JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2025MS005456

Rare Event Sampling for Moving Targets: Extremes of Temperature and Daily Precipitation in a General Circulation Model

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

[1]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA, [2]Department of Geophysical Sciences and the Data Science Institute, University of Chicago, Chicago, IL, USA

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

Correspondence to:
J. Finkel,
jfinkel@uchicago.edu

Citation:
Finkel, J., & O'Gorman, P. A. (2026). Rare event sampling for moving targets: Extremes of temperature and daily precipitation in a general circulation model. *Journal of Advances in Modeling Earth Systems*, *18*, e2025MS005456. https://doi.org/10.1029/2025MS005456

Received 25 AUG 2025
Accepted 12 FEB 2026

Author Contributions:
Conceptualization: Justin Finkel, Paul A. O'Gorman
Data curation: Justin Finkel, Paul A. O'Gorman
Funding acquisition: Paul A. O'Gorman
Investigation: Justin Finkel, Paul A. O'Gorman
Methodology: Justin Finkel, Paul A. O'Gorman
Project administration: Paul A. O'Gorman
Resources: Paul A. O'Gorman
Software: Justin Finkel, Paul A. O'Gorman
Supervision: Paul A. O'Gorman
Validation: Justin Finkel, Paul A. O'Gorman
Visualization: Justin Finkel, Paul A. O'Gorman

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, for example, 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., a return period of 150 years from 20 years of simulation). The algorithm, called TEAMS ("trying-early adaptive multilevel splitting"), was developed previously with 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 toward 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 that used 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 toward the extremes by repeated perturbation, pruning, and cloning steps, while adjusting probability weights on ensemble members to compensate for preferential sampling and thus enable unbiased 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; Webber et al., 2019; Wouters & Bouchet, 2016). Rare event algorithms are attractive for being agnostic to the model: importantly, they can operate on models grounded in physics and potentially

Writing – original draft: Justin Finkel, Paul A. O'Gorman
Writing – review & editing: Justin Finkel, Paul A. O'Gorman

could also be applied to faster, data driven models with the alluring possibility of generating abundant extreme events at will (Mahesh et al., 2024a, 2024b).

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: an overlap of timescales between ensemble dispersion and event duration. If the event duration is too short, the ensemble has too little time to spread out and sample the tails before the event is over. This is not a problem 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, with enough time for dispersion between each round, so that extreme anomalies can be achieved 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 (Finkel & O'Gorman, 2024; Lestang et al., 2020; Rolland, 2022). This is a major limitation given that transient cyclones and anticyclones 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 by adapting the classical Adaptive Multilevel Splitting (AMS) algorithm (Cérou & Guyader, 2007; Lestang et al., 2018) in which a level of extremity is progressively raised and only ensemble members reaching that level are retained and split at a level crossing. We modified AMS to perturb simulations in advance of the level crossing, thus giving more time for ensemble members to separate before reaching an extreme event. Splitting in advance requires an additional acceptance/rejection step that we include through the formalism of subset simulation (Au & Beck, 2001). The resulting algorithm, TEAMS ("trying-early adaptive multilevel splitting"), introduces a key hyperparameter, the *advance split time* (AST), 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). TEAMS draws inspiration from the related approach of *ensemble boosting* (Gessner, 2022; Gessner et al., 2021) which has recently been extended to include probability estimates (Bloin-Wibe et al., 2025; Finkel & O'Gorman, 2025). Ensemble boosting differs from TEAMS in that it perturbs before the extreme events in an existing long simulation, and perturbs them all the same number of times without sub-selection based on a level-raising protocol. In Finkel and O'Gorman (2024) we demonstrated TEAMS on the Lorenz-96 system, a relatively simple model of spatiotemporal chaos that nevertheless captures the essence of baroclinic waves. 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 aquaplanet 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 the 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 horizontal spectral resolution to T21, meaning a triangular truncation of spherical harmonics with maximum wavenumbers 21 in both zonal and meridional directions (Krishnamurti et al., 2006).

We also reduce the temporal and vertical resolution, using a 40-min timestep and six σ-levels in the vertical (half levels at $\sigma = 0.0, 0.0343, 0.15, 0.4, 0.7, 0.966, 1.0$), where σ is pressure normalized by surface pressure. We also present some limited results at a higher horizontal resolution of T42, with 30 vertical levels and a 10-min timestep. The simulations were performed with MPI on four Intel Xeon cores, completing 60 days of simulation in roughly 20 s at T21 resolution, or 15 min at T42 resolution. All results are at the default resolution of T21 unless otherwise noted.

The second modification is to introduce the randomness needed to induce variability between ensemble members. Other rare event sampling methods (e.g., Abbot et al., 2021; Bloin-Wibe et al., 2025; Ragone et al., 2018) used single-time perturbations. Here, in line with the continuous-time forcing used in Finkel and O'Gorman (2024), we implemented a stochastic parameterization scheme known as stochastically perturbed parameterized tendencies (SPPT). SPPT was developed in numerical weather prediction to enhance ensemble spread to more likely capture the observed evolution (Berner et al., 2009, 2015; Palmer et al., 2009), and here we can use it to discover unlikely paths toward extremes. Our implementation of SPPT closely follows the specification in Palmer et al. (2009), which contains further details and background. In brief, SPPT randomly perturbs the total parameterized tendencies (i.e., contributions from large-scale condensation, vertical turbulent diffusion, and radiation) of horizontal winds, humidity, and temperature. The perturbation acts every timestep through a multiplicative factor of $1 + r_{\mathrm{SPPT}}(x, y, z, t)$, 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. There are three tunable parameters: characteristic autocorrelation timescale τ_{SPPT}, length scale L_{SPPT} which specifies how quickly amplitude drops off with wavenumber, and an overall multiplier σ_{SPPT}. To prevent unrealistically large fluctuations, r_{SPPT} is clipped to the range of ± 2 standard deviations. Sensitivity analysis led us to select $L_{\mathrm{SPPT}} = 500$ km, $\tau_{\mathrm{SPPT}} = 6$ hours, and $\sigma_{\mathrm{SPPT}} = 0.3$. We gave especially careful consideration to σ_{SPPT}, the overall noise amplitude, because the analogous stochastic forcing strength that Finkel and O'Gorman (2024) used on Lorenz-96 was found to strongly affect optimal advance split time and the extent to which TEAMS improved on AMS. The choice of $\sigma_{\mathrm{SPPT}} = 0.3$ will be justified in Figure 1, and is 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, 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.

