed TEAMS on two different target variables (left: precipitation and right: temperature) with a sequence of advance split times (ASTs) of 0, 4, 6, 8, 10, 12, 14, 16, 20, and 24. Each case was repeated 48 times with different random seeds. The finer AST spacing of 2 days between 6 and 16 was done after an initial sweep with 4-day spacing to identify a broadly optimal region. Optimality is assessed by the two diagnostics shown: (top) L^2 error between TEAMS and DNS return level curves, equivalent to the root-mean-square distance between red and black curves in Fig. 2 (smaller is better); and (bottom) the Boost, defined as the maximum increase in severity between an ensemble member and all of its descendants (or zero if all its descendants are less severe), which is averaged over all members in a TEAMS run. Both L^2 and Boost are defined for a single TEAMS run, and there are 48 runs performed at each AST, whose (mean, median, interquartile range) are plotted as (black lines, red lines, and red bands) respectively.

in estimating root-mean-squared error (RMSE) as defined below. We set $T_B = 50$ days which is long enough for the RMSE to saturate (it is similar to the TEAMS time horizon, $T = 60$ days).

3. Continue a simulation from $X(0)$ for an *equilibration interval* T_E, and split $X(T_E)$ into B more branches.

4. Repeat step 3 (but starting from the most recent split time) W times to create W ensembles, resulting in a dataset

$$\{X_{b,w}(r) : 1 \le b \le B, 1 \le w \le W, 0 \le r \le T_B\} \tag{6}$$

(W stands for "whorls", a botanical term for a point on a stem from which multiple branches emanate). We set $W = 20$. r denotes the time since the split, equivalent to $t - (w - 1)T_E$ for the wth whorl.

5. Measure the ensemble dispersion from each whorl $w = 1, \dots, W$ in terms of the RMSE as a function of the elapsed time r since the split:

$$\mathrm{RMSE}_w(r) = \sqrt{\frac{1}{B}\sum_{b=1}^{B} D\Big(X_{w,b}(r), X_{w,0}(r)\Big)^2} \tag{7}$$

Here $X_{w,b}$ refers to the bth branch from the wth whorl, while $b = 0$ denotes the "tree trunk" which spawns these branches. The distance function $D(X, Y)$ is Euclidean distance in the physical field of interest calculated over a region, chosen here to be the entire Northern Hemisphere. Other distance metrics could be used, for example by restricting to the region around the target location, which would be possibly more relevant to the event of interest but also more noisy. Fig. 5 displays the results of one whorl from this branching experiment, in the form of local precipitation and temperature timeseries (top row) and RMSEs of these two fields, respectively (bottom row). Individual branches, plotted in red, show the impact of different stochastic parameterization realizations.

6. Because different initial conditions spread at different rates, RMSE_w might have different shapes for different whorls, but each will eventually saturate to the same asymptotic value. The RMS of RMSE_w across all ws—i.e., $\sqrt{\frac{1}{W}\sum_w \mathrm{RMSE}_w^2}$, denoted $\mathrm{RMSE}(r)$—is displayed as purple lines in Fig. 5c,d, and we estimate the asymptotic RMSE by its final 15-day average. Define the fractional saturation time $t_{\epsilon,w}$ as the time s at which $\mathrm{RMSE}_w(r)$ reaches a fraction ϵ of the asymptotic value. Following the prescription from Finkel and O'Gorman (2024), δ should be approximated by $\overline{t_{3/8}} := \frac{1}{W}\sum_{w=1}^{W} t_{3/8,w}$. This is not exactly the same as the time that $\mathrm{RMSE}(r)$ crosses the threshold, but they are practically indistinguishable for the full-NH Euclidean distance (smaller-area distances lead to a dis-

crepancy). A benefit of averaging times first instead of RMSEs first is that it gives a straightforward estimate of standard deviation of $t_{3/8}$ across ws, which is denoted in the legends along with the mean [$t_{3/8} = \overline{t_{3/8}} \pm \mathrm{std}(t_{3/8})$].

