Editors' Suggestion | Featured in Physics

AI-Boosted Rare Event Sampling to Characterize Extreme Weather

Amaury Lancelin,[1,2,*] Alexander Wikner,[3,*] Laurent Dubus,[2,4] Clément Le Priol,[1,†] Dorian S. Abbot,[3,‡] Freddy Bouchet,[1,§] Pedram Hassanzadeh,[3,‖] and Jonathan Weare[5,¶]

[1]*LMD/IPSL, CNRS, ENS, Université PSL, École Polytechnique, Institut Polytechnique de Paris, Sorbonne Université, Paris, France*
[2]*Réseau de Transport d'Électricité (RTE), Paris, France*
[3]*Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois, USA*
[4]*World Energy & Meteorology Council (WEMC), Norwich, United Kingdom*
[5]*Courant Institute, New York University, New York, New York, USA*

(Received 21 January 2026; accepted 8 June 2026; published 5 August 2026; corrected 5 November 2026)

Weather extremes pose major societal risks, especially in a changing climate, but due to their rarity, they are difficult to study using limited observations or complex climate models. We introduce AI + RES, a framework coupling fast AI weather forecasts with a high-fidelity physics model using a rare-event algorithm to efficiently characterize extremes. This approach enables the study of the statistics and physics of very rare events, such as once per millennium heat waves at two orders-of-magnitude lower computational cost. AI + RES can be applied broadly across climate science and other fields concerned with rare events.

DOI: 10.1103/b1gc-9c2q

Rare, extreme weather and climate events—such as heat waves, floods, and hurricanes—have enormous socioeconomic impacts [1,2]. Accurately assessing how their frequency and intensity change in a warming climate is essential for developing effective mitigation and adaptation strategies [3–7]. However, it is difficult to estimate the frequency and potential impact of the rarest events from observational data [8]. Extreme value theory [9] can be used to extrapolate from short historical records, but it is often overwhelmed by uncertainty [8,10–12] and does not provide examples of the event of interest for physical interrogation. High-fidelity physics-based global climate model (GCM) simulations can provide additional data, but accurate sampling requires simulation lengths at least ten times longer than the return time of the event of interest, which is computationally infeasible. In response to these computational constraints, rare event sampling (RES) techniques and artificial intelligence (AI) emulators have been developed. While promising, both methods have proven insufficient on their own to fully resolve the problem.

RES is a set of importance sampling tools that focus computation on the rare event of interest [12–18]. RES entails scheduled duplication ("splitting") or termination ("killing") of members of an ensemble of independently evolving model simulations to promote efficient sampling of the targeted rare event. The probability that any one ensemble member (a trajectory) is duplicated or terminated is determined by a user-chosen *score function* evaluated at the scheduled resampling times. Choosing a score function typically requires extensive domain knowledge, trial and error, and parameter optimization with costs that may outweigh any advantages of using RES over direct simulation. Recent work has also explored using ensemble boosting, a cousin of RES, to study unprecedented extreme events [19,20]. This approach has recently been extended to allow for probability estimation by adopting an RES approach with a one-step splitting scheme [20,21].

RES has proven highly effective for sampling extremes of long-duration weather, e.g., seasonal means [12–14,18,22], by using *persistence*, i.e., the *current* value of the index variable defining the event, as a score function. This approach, however, fails catastrophically when the current value of the index variable is a poor predictor of its future value [14,16]. This is the case for many phenomena in complex atmospheric dynamics, such as blocking-driven heat waves, which have led to some of the most socioeconomically impactful extreme events in recent memory in France, Russia, and the U.S. Pacific Northwest [23–26]. Previous work has therefore established the strong potential of RES, but only if an efficient

*These authors contributed equally to this work.
†Now at AXA Climate, Paris, France.
‡Contact author: abbot@uchicago.edu
§Contact author: freddy.bouchet@lmd.ipsl.fr
‖Contact author: pedramh@uchicago.edu
¶Contact author: weare@nyu.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