Figure 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 (Figures 1a and 1b), which grow and decay over synoptic $\sim$5-day timescales (indicated by the Hovmöller diagrams in Figures 1c and 1d). Our aim is to characterize—using rare event sampling—the extreme, local fluctuations in these fields at the storm track's center. We therefore fix a target latitude of 45°N and a target longitude of 180°E, taking the field value in a single grid cell ($\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 11 longitudinal rotations in 30° increments for more stable estimation with 12 times the data. Figures 1e and 1f displays the long-term climate statistics of precipitation and temperature at the target location, revealing $\sigma_{\mathrm{SPPT}} \approx 0.3$ to be near the upper limit of noise level that still avoids disrupting the deterministic model's statistics too severely. The effect of noise is to increase precipitation extremes but decrease mean temperatures. Tagle et al. (2016) found stochastic parameterization increased mean temperatures in the Community Atmosphere Model, and the different result for temperature found here may relate to the idealized GCM we use which does not include land or cloud radiative effects.

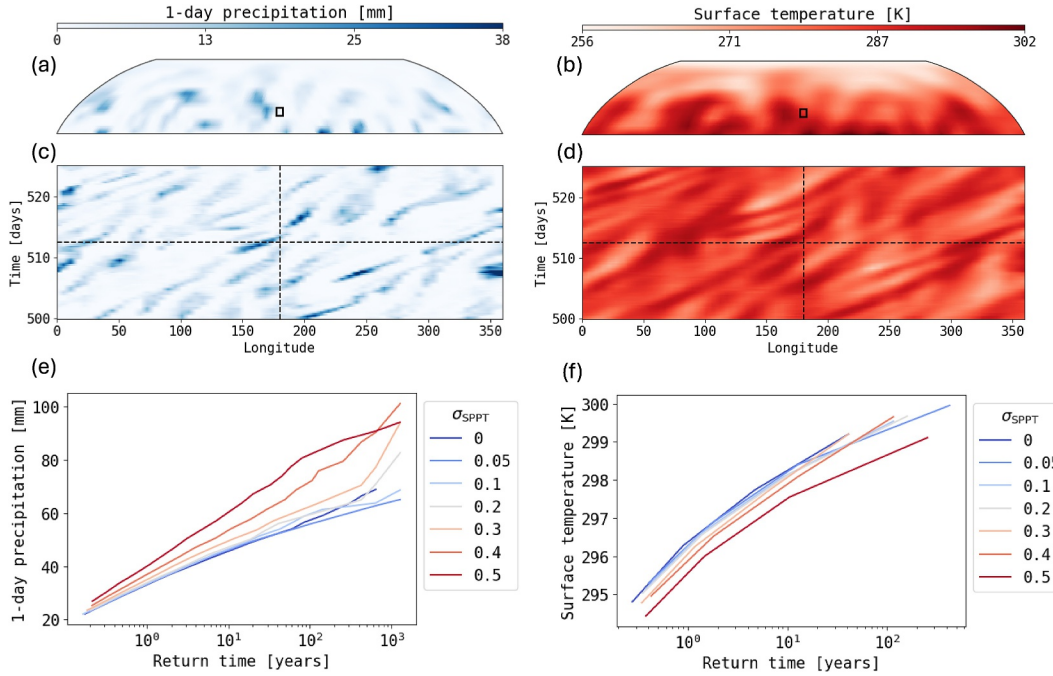

Figure 1. Simulated precipitation and surface temperature fields and their return levels at T21 resolution. After a spin-up of 500 days, the aquaplanet GCM produces physically plausible large-scale storm track dynamics: a sequence of extratropical cyclones and anticyclones bringing packets of precipitation (a) and temperature fluctuations (b), propagating eastward with lifetimes of ∼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 versus 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 $\sigma_{SPPT} \lesssim 0.3$ and start deviating substantially for larger values, which is why we adhere to $\sigma_{SPPT} = 0.3$ in panels (a–d) and hereafter.

The results in Figure 1 at T21 resolution are based on a long run of 36,500 days (100 years, or 1,200 years including longitudinal rotation) after spinup, which we refer to as a direct numerical simulation (DNS). For validation of TEAMS results shown later, we extended the data set at $\sigma_{SPPT} = 0.3$ even further: at T21 resolution we generated 272.8 years of DNS with 30 longitudinal rotations for 8,184 years total; and at T42 we generated 8.4 years of DNS with 30 longitudinal rotations for 253 years total. 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. Figure 2 serves as a visual reference for the key elements of the procedure.

1. Ancestor initialization: Sample N initial conditions $\{X_1(0), X_2(0), \ldots, X_N(0)\}$ from the distribution of interest, denoted ρ_0. For us, ρ_0 is the distribution at statistical steady state, that is, the limiting distribution of a very long GCM simulation. Other applications might restrict the initial conditions to specific phases of an 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 $\Delta T_{init} = 60$ days, chosen as roughly twice the *error saturation* timescale (over which two branched simulations decorrelate).

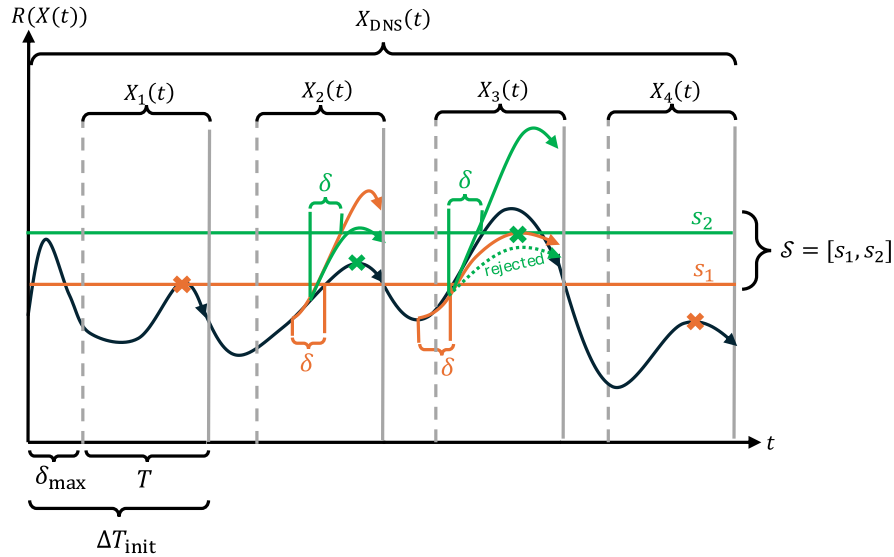

Figure 2. Schematic of TEAMS in a simplified setup with $N = 4$ ancestors and only two level-raisings. Orange elements pertain to the first culling and cloning step; green elements pertain to the second. Horizontal lines are levels s set by the algorithm, and crosses denote maxima that are eliminated as a result of falling below the level. The black line is the DNS used to generate the ancestors. Splits occur a time δ before a level crossing. Severities are measured over a time horizon T that follows a runway $\delta_{\max}$. See text for further details.