For precipitation, we find $t_{3/8} = 6.3 \pm 1.0$ days, and for temperature 10.4 ± 0.7 days. A longer timescale for temperature is not surprising given that temperature is a smoother field than precipitation, and it correctly predicts that a longer δ is optimal for temperature as compared to precipitation. Comparing to the grid search over AST in Fig. 4, $\overline{t_{3/8}}$ systematically underestimates the optimal AST: for precipitation, $\overline{t_{3/8}} = 6.3$ days compared to an optimal time of 10 days, and for temperature, $\overline{t_{3/8}} = 10.4$ days compared to an optimal time of 12 days. In both cases, the optimal time roughly matches $\overline{t_{2/3}}$. We used the full Northern hemisphere to estimate ensemble dispersion which is roughly analagous to what we considered in the Lorenz 96 system and reduces noise. However, the bias reduces with smaller-area averages, with $\overline{t_{3/8}} = 8.5$ and 11.7 days respectively when restricting the averaging region for the distance metric D to $40° \times 10°$ lon×lat centered on the target. This suggests that more localized measurements of ensemble dispersion may be needed when the domain is higher-dimensional. These are all important nuances to bear in mind when expanding to other applications, especially those with different spatiotemporal scales such as mesoscale convective systems.

5 Conclusion

Extreme weather events have long been recognized as a major challenge for risk assessment, which motivates the use and development of suitable rare event algorithms: protocols to perturb simulations, over-sample the extremes, and then correct for the statistical bias introduced. The subclass of extremes which are *sudden* and *transient* resist standard rare event algorithms by simply running their course before the perturbations can take effect. We augmented a standard algorithm, adaptive multilevel splitting (AMS) with *early perturbations*, resulting in "trying early AMS" (TEAMS), and after developing the method on the benchmark Lorenz-96 system in Finkel and O'Gorman (2024), here we have successfully applied the algorithm to a three-dimensional model of the atmosphere's general circulation, extending the estimable range of return periods to $100 - 150$ years with only ~ 20 years of simulation and $300 - 500$ years with only ~ 40 years of simulation.

The key hyperparameter of this algorithm is the *advance split time*: how far ahead of time to perturb a simulated extreme event to optimally sample the range of how much more or less se-