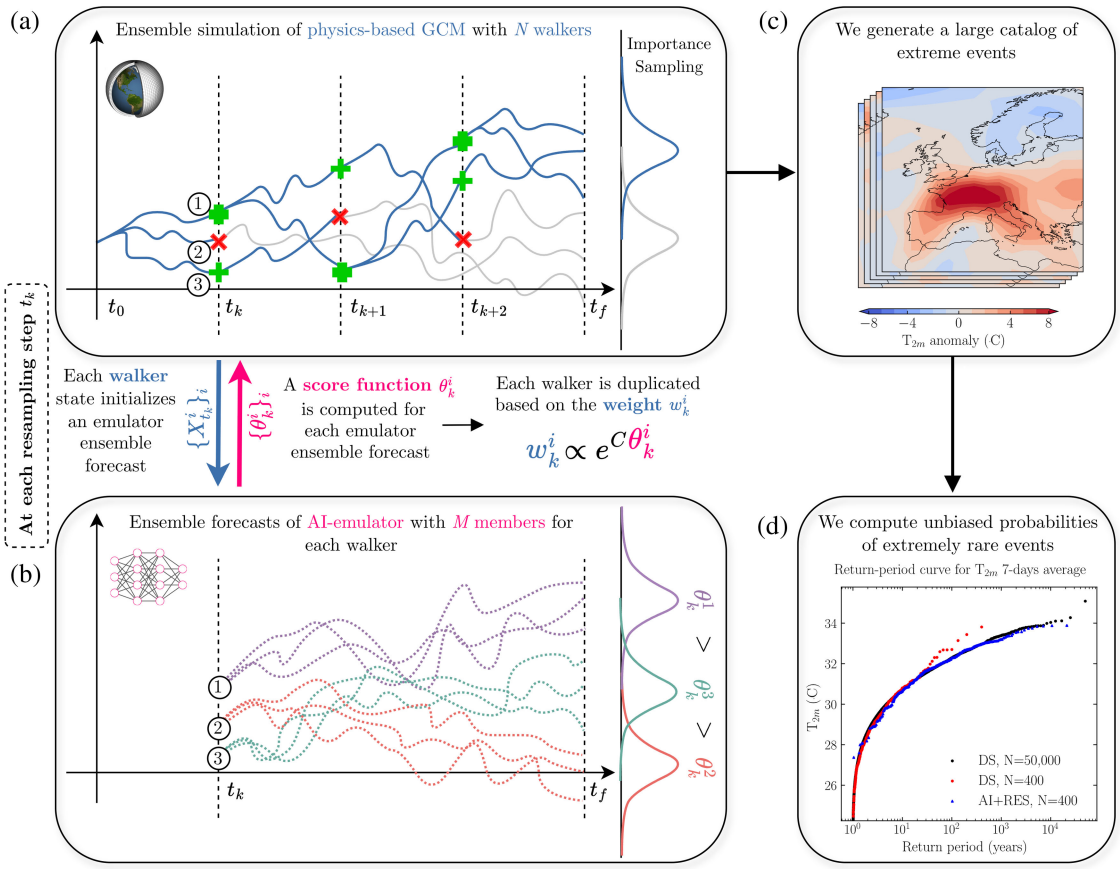


FIG. 1. Schematic of the AI + RES framework. We first train an AI weather emulator on 100 years of PlaSim GCM data and then use it to guide a rare event sampling (RES) algorithm. (a) In AI + RES, we run a PlaSim ensemble simulation with N parallel simulations, called *walkers* and denoted X^i_t. (b) At each resampling time t_k in the algorithm (vertical dashed lines), we perform an ensemble forecast with the AI emulator for each PlaSim walker until the target time t_f and then duplicate the more promising walkers based on these forecasts [Algorithm 1 and Eq. (C1)]. This allows us to both generate a large catalog of extreme events from physics-based simulations (c) and compute unbiased probabilities for these rare events [Eq. (B1)] (d) at a fraction of the cost of direct sampling (DS). An example of schematic (a) with actual data from an AI + RES experiment can be found in Supplemental Material (SM) Fig. S9 [46].

method for finding an effective and accessible score function can be identified.