2. Ancestor simulation: Run the dynamics forward for a *time horizon* T from each ancestral initial condition. We set $T = 35$ days, long enough to contain 1-2 events and ample time for a branched ensemble to decorrelate. The resulting data set is $\{X_n(t) : 1 \le n \le N, 0 \le t \le T\}$, where we have re-indexed time for convenience. In our setup, no new simulation is necessary to create the ancestor simulations: we just use the DNS segments directly. We reserve $\delta_{\max}$ days before each initial condition and after the previous terminal condition as a "runway," where $\delta_{\max} = 25$ was chosen as an upper bound on the advance split times considered. The runway is needed for cases in which the split time occurs before the ancestral initial condition. Both the runway and the remaining time horizon of $T = \Delta T_{\mathrm{init}} - \delta_{\max} = 35$ days helps ancestors be more independent, which is not a strict requirement (they need only be identically distributed), but improves statistical robustness. Thus, each ancestor simulation takes the form

$$X_n(t) = X_{\mathrm{DNS}}(\delta_{\max} + (n-1)\Delta T_{\mathrm{init}} + t). \tag{1}$$

Figure 2 depicts how the ancestor simulations are laid out relative to the DNS simulation.
Assign each ancestor a probability weight $W_n = 1$. Furthermore, initialize a set of *active members*

$$\mathcal{A} = \{1, \ldots, N\} =: \{a_1, \ldots, a_A\} \tag{2}$$

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 four snapshots = one day) or surface temperature (measured at a single six-hourly snapshot) at the target grid box indicated in Figure 1. Choose a number $K < A$ and cull 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)). We denote s as the level that will be progressively raised as the TEAMS algorithm proceeds. We set s equal to he K-th smallest severity such that it 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 $\mathcal{A} = \{a_1, \ldots, a_{A-K}\}$, 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 after $\delta_{\max}$ (in six-hourly outputs) that $R_a(t) > s$ and call this time t_a^s. At an *earlier* time $t_a^s - \delta$ (which can be in the initial length-$\delta_{\max}$ runway), spawn a new "child" $\tilde{X}$ which shares its parent's history up until $t_a^s - \delta$, but then gets perturbed by use of a new seed for random number generation in the stochastic parameterization scheme (or a small random kick if the model is deterministic). Thereafter, the child diverges from its parent for the remainder of the simulation until the time horizon ends at T. δ is the key *advance split time* (AST) parameter, which we vary systematically in this study from 0 to 20 days. Calculate the child's severity $\tilde{S}$ as the maximum of its intensity $\tilde{R}(t)$ over $0 \le t \le T$, the same time horizon as used for the parent, which excludes the initial $\delta_{\max}$ runway. The next step depends on whether the child's severity exceeds s:
   a. If the child's severity $\tilde{S}$ exceeds s, we call this "success" and officially admit the child into the active population: $X_{a_{A+1}} = \tilde{X}$, with the same probability weight as its parent, and $S_{a_{A+1}} = \tilde{S}$. 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 Figure 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 with $N = 16$, we set $M_{\max} = 150$. For $N = 32$, we set $M_{\max} = 300$.
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 $\{X(t) : 0 < t \le T\}$ with $X(0)$ drawn from ρ_0, and F is a generic functional, estimate its expectation as

$$\hat{F} = \frac{\sum_{m=1}^{M} W_m F(X_m)}{\sum_{m=1}^{M} W_m}. \tag{3}$$

The denominator is always equal to N. In particular, for any given severity s, an estimate $\hat{\mathbb{P}}\{S > s\}$ 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

$$\hat{\tau}(s) = -\frac{T}{\log\left[1 - \hat{\mathbb{P}}\{S > s\}\right]}, \tag{4}$$

where T is the time horizon.

The estimator (3) is unbiased, meaning correct *in expectation* over different independent repetitions ("runs") of the entire procedure, each of which is itself random. In the results to follow, we have generated 48 independent runs with different ancestors and random seeds to get an accurate sense of variability across runs. The total cost of a single TEAMS run, as reported below in Figure 3, is taken to be $M(\delta + T)$, where M is the total number of members generated in the run ($\le M_{\max}$; 150 for $N = 16$ and 300 for $N = 32$), δ is the advance split time ($\le \delta_{\max} = 25$ days) and T is the time horizon ($= 35$ days). These costs range from 20 to 40 years per run in the experiments shown.