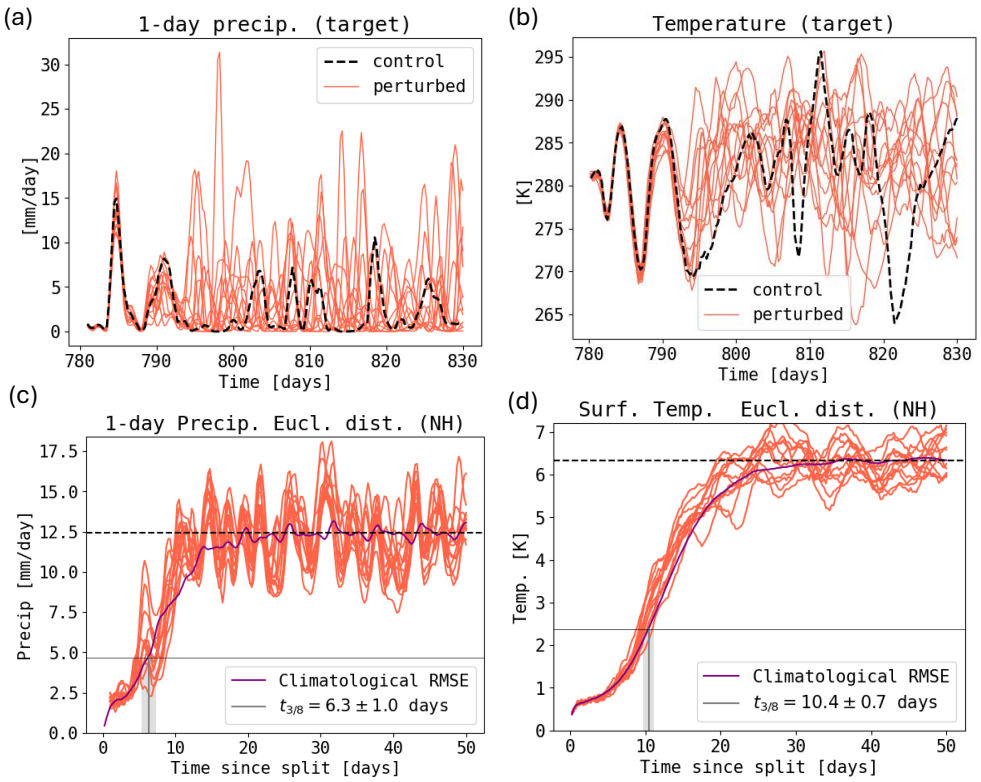

Figure 5. Ensemble dispersion. (a) Precipitation over the target region during a 50-day stretch. The control simulation (black) and 12 perturbed ensemble members (red) are all subject to different realizations of stochastic parameterization (SPPT) with the same statistics; see text for details. (c) Area-weighted Euclidean distance (RMSE) between each realization and the control (red) for the same time span, using the full northern hemisphere precipitation field. The RMSE over different initial conditions (i.e. different whorls) is shown in purple (denoted RMSE(r) in the text). The long-term average, or "saturation RMSE", is shown as a horizontal black dashed line. The horizontal gray threshold marks the fraction $3/8$ of saturation, and the vertical gray line with error bars delineates the mean and standard deviation of $t_{3/8}$, the threshold-crossing time, across whorls as $\overline{t_{3/8}} \pm \mathrm{std}(t_{3/8})$. (b,d): Same as (a,c), but using surface temperature instead of precipitation fields.

vere that event could have been. Exhaustive experiments with Lorenz-96 informed a heuristic rule to set the advance split time based on ensemble dispersion rates (Finkel & O'Gorman, 2024), and here we verified the same rule as a good approximation to the optimal choice in this much more complex, albeit idealized, atmospheric model, and for two different target variables: heavy precipitation and heat extremes. This first evidence of generalizability leads us to conjecture that a similar rule holds in more complex, realistic GCMs.

There are several wide avenues for advancing this research. An obvious next step is do testing at higher resolution and/or more realistic GCMs or regional climate models. However, algorithmic improvements are still needed for broad application. In particular, we need improved guidance in how to choose the time horizon T and the population control parameters: ancestor pool N, killing rate K, and computational budget $M_{\max}$. More interestingly, the appropriate choice of perturbation space is quite open-ended as a general question, especially when stochastic parameterization is not intrinsically a part of the model. Others have conjectured that the perturbation space is inconsequential provided the magnitude is small (Ragone et al., 2018), but this remains to be tested, as we are doing in separate ongoing work. Moreover, utilizing *deterministic optimization* to design a more structured sequence of perturbations (in a similar fashion as Farazmand and Sapsis (2017) and Sapsis (2020)) may be a route toward more efficient sampling strategies.

Another immediate goal—beyond our current scope of establishing the TEAMS algorithm, but more and more relevant with more realistic models—is to physically interpret the algorithm's output, which differs from typical datasets in that ensemble members are weighted unequally and grouped into "families". Spatial composites of relevant fields, like column water vapor, can be extracted by applying the weighted-average formula (2) pointwise to maps, which has been done for seasonal heat extremes in, e.g., Ragone et al. (2018); Ragone and Bouchet (2021); Miloshevich et al. (2024); Le Priol et al. (2024). In particular, visualizing *differences* between an ancestor and its descendants in this way will reveal mechanisms for physical drivers that strengthen or dampen extremes, and can be compared with traditional perturbations used in numerical weather prediction like Lyapunov, singular, and bred vectors (e.g., Norwood et al., 2013; Palmer & Zanna, 2013). The value added by rare event algorithms is the chance to greatly enhance statistical confidence in composite maps and other diagnostics.

Overall, we wish to convey simultaneous signals of caution and optimism. "Extreme weather events" do not comprise a monolithic category, but are tremendously diverse in spatiotemporal

scales, and one rare event algorithm off the shelf cannot be expected to successfully sample all of them. Here we have identified one particular dimension of challenge—relative timescales of ensemble dispersion and the event itself—and successfully remedied it using insight from a simpler model. The specific algorithm, and the general strategy for leveraging a model hierarchy, will help guide the community's continued exploration of extreme events, a growing frontier of climate research.