One of the most exciting recent developments in climate science and scientific AI has been the introduction of autoregressive AI weather emulators [27–30]. When trained on high-resolution reanalysis data, these models can outperform state-of-the-art physics-based numerical weather prediction systems for short- and medium-range (∼10–15 day) forecasts [28,29,31]. Their extreme computational efficiency (up to 10^4 times faster than traditional models [27]) makes them appealing for studying rare extreme weather. More recently, stable AI emulators have been extended to climate timescales, enabling extremely long simulations or large ensembles in which rare events could, in principle, be sampled directly [32–36]. However, because these models are trained to predict at short lead times, and only on historical data or expensive high-fidelity simulations that contain few examples of the most extreme events, they often struggle to extrapolate reliably and to reproduce the correct frequencies of very rare extremes [37–40]. More broadly, beyond having good standard forecast metrics and reproducing stationary statistics, emulators may struggle to capture physically meaningful forced responses under perturbation—arguing for the need for response-theoretic diagnostics [41].

In this Letter, we present AI + RES, a robust algorithm that leverages the strong forecast capabilities of AI emulators to build an effective score function for RES to characterize rare, extreme events. This algorithm is described in Fig. 1 and discussed in detail in the End Matter. We demonstrate this new framework by using it to simulate extreme heat waves in an intermediate complexity GCM, PlaSim [42], which allows for extensive long-term control runs to compute baseline rare event statistics and rigorously evaluate the performance of our approach. For validation purposes, PlaSim is run in a stationary climate configuration. Heat waves are a well-motivated application area because they are the deadliest extreme weather events and are expected to worsen under climate change [43–45]. We show that AI + RES provides accurate rare event

statistics at a numerical speedup of up to a factor of $\mathcal{O}(100)$, whereas standard RES fails to even yield examples of the rarest events of interest. This Letter demonstrates a novel methodology that addresses a central difficulty of RES and may broaden its applicability in fields where a fast AI emulator can be constructed.

Midlatitude heat waves—As mentioned, we focus on sampling rare midlatitude heat waves due to their societal impact and the challenge that they pose to standard RES. Following previous studies [13,47], we consider heat wave events defined by large values of the spatiotemporal average of the 2-meter air temperature, T_{2m}, over a prescribed geographic region and a time interval typically spanning one week during the summer season [see Eq. (A1)]. We consider two such regions, centered over France and over the U.S. Midwest, respectively (SM Fig. S7 [46]). In this Letter, we use *rare* to denote low-probability events and *extreme* to denote very large values of the chosen observable, with substantial overlap between the two. We refer to [48,49] for more formal definitions.

Ground truth and baselines—To evaluate the performance of AI + RES, we compare it to a ground truth and five baselines. The ground truth, *direct sampling* (DS) with $N = 50,000$, is a very large ensemble that is only possible due to the computational efficiency of PlaSim. The first baseline, DS, $N = 400$, is a DS with PlaSim using the same ensemble size as the number of walkers used in the AI + RES algorithm. The second baseline, *Standard*-RES, uses the RES algorithm but with the traditional persistence score function used in state-of-the-art weather and climate applications [12–14,18,22]. The third baseline, AI-DS, is a DS ensemble using the AI emulator started from the same initial conditions as DS and RES, with the same ensemble size as the ground truth ($N = 50,000$). This allows us to determine whether the AI emulator has learned enough from its 100-year training period to extrapolate to tail events in PlaSim. The fourth baseline, EVT, applies extreme value theory, specifically the generalized Pareto distribution (GPD) within a peak-over-threshold framework, to PlaSim datasets of the same size as the AI + RES experiments ($N = 400$). We estimate uncertainty in this method by performing the fits on 100 different datasets of equal size. We discuss a fifth baseline, Perfect-Forecast-System+RES (PFS + RES), in SM Sec. S8.2 [46]. This baseline uses a PlaSim ensemble forecast for the score function and yields an upper bound on algorithm performance given a "perfect emulator." This baseline is only possible with GCMs like PlaSim that are computationally inexpensive.