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

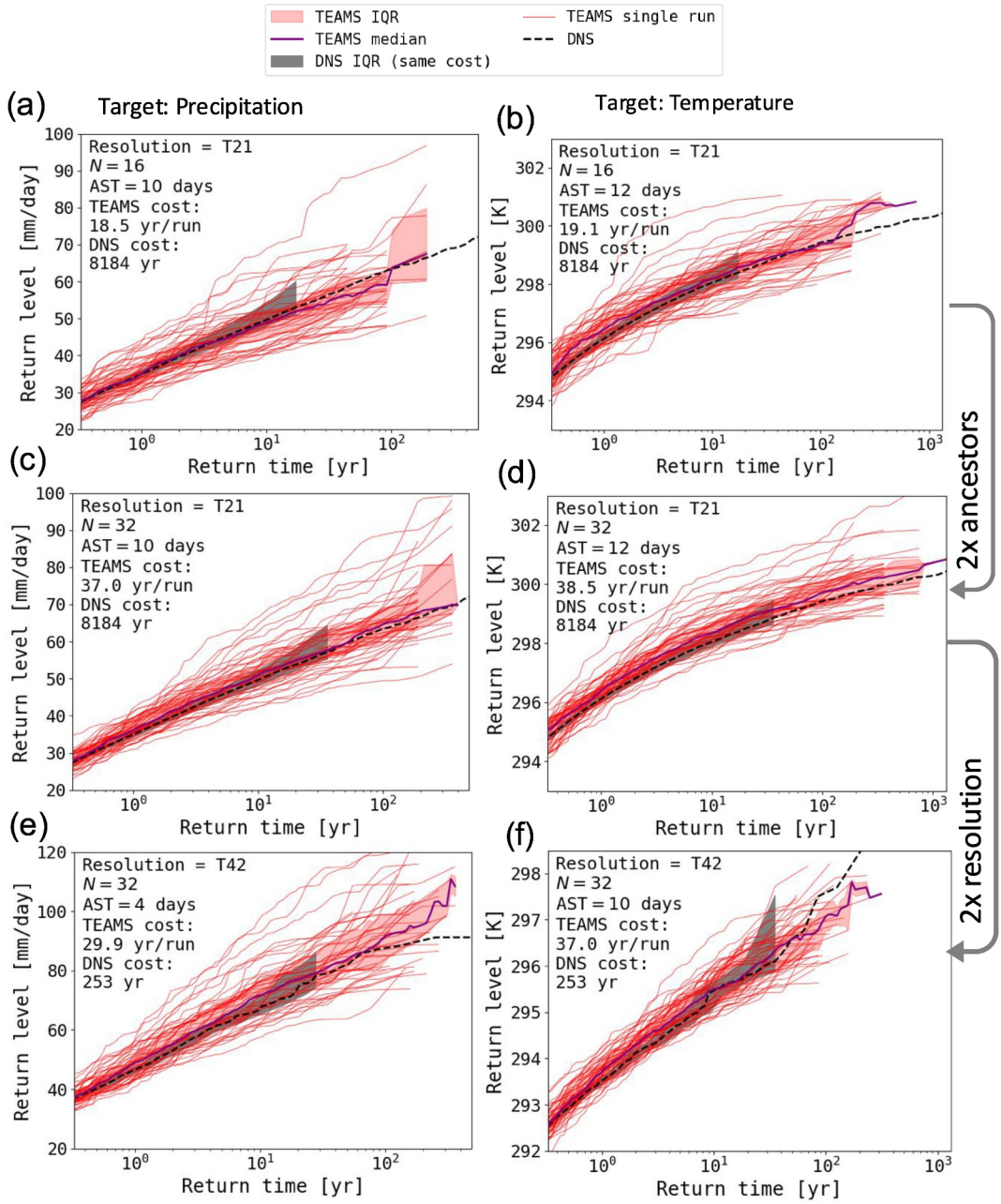

Figure 3. Performance of the rare event algorithm (TEAMS) against the benchmark direct numerical simulation (DNS) in terms of accuracy, uncertainty, and computational cost. Target variables are precipitation (left) and surface temperature (right). Panels (a, b) show the baseline setting: T21 resolution with $N = 16$ ancestors. We further perform two "pivot" experiments: doubling the ancestors to 32 (c, d) and doubling the resolution to T42 (e, f). Note the difference in vertical axis scale for T42, as resolution can strongly influence the possible ranges. All curves are estimates of return level (event severity) as a function of return period (average inter-event time). Black dashed lines come from a long DNS ("ground truth"). Each thin red line comes from a run of TEAMS with a different random seed (48 in total). Purple lines and red bands indicate medians and inter-quartile ranges (25th–75th percentiles) across the 48 runs, or somewhat fewer in the far tail, restricting to runs with enough splits 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, equal in cost to a single TEAMS run. Each panel contains a table of corresponding parameters and costs.

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 $M_{\max} = 150$ (when $N = 16$) or $M_{\max} = 300$ (when $N = 32$) as the second major difference from Finkel and O'Gorman (2024), where the budget was 1,024 (with $N = 128$) 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 estimator's variance as well as its "apparent bias": the same phenomenon that causes a coin with true "tails" probability of 1/100 to underestimate it at zero on 99% of flips, even though each flip is unbiased. The previous version of TEAMS had a similar high probability of underestimating return levels in any given run, despite being unbiased, and the algorithmic modification was critical for extending this algorithm from a toy model (Lorenz-96) to a GCM.

We highlight two connections between this new version of TEAMS and other existing algorithms. First, 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. Second, with particular choices of culling schedule, TEAMS could resemble ensemble boosting with probability estimates as laid out in Bloin-Wibe et al. (2025) and Finkel and O'Gorman (2025). Ensemble boosting starts by immediately selecting ancestors exceeding an already-extreme threshold, such as the 0.9 quantile. In TEAMS, one could customize the culling schedule by changing K as the level is raised, for example, $K = 0.95N$ for the first level and $K = 1$ subsequently. One could furthermore halt the level-raising at a single level, and increase the population size to several times the initial N (with proper re-weighting as in step 4a with each new descendant), and this would make TEAMS very much resemble ensemble boosting. We don't make such drastic modifications here, as there is some benefit to raising the level more modestly and giving the moderately extreme ancestors more chances to boost, but it is helpful to view these algorithms as existing on a continuum, related to the tradeoff between exploration and exploitation in optimization methods (Rose et al., 2021).

The 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 4.4, 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, and at two resolutions, to demonstrate its ability to sample extreme events in the GCM.

4. Results

4.1. TEAMS Performance

In our default configuration of T21 resolution and $N = 16$ ancestors, we ran TEAMS for a range of advance split times $\delta \in \{0, 4, 6, 8, 10, 12, 14, 16, 20, 24\}$ days. Figures 3a and 3b displays the resulting estimates of return level versus 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. 4.3. In addition, as a test of the algorithm's scaling behavior, we performed two "pivot" experiments: doubling the ancestor pool to $N = 32$ in Figures 3c and 3d, and doubling horizontal resolution to T42 in Figures 3e and 3f. Resolution-doubling tests the algorithm's robustness with more expensive, realistic models, and to what extent lower-resolution versions can inform best practices. To maximize the chances of success at T42 we retained $N = 32$ (which proved well worth the extra cost at T21), and re-calibrated the advance split time in light of differing dispersion timescales at higher resolution (see Section 4.4).

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 Figure 3, 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 Figure 3 assess reliability by how close to the ground truth one can expect a single TEAMS run to land with 50% probability. Clearly, individual runs of TEAMS can deviate substantially from the ground truth—a problem that

should abate with larger N, but perhaps slowly—making it advisable for practitioners to perform several independent runs to assess uncertainty. However, most runs do cluster near the ground truth, as conveyed by the red error bar.

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 accuracy in the bulk and accuracy in the tail of the distribution. For the default case of $N = 16$ (Figures 3a and 3b), one run of TEAMS is equivalent to $\sim$19 years of DNS in computational cost. TEAMS is less certain than DNS in return periods $\lesssim$19 years (the TEAMS computational budget), as indicated by its wider error bars. But TEAMS 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). 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. Furthermore, extreme value theory can struggle to capture the shape parameter and hence the far tails correctly, sometimes assigning zero probability to events that are dynamically possible (e.g., Zeder et al., 2023), which TEAMS can overcome by using the dynamics itself.

Doubling the ancestor pool from $N = 16$ to 32 (Figures 3c and 3d) 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 estimate for return periods from 40 to 300 years for precipitation and 40–500 years for temperature, which a 40 years DNS could not estimate.

Doubling the resolution to T42 also leads to good performance and speedup (Figures 3e and 3f) albeit not quite as good as for T21. We chose shorter advance split times for T42 based on ensemble dispersion experiments as discussed in detail in Section 4.4. The T42 runs are significantly more expensive: besides doubling horizontal resolution, we also increased vertical levels from 6 to 30 and reduced the timestep from 2,400 to 600 s. We expect that further experimentation with advance split times and population control parameters (such as N, K) should make improvements possible at this and much higher resolutions. Although the performance is contingent on hyperparameters, we have demonstrated generalization to higher resolution, which is enough to draw cautious optimism for the algorithm's scalability.