Data availability statement

The code to run the climate model and the rare event algorithm is publicly accessible in two repositories:

1. "`jf_conv_gray_smooth`" (justinfocus12, 2025, available at `https://doi.org/10.5281/zenodo.16878347`) contains the core Fortran model code
2. "`TEAMS`" (justinfocus12, n.d., available at `https://doi.org/10.5281/zenodo.16878339`) contains Python code for the rare event algorithm that wraps the Fortran code as well as some other example systems (including Lorenz-96).

Interested readers should contact J. F. (`ju26596@mit.edu`) for guidance on using and extending the code.

Conflict of interest

The authors declare no conflicts of interest relevant to this study.

Acknowledgments

We thank Judith Berner for assistance with implementing the stochastic parameterization. Computations for this project were performed on the MIT Engaging cluster. This research is part of the MIT Climate Grand Challenge on Weather and Climate Extremes. Support was provided by Schmidt Sciences, LLC.

References

Anderson, J. L., et al. (2004). The new gfdl global atmosphere and land model am2–lm2: Evaluation with prescribed sst simulations. *Journal of Climate*, *17*(24), 4641 - 4673. Retrieved from `https://journals.ametsoc.org/view/journals/clim/17/24/`

jcli-3223.1.xml doi: 10.1175/JCLI-3223.1

Au, S.-K., & Beck, J. L. (2001). Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, *16*(4), 263-277. Retrieved from https://www.sciencedirect.com/science/article/pii/S0266892001000194 doi: https://doi.org/10.1016/S0266-8920(01)00019-4

Berner, J., Fossell, K. R., Ha, S.-Y., Hacker, J. P., & Snyder, C. (2015). Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Monthly Weather Review*, *143*(4), 1295 - 1320. Retrieved from https://journals.ametsoc.org/view/journals/mwre/143/4/mwr-d-14-00091.1.xml doi: 10.1175/MWR-D-14-00091.1

Berner, J., Shutts, G. J., Leutbecher, M., & Palmer, T. N. (2009). A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ecmwf ensemble prediction system. *Journal of the Atmospheric Sciences*, *66*(3), 603 - 626. Retrieved from https://journals.ametsoc.org/view/journals/atsc/66/3/2008jas2677.1.xml doi: 10.1175/2008JAS2677.1

Bloin-Wibe, L., Noyelle, R., Humphrey, V., Beyerle, U., Knutti, R., & Fischer, E. (2025). Estimating return periods for extreme events in climate models through ensemble boosting. *EGUsphere*, *2025*, 1–40. Retrieved from https://egusphere.copernicus.org/preprints/2025/egusphere-2025-525/ doi: 10.5194/egusphere-2025-525

Farazmand, M., & Sapsis, T. P. (2017). A variational approach to probing extreme events in turbulent dynamical systems. *Science Advances*, *3*(9), e1701533. Retrieved from https://www.science.org/doi/abs/10.1126/sciadv.1701533 doi: 10.1126/sciadv.1701533

Finkel, J., & O'Gorman, P. A. (2025). Boosting ensembles for statistics of tails at conditionally optimal advance split times. Retrieved from https://arxiv.org/abs/2507.22310

Finkel, J., & O'Gorman, P. A. (2024). Bringing statistics to storylines: Rare event sampling for sudden, transient extreme events. *Journal of Advances in Modeling Earth Systems*, *16*(6), e2024MS004264. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024MS004264 (e2024MS004264 2024MS004264) doi: https://doi.org/10.1029/2024MS004264

Frierson, D. M. W., Held, I. M., & Zurita-Gotor, P. (2006). A gray-radiation aquaplanet

moist gcm. part i: Static stability and eddy scale. *Journal of the Atmospheric Sciences*, *63*(10), 2548 - 2566. Retrieved from https://journals.ametsoc.org/view/journals/atsc/63/10/jas3753.1.xml doi: 10.1175/JAS3753.1

Gessner, C. (2022). *Physical storylines for very rare climate extremes* (Unpublished doctoral dissertation). ETH Zurich.