AI + RES *accurately estimates long return period events*—Figure 2 shows that, for both France and the U.S. Midwest, AI + RES produces accurate, unbiased return period estimates up to our reference ensemble limit of 50,000 years, despite using ensembles with only $N = 400$ walkers. In stark contrast, DS of the same computational cost ($N = 400$) can only produce accurate return

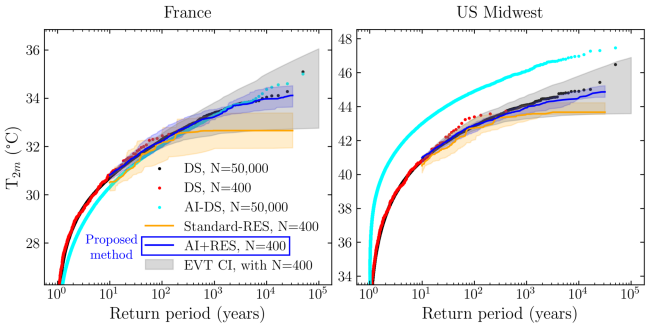


FIG. 2. Return-period curve for a 7-day heat wave (T_{2m} averaged over 7-days) over France (left panel) and the U.S. Midwest (right panel). The solid blue and yellow lines show the median return-period curve from 10 independent realizations of the AI + RES and standard-RES algorithm, respectively; shaded regions indicate the 10th–90th percentiles. Red and Black dots show DS with $N = 50,000$ and $N = 400$. The gray shaded area (EVT) shows the 10th and 90th percentiles obtained by fitting different GPD distributions with 100 independent training datasets with $N = 400$.

period estimates up to about 50 years, 1,000 times shorter. This dramatic difference underscores the utility of RES when an *effective score function* can be identified. Moreover, Standard-RES saturates for return periods longer than about 100 years, falsely suggesting an upper limit on heat wave intensity. This indicates that persistence misses important trajectories to rare heat waves that the AI-based score function is able to identify. The AI-DS baseline is biased both in the mean and variability of the observable, producing erroneous return period estimates even on a decadal timescale. The correct AI-DS estimate at a return period of about 2,000 years in France is an accident of the curves crossing. It is notable that, despite the errors in AI-DS, the AI emulator's weather forecast skill can still serve as a useful score function by effectively ranking progress of walkers toward strong heat waves. Finally, the correct return periods do fall within the uncertainty of the EVT baseline even at the longest timescales, but the uncertainty estimate is up to four times larger than that from AI + RES for return periods on the order of 10,000 years. The EVT estimates depend so strongly on the particular years fed into them that they cannot be relied on to produce accurate return period estimates. Also, EVT, unlike AI + RES, does not provide information about the identified extremes and their dynamics.

AI + RES *provides a* 100× *computational speedup for rare events*—Having established that AI + RES is unbiased, we can quantify its computational speedup compared to DS using the relative efficiency of sampling extreme events [SM Eq. (S7) [46]] and the relative variance of rare probability estimators [SM Eq. (S10)]. Both metrics tend to yield greater computational gains for rarer events (Fig. 3), with speedups of more than 100 for a return period of 10,000 years. The only exception is that the variance

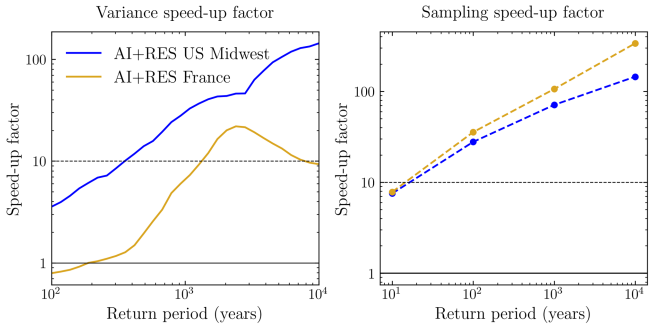


FIG. 3. Speedup factors for the AI + RES algorithm over France and the U.S. Midwest as a function of return period for reducing variance in estimated probability and producing additional samples.

speedup estimate for the France region peaks at a factor of 25 around the 2,500-year return period and then gradually decreases to 10 for the rarest events. This nonmonotonic trend is most likely due to a sampling error and is discussed in more detail in SM Sec. S5.2 [46]. Moreover, we hypothesize that the difference in variance speedup for the two regions is due to regional differences in AI emulator weather forecast skill, potentially due to differences in soil moisture variability, as pointed out in SM Fig. S8 [46].