4.2. Case Studies and Population Dynamics

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. Figure 4 displays one case study for each target variable (precipitation and temperature), with the same advance split times as used at T21 in Figures 3a–3d (10 and 12 days, respectively). Boosting happens either by amplifying an existing spike, or by materializing a new spike where none existed before. In Figure 4a, 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 Figure 4c, descendants build on the original spike leading to higher and higher severities. Intuitively, this is the more desirable behavior for TEAMS, going by the heuristic guidance that "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 it would limit the discovery of new events that do, in practice, appear to contribute to the best successes of TEAMS so far by allowing later generations to distinguish themselves. We surmise that employing more deterministic optimization strategies, such as Newton's method in the space of perturbations, might help to get the most possible improvement out of existing peaks and obviating the need to get lucky by discovering completely new events. This is a primary direction of our ongoing research.

The "population dynamics" of TEAMS offers another window into its behavior and a diagnostic for possible improvements. Figures 4b and 4d shows aspects of the ensemble members' progress through generations for the

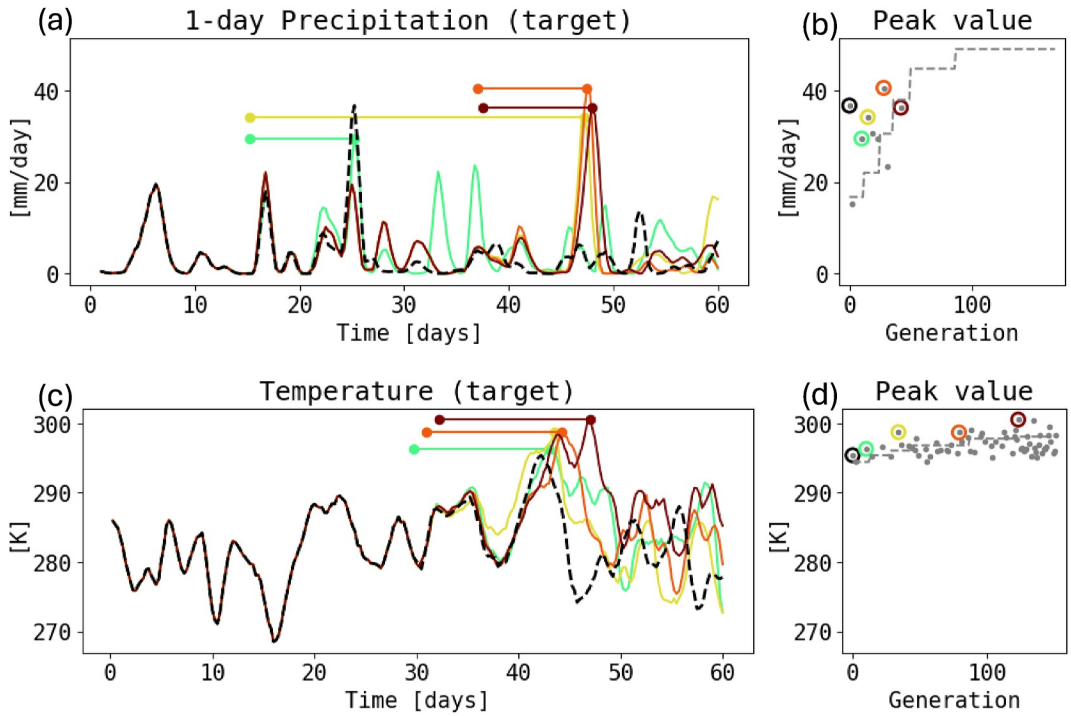

Figure 4. Examples of boosted simulations produced by TEAMS at T21 resolution. Results are shown for (a, b) precipitation with advance split time 10 days, and (c, d) temperature with advance split times 12 days—the values found to be optimal at this resolution. In panels (a, c), black dashed curves are 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. There are more gray dots in (d) because the family in (c, d) happened to survive for more rounds of level-raising than the family in (a, b).

same case study. The level s rises in a stepwise manner while descendant scores rise, on average, only gradually over successive generations, eventually falling systematically below the levels and increasing the rejection rate in later stages of the algorithm. The same story plays out when averaging over runs in Figures 5a and 5b: severities reached by new ensemble members rise at a slower rate than the levels. The curves cross at generation 3, both for precipitation and temperature despite the differently shaped curves. This coincides with the acceptance rate, shown as the black lines in panels c–d, first dropping below 1/2. Total population growth accelerates from generations 1–5 as the algorithm has to try more to produce successful children, and slows down thereafter. We speculate that raising the acceptance rate might improve the overall efficiency and therefore return period. This might be done by adaptively decreasing the advance split time as the algorithm progresses. More deliberate choices of perturbations might also help to increase acceptance, but with important implications on the assignment of weights. These strategies are beyond our current scope, but we suggest the diagnostics in Figure 5 as helpful in pursuit of them.

4.3. Sensitivity Analysis of Advance Split Time

Figure 6 quantifies the variation in performance with δ using two simple performance indicators. The first measures *statistical* accuracy in high return levels:

$$L^2 \text{ error} = \left(\frac{1}{\log(\tau_{\max}/\tau_{\min})} \int_{\tau_{\min}}^{\tau_{\max}} [\hat{s}_{\text{DNS}}(\tau) - \hat{s}_{\text{TEAMS}}(\tau)]^2 \, d[\log \tau] \right)^{1/2} \quad (5)$$