Gessner, C., Fischer, E. M., Beyerle, U., & Knutti, R. (2021). Very rare heat extremes: Quantifying and understanding using ensemble reinitialization. *Journal of Climate*, *34*(16), 6619 - 6634. Retrieved from https://journals.ametsoc.org/view/journals/clim/34/16/JCLI-D-20-0916.1.xml doi: 10.1175/JCLI-D-20-0916.1

Held, I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, *86*(11), 1609 - 1614. Retrieved from https://journals.ametsoc.org/view/journals/bams/86/11/bams-86-11-1609.xml doi: 10.1175/BAMS-86-11-1609

Huang, X., Chen, J., & Zhu, H. (2016). Assessing small failure probabilities by ak–ss: An active learning method combining kriging and subset simulation. *Structural Safety*, *59*, 86-95. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167473016000035 doi: https://doi.org/10.1016/j.strusafe.2015.12.003

justinfocus12. (n.d.). *justinfocus12/teams: Initial release 2320 for submission 2321.*

justinfocus12. (2025, August). *justinfocus12/jf_conv_gray_smooth: Initial release for submission.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.16878347 doi: 10.5281/zenodo.16878347

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

Krakauer, N. Y. (2024). It is normal: The probability distribution of temperature extremes. *Climate*, *12*(12). Retrieved from https://www.mdpi.com/2225-1154/12/12/204 doi: 10.3390/cli12120204

Le Priol, C., Monteiro, J. M., & Bouchet, F. (2024, oct). Using rare event algorithms to understand the statistics and dynamics of extreme heatwave seasons in south asia. *Environmental Research: Climate*, *3*(4), 045016. Retrieved from https://dx.doi.org/10.1088/2752-5295/ad8027 doi: 10.1088/2752-5295/ad8027

Lestang, T., Bouchet, F., & Lévêque, E. (2020). Numerical study of extreme mechanical force exerted by a turbulent flow on a bluff body by direct and rare-event sampling

techniques. *Journal of Fluid Mechanics*, *895*, A19. doi: 10.1017/jfm.2020.293

Lestang, T., Ragone, F., Bréhier, C.-E., Herbert, C., & Bouchet, F. (2018, Apr). Computing return times or return periods with rare event algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, *2018*(4), 043213. Retrieved from https://doi.org/10.1088/1742-5468/aab856 doi: 10.1088/1742-5468/aab856

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Elms, J., … Willard, J. (2024). Huge ensembles part i: Design of ensemble weather forecasts using spherical fourier neural operators. Retrieved from https://arxiv.org/abs/2408.03100

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Harrington, P., … Willard, J. (2024). Huge ensembles part ii: Properties of a huge ensemble of hindcasts generated with spherical fourier neural operators. Retrieved from https://arxiv.org/abs/2408.01581

Miloshevich, G., Lucente, D., Yiou, P., & Bouchet, F. (2024). Extreme heat wave sampling and prediction with analog markov chain and comparisons with deep learning. *Environmental Data Science*, *3*, e9.

Norwood, A., Kalnay, E., Ide, K., Yang, S.-C., & Wolfe, C. (2013, Jun). Lyapunov, singular and bred vectors in a multi-scale system: an empirical exploration of vectors related to instabilities. *Journal of Physics A: Mathematical and Theoretical*, *46*(25), 254021. Retrieved from https://dx.doi.org/10.1088/1751-8113/46/25/254021 doi: 10.1088/1751-8113/46/25/254021

O'Gorman, P. A., & Schneider, T. (2008). The hydrological cycle over a wide range of climates simulated with an idealized gcm. *Journal of Climate*, *21*(15), 3815 - 3832. Retrieved from https://journals.ametsoc.org/view/journals/clim/21/15/2007jcli2065.1.xml doi: 10.1175/2007JCLI2065.1

O'Gorman, P. A., & Schneider, T. (2009). Scaling of precipitation extremes over a wide range of climates simulated with an idealized gcm. *Journal of Climate*, *22*(21), 5676 - 5685. Retrieved from https://journals.ametsoc.org/view/journals/clim/22/21/2009jcli2701.1.xml doi: 10.1175/2009JCLI2701.1

Palmer, T. N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G. J., … Weisheimer, A. (2009). Stochastic parametrization and model uncertainty. *ECMWF Technical Memoranda*.