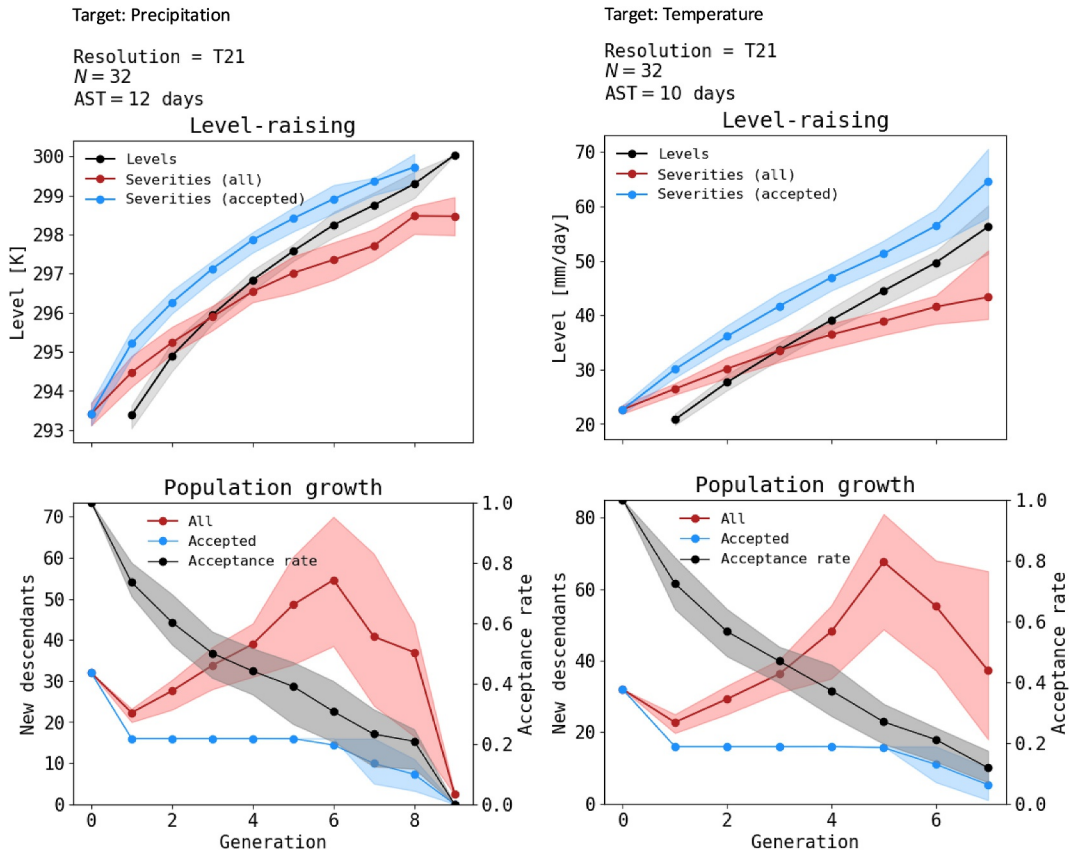

Figure 5. Progression of the population size and severities during TEAMS for both precipitation (a, c) and temperature (b, d) targets at T21 resolution and $N = 32$. (a, b) The levels s (gray), the mean severities across all new descendants born that generation (red), and the mean severities across accepted new descendants only (blue) in a run, as a function of "generation" or how many levels have been set so far. Lines and shaded bands show means and interquartile ranges across the 48 runs. (c, d) Population growth per generation (red), accepted population growth (blue), and acceptance rate (black) at each round of level-raising. Population growth accelerates from generations 1–5 and declines thereafter.

where τ is a return period running from $\tau_{\min} = 50$ days to $\tau_{\max} = 1.6 \times 10^4$ years, and $\hat{s}_{(\mathrm{DNS,TEAMS})}(\tau)$ represents the corresponding severity return level estimated by (DNS, TEAMS) by inverting the estimator $\hat{\tau}(s)$ in Equation 4 with linear (in log τ 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 Figure 3), we extrapolate $\hat{s}_{\mathrm{TEAMS}}$ 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:

$$\mathrm{Boost} = \frac{1}{M}\sum_{m=1}^{M} \max\{\max(S_\ell - S_m, 0) : X_\ell \text{ is a descendant of } X_m\} \tag{6}$$

where M is the total number of ensemble members, including all ancestors and all accepted descendants (but not rejects). Figure 6 shows both performance indicators' δ-dependance, and adds to the growing collection of examples (Bloin-Wibe et al., 2025; Finkel & O'Gorman, 2024, 2025) demonstrating 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 Figure 7. Thus it appears that the appropriate target time is not universal but rather depends, at least weakly, on the choice of target variable. Heuristically, smoother fields like temperature should take longer to decorrelate, and

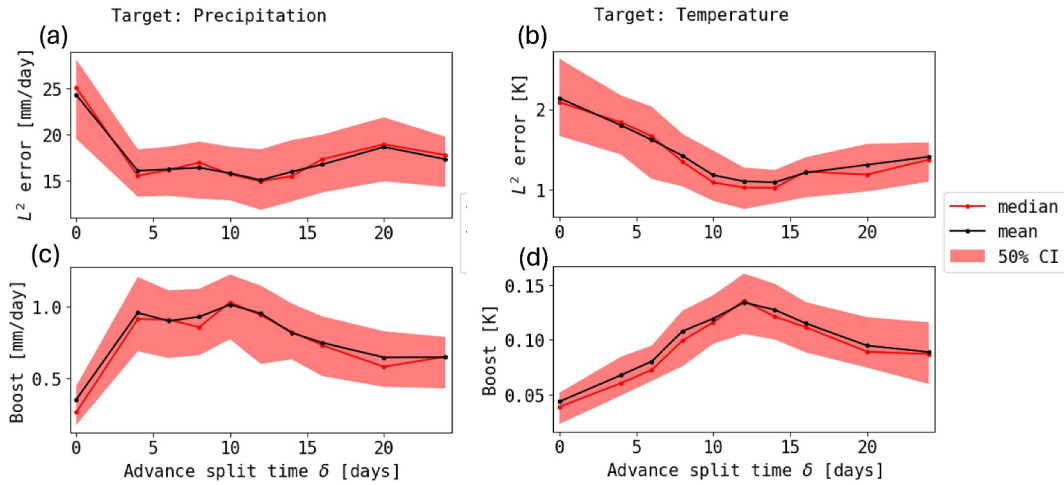

Figure 6. TEAMS performance diagnostics as functions of advance split time for T21 resolution and $N = 16$. We deployed TEAMS on two different target variables (left: precipitation and right: temperature) with a sequence of ASTs of 0, 4, 6, 8, 10, 12, 14, 16, 20, and 24. Each case was repeated 48 times 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 Figure 3 (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.

therefore call for a longer advance split time—at least, when our event of interest is a *single-time maximum* which is the setting where TEAMS is helpful. The next section bears this out quantitatively.

4.4. Ensemble Spreading Rate

Finkel and O'Gorman (2024) found that the optimal δ was well estimated as the time $t_{3/8}$ after which a perturbed ensemble disperses to a fraction 3/8 of its saturation dispersion. To test the generalization of this rule, we now compare the optimal AST found by grid search in the previous section with $t_{3/8}$ for the GCM, computed by the branching procedure specified below (same as in Finkel and O'Gorman (2024)). The results demonstrate that saturation dispersion remains a useful guide, but the specific value 3/8 needs quantitative adjustment. The branching procedure is as follows:

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 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.
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 data set

$$\{X_{b,w}(r) : 1 \le b \le B, 1 \le w \le W, 0 \le r \le T_B\} \tag{7}$$

(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 $w = 1, \ldots, W$ in terms of the RMSE as a function of the elapsed time r since the split:

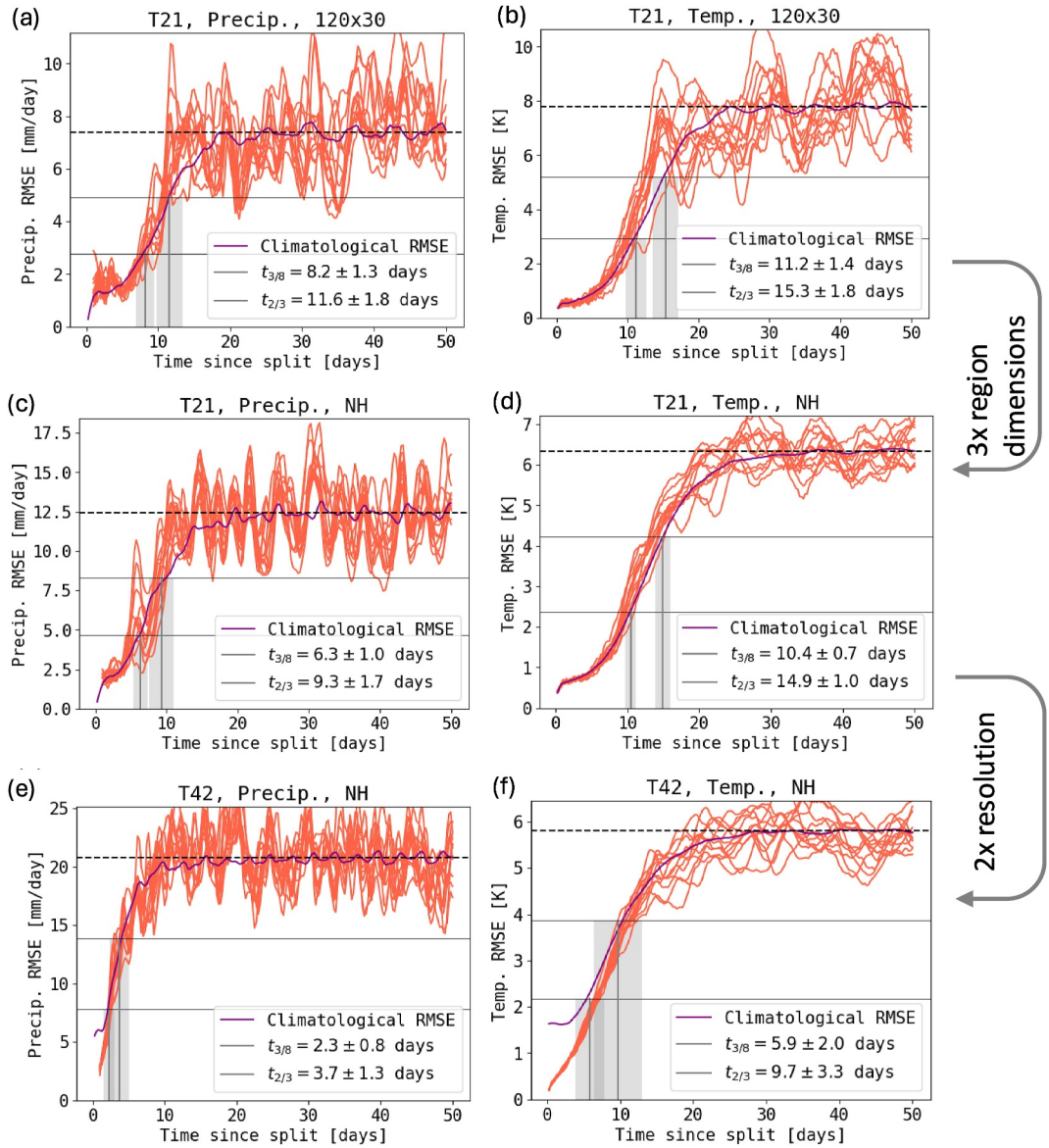

Figure 7. Ensemble dispersion measured using precipitation (a, c, e) and surface temperature (b, d, f) fields. Shown in red is the area-weighted Euclidean distance (RMSE) between each realization and the control based on an average over a $120^\circ \times 30^\circ$ longitude × latitude region centered on the target (a, b) or the full northern hemisphere (c–f). The RMS of 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 two horizontal gray thresholds mark the fractions 3/8 and 2/3 of saturation, and the vertical gray lines with error bars delineate the means and standard deviations of $t_{3/8}$ and $t_{2/3}$, the threshold-crossing time, across whorls as $\overline{t_\epsilon} \pm \text{std}(t_\epsilon)$. Results are shown for T21 resolution (a–d) and T42 resolution (e, f).

$$\text{RMSE}_w(r) = \sqrt{\frac{1}{B}\sum_{b=1}^{B} D(X_{w,b}(r), X_{w,0}(r))^2} \tag{8}$$

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 either the entire Northern Hemisphere or a smaller region centered on the target location ($120^\circ \times 30^\circ$ longitude × latitude), which might be more relevant to the event of interest but

also more noisy. Figure 7 displays the results in the form of RMSEs, both with the smaller area at T21 (a, b), the full northern hemisphere at T21 (c, d) and the full northern hemisphere at T42 (e, f). 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 $\mathrm{RMSE}_w(r)$ across all ws—i.e., $\sqrt{\frac{1}{W}\sum_w \mathrm{RMSE}_w^2(r)}$, denoted $\mathrm{RMSE}(r)$—is displayed as purple lines in Figure 7, and we estimate the asymptotic RMSE by its final 15-day average. Define the fractional saturation time $t_{\epsilon,w}$ as the time r 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 $\overline{t_{3/8}} := \frac{1}{W}\sum_{w=1}^{W} t_{3/8,w}$. For the GCM, we actually find $\overline{t_{2/3}}$ to be a better guide, and both are marked in Figure 7. The order of averaging is important: $\overline{t_\epsilon}$ is not exactly the same as the time that $\mathrm{RMSE}(r)$ crosses $\epsilon \times$(saturation RMSE), but they are practically indistinguishable for the regions considered here. A benefit of averaging times first instead of RMSEs first is that it gives a straightforward estimate of standard deviation of t_ϵ across ws, which is denoted in the legends along with the mean $\left[t_\epsilon = \overline{t_\epsilon} \pm \mathrm{std}(t_\epsilon)\right]$ for both $\epsilon = 3/8$ and $\epsilon = 2/3$.

The first thing to notice is that RMSE saturates more slowly when measured by temperature instead of precipitation, which is consistent with the previous section's result that optimal AST is longer for temperature than for precipitation. This is unsurprising as temperature is generally a smoother field. In addition, the saturation timescales change with the area chosen for averaging. At T21, going from the full NH to the $120^\circ \times 30^\circ$ region, $t_{3/8}$ increases from 6.3 to 8.2 days for precipitation, and from 10.4 to 11.2 days for temperature. The true optimal AST is in fact *longer* than both these values: 10 days for precipitation and 12 days for temperature. We might shrink the averaging region further to make $t_{3/8}$ the right predictor, but this would be ad hoc and subject to increasing noise. Instead we adhere to a full-NH notion of distance, which is roughly analogous to what we considered in the Lorenz 96 system. This choice gives a new empirical saturation fraction of $\sim 2/3$ as a better match for optimal AST. For $t_{2/3}$ we find 9.3 days for precipitation and 14.9 for temperature, which is roughly consistent with the grid search optimal values given the breadth of the valleys in L^2 error in Figure 6.

It is with this new rule that we deployed TEAMS on the most expensive test case: T42 resolution and $N = 32$ ancestors. Figures 7e and 7f shows that the dispersion timescale at T42 is well shorter than at T21. We selected $\delta = 4$ and 10 days as the approximate $t_{2/3}$ values for precipitation and temperature, respectively, which yielded the results shown in Figures 3e and 3f. We take encouragement from the fact that even rough estimates for δ, with rounding, give substantial speedups.

Clearly, the general problem of optimizing AST is not yet solved. The leading Lyapunov exponent (Cencini & Ginelli, 2013) is a natural first guess for optimal AST, but it only pertains to infinitesimal errors, whereas we aim for finite-amplitude boosts. Furthermore, the Lyapunov exponents' property of being intrinsic to the system actually make them incapable of adjusting to different targets, which we have clearly shown is necessary. Note also that ensemble dispersion timescale depends upon the magnitude of stochastic forcing, as shown in Finkel and O'Gorman (2024), and would also change if we were to use a deterministic model with small one-off kicks instead. Such factors raise doubts on whether Lyapunov exponents are applicable, but ensemble dispersion is at least clearly defined and measurable in all these cases. There are theoretical results emerging related to an optimal AST for maximizing chosen combinations of moments of a boosted distribution (appendix B of Bloin-Wibe et al., 2025), as well as tentative general rules for an optimal AST based on Bayesian optimization (Finkel & O'Gorman, 2025), but the results here show the need to customize AST for the model, the resolution, and the target, which are important nuances to bear in mind when expanding to other applications, especially those with different spatiotemporal scales such as mesoscale convective systems. Yet our results hint at *scaling relations* between resolution, regions for calculating RMSE, fractional saturation time, and optimal AST. Here we only aim to show that strong speedups are achievable in a GCM, but a more general calibration method is worth pursuing.

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 addressed this problem by augmenting a standard algorithm, adaptive multilevel splitting (AMS) with

early perturbations, resulting in "trying early AMS" (TEAMS). 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 $\sim$20 years of simulation and $300 - 500$ years with only $\sim$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 severe 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 found a similar rule gave good performance for this more complex, albeit idealized, atmospheric model. The performance held for two different target variables (heavy precipitation and heat extremes), and two different resolutions (T21 and T42 horizontal triangular spectral truncation). 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 data sets 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 (Equation 3) pointwise to maps, which has been done for seasonal heat extremes in, for example, Ragone et al. (2018), Ragone and Bouchet (2021), Miloshevich et al. (2024), Le Priol et al. (2024), and Noyelle et al. (2025). 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. For the sake of brevity and to focus on the main point of algorithm development, we leave detailed spatial dynamical analysis to future work.

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.

Conflict of Interest

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

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, 2025a, available at https://doi.org/10.5281/zenodo.16878347), contains the core Fortran model code

2. "TEAMS" (justinfocus12, 2025b, 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.

Acknowledgments
We thank Judith Berner for assistance with implementing the stochastic parameterization. We also extend thanks to three anonymous reviewers for insightful feedback that helped to strengthen the paper substantially. 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.

References

Abbot, D. S., Webber, R. J., Hadden, S., Seligman, D., & Weare, J. (2021). Rare event sampling improves Mercury instability statistics. *The Astrophysical Journal*, *923*(2), 236. https://doi.org/10.3847/1538-4357/ac2fa8

Anderson, J. L., Balaji, V., Broccoli, A. J., Cooke, W. F., Delworth, T. L., Dixon, K. W., 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. https://doi.org/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. 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. https://doi.org/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. https://doi.org/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. https://doi.org/10.5194/egusphere-2025-525

Cencini, M., & Ginelli, F. (2013). Lyapunov analysis: From dynamical systems theory to applications. *Journal of Physics A: Mathematical and Theoretical*, *46*(25), 250301. https://doi.org/10.1088/1751-8113/46/25/250301

Cérou, F., & Guyader, A. (2007). Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, *25*(2), 417–443. https://doi.org/10.1080/07362990601139628

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

Finkel, 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. https://doi.org/10.1029/2024MS004264

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

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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.strusafe.2015.12.003

justinfocus12. (2025a). justinfocus12/jf_conv_gray_smooth: Initial release for submission [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.16878347

justinfocus12. (2025b). justinfocus12/teams: Initial release for submission [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.16878339

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), 204. https://doi.org/10.3390/cli12120204

Krishnamurti, T. N., Hardiker, V., Bedi, H., & Ramaswamy, L. (2006). *An introduction to global spectral modeling*. Springer.

Le Priol, C., Monteiro, J. M., & Bouchet, F. (2024). Using rare event algorithms to understand the statistics and dynamics of extreme heatwave seasons in South Asia. *Environmental Research: Climate*, *3*(4), 045016. https://doi.org/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. https://doi.org/10.1017/jfm.2020.293

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

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Elms, J., et al. (2024a). 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., et al. (2024b). 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. https://doi.org/10.1017/eds.2024.7

Norwood, A., Kalnay, E., Ide, K., Yang, S.-C., & Wolfe, C. (2013). 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. https://doi.org/10.1088/1751-8113/46/25/254021

Noyelle, R., Caubel, A., Meurdesoif, Y., Faranda, D., & Yiou, P. (2025). Evolution of the dynamics of centennial hot summers in Western Europe with climate change. *Geophysical Research Letters*, *52*(14), e2025GL115552. https://doi.org/10.1029/2025GL115552

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. https://doi.org/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. https://doi.org/10.1175/2009JCLI2701.1

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

Palmer, T. N., & Zanna, L. (2013). Singular vectors, predictability and ensemble forecasting for weather and climate. *Journal of Physics A: Mathematical and Theoretical*, *46*(25), 254018. https://doi.org/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. 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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1017/jfm.2021.957

Rose, D. C., Mair, J. F., & Garrahan, J. P. (2021). A reinforcement learning approach to rare trajectory sampling. *New Journal of Physics*, *23*(1), 013013. https://doi.org/10.1088/1367-2630/abd7bd

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. https://doi.org/10.1098/rspa.2019.0834

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L. V., Hegerl, G., et al. (2017). Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and Climate Extremes*, *18*, 65–74. https://doi.org/10.1016/j.wace.2017.10.003

Tagle, F., Berner, J., Grigoriu, M. D., Mahowald, N. M., & Samorodnitsky, G. (2016). Temperature extremes in the community atmosphere model with stochastic parameterizations. *Journal of Climate*, *29*(1), 241–258. https://doi.org/10.1175/JCLI-D-15-0314.1

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. https://doi.org/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. https://doi.org/10.1063/1.5081461

Winget, M., & Persky, A. M. (2022). A practical review of mastery learning. *American Journal of Pharmaceutical Education*, *86*(10), ajpe8906. https://doi.org/10.5688/ajpe8906

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

Zeder, J., Sippel, S., Pasche, O. C., Engelke, S., & Fischer, E. M. (2023). The effect of a short observational record on the statistics of temperature extremes. *Geophysical Research Letters*, *50*(16), e2023GL104090. https://doi.org/10.1029/2023GL104090

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. 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. https://doi.org/10.1146/annurev-biophys-070816-033834