Palmer, T. N., & Zanna, L. (2013, Jun). Singular vectors, predictability and ensemble forecasting for weather and climate. *Journal of Physics A: Mathematical and Theoretical*,

46(25), 254018. Retrieved from https://dx.doi.org/10.1088/1751-8113/46/25/254018 doi: 10.1088/1751-8113/46/25/254018

Ragone, F., & Bouchet, F. (2021). Rare event algorithm study of extreme warm summers and heatwaves over europe. *Geophysical Research Letters*, *48*(12), e2020GL091197. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020GL091197 (e2020GL091197 2020GL091197) doi: https://doi.org/10.1029/2020GL091197

Ragone, F., Wouters, J., & Bouchet, F. (2018). Computation of extreme heat waves in climate models using a large deviation algorithm. *Proceedings of the National Academy of Sciences*, *115*(1), 24–29. Retrieved from https://www.pnas.org/content/115/1/24 doi: 10.1073/pnas.1712645115

Ragulina, G., & Reitan, T. (2017). Generalized extreme value shape parameter and its nature for extreme precipitation using long time series and the bayesian approach. *Hydrological Sciences Journal*, *62*(6), 863–879. Retrieved from https://doi.org/10.1080/02626667.2016.1260134 doi: 10.1080/02626667.2016.1260134

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

Sapsis, T. P. (2020). 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*(2234), 20190834. Retrieved from https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2019.0834 doi: 10.1098/rspa.2019.0834

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L. V., Hegerl, G., ... Zwiers, F. W. (2017). Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and Climate Extremes*, *18*, 65-74. Retrieved from https://www.sciencedirect.com/science/article/pii/S2212094717300440 doi: https://doi.org/10.1016/j.wace.2017.10.003

Uribe, F., Papaioannou, I., Marzouk, Y. M., & Straub, D. (2021). Cross-entropy-based importance sampling with failure-informed dimension reduction for rare event simulation. *SIAM/ASA Journal on Uncertainty Quantification*, *9*(2), 818-847. Retrieved from https://doi.org/10.1137/20M1344585 doi: 10.1137/20M1344585

Webber, R. J., Plotkin, D. A., O'Neill, M. E., Abbot, D. S., & Weare, J. (2019). Practical rare event sampling for extreme mesoscale weather. *Chaos: An Interdisciplinary Journal*

of Nonlinear Science, *29*(5), 053109. Retrieved from https://doi.org/10.1063/1.5081461 doi: 10.1063/1.5081461

Winget, M., & Persky, A. M. (2022). A practical review of mastery learning. *American Journal of Pharmaceutical Education*, *86*(10), ajpe8906. Retrieved from https://www.sciencedirect.com/science/article/pii/S0002945923007386 doi: https://doi.org/10.5688/ajpe8906

Wouters, J., & Bouchet, F. (2016, Aug). Rare event computation in deterministic chaotic systems using genealogical particle analysis. *Journal of Physics A: Mathematical and Theoretical*, *49*(37), 374002. Retrieved from https://dx.doi.org/10.1088/1751-8113/49/37/374002 doi: 10.1088/1751-8113/49/37/374002

Zhang, B. J., Sahai, T., & Marzouk, Y. M. (2022). A koopman framework for rare event simulation in stochastic differential equations. *Journal of Computational Physics*, *456*, 111025. Retrieved from https://www.sciencedirect.com/science/article/pii/S0021999122000870 doi: https://doi.org/10.1016/j.jcp.2022.111025

Zuckerman, D. M., & Chong, L. T. (2017). Weighted ensemble simulation: Review of methodology, applications, and software. *Annual Review of Biophysics*, *46*(1), 43-57. Retrieved from https://doi.org/10.1146/annurev-biophys-070816-033834 (PMID: 28301772) doi: 10.1146/annurev-biophys-070816-033834