The rare events generated by AI + RES *are physically realistic*—A major advantage of AI + RES over the EVT baseline is that it generates full physical trajectories leading to the rare event of interest. This is particularly valuable for investigating the dynamics of the rare event and its precursors [18].

In Fig. 4, we illustrate this capability by presenting composite maps associated with heat waves over France with a return period longer than 100 years from the ground-truth DS ($N = 50,000$), samples generated with AI + RES ($N = 400$), and samples generated from the DS with the same ensemble ($N = 400$). A strong precursor pattern emerges in the ground-truth DS three days before the heat wave onset, marked by a pronounced positive 500-hPa geopotential height anomaly over the British Isles, flanked by a negative anomaly over the Labrador Sea. A weaker negative anomaly off the coast of Portugal suggests potential influence from a cutoff low, while a weak negative anomaly is also visible over Eastern Europe. At the heat wave peak, this synoptic pattern shifts: the high-pressure anomaly migrates eastward, the Labrador Sea anomaly propagates into the North Atlantic, and the Eastern European anomaly intensifies. The AI + RES composite shows similar patterns and captures the correct evolution, although it is slightly noisier than the ground truth due to fewer samples. By contrast, the composite from the DS with the same computational budget as AI + RES is significantly noisier and contains incorrect physical patterns. For example, the 3-day-lagged Z_{500} anomaly north of Scandinavia in this composite is a spurious teleconnection likely due to sampling error.

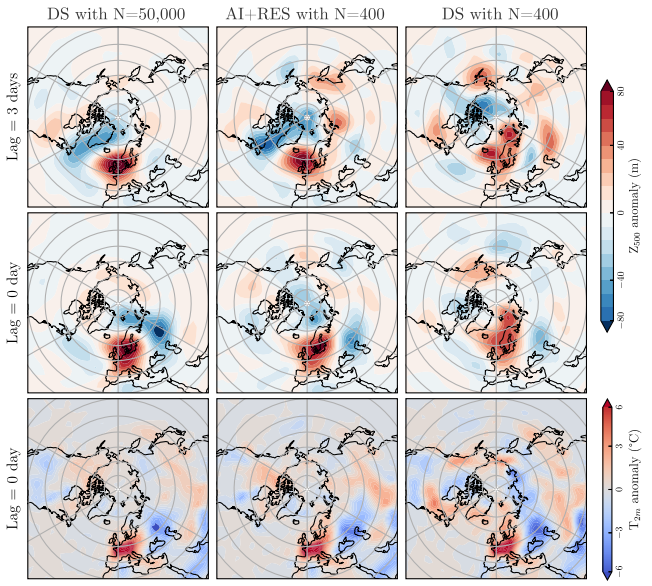


FIG. 4. Composite maps of heat waves over France with return periods exceeding 100 years. Events are defined by Eq. (A1) with $L = 3$ days. From top to bottom, the rows show daily mean Z_{500} anomaly composites three days before the heat wave onset, the L-day average Z_{500} anomaly composites during the heat wave, and the L-day average T_{2m} anomaly composites during the heat wave.

Similarly, SM Fig. S13 [46] shows that AI + RES generates physically realistic samples over the U.S. Midwest that match those from ground truth, while DS with the same computational budget produces samples with spurious features, obscuring the underlying dynamics.

Discussion—To address the challenge posed by the scarcity of data for studying extreme events, we introduced AI + RES, a novel framework that synergistically combines RES, AI emulators, and GCMs such that they complement each other and alleviate each individual approach's shortcomings. AI + RES provides a practical way to address the long-standing challenge of finding an appropriate score function for RES in a high-dimensional chaotic system. As the first proof-of-concept, we applied this framework on midlatitude heat waves in the PlaSim GCM. We showed that this approach enables the sampling of extremely rare events and yields accurate estimates of their return periods up to 50,000 years with an ensemble size of only 400, with a reduction in computational cost of several hundred compared with direct sampling. We also showed that, even with an imperfect emulator, AI + RES significantly outperforms the current state-of-the-art RES that relies on simple score functions.

Our work motivates several avenues for further improvement. First, the deterministic nature of the current emulator limits the quality of its ensemble forecasts. Having an emulator trained directly in a probabilistic manner [30,50–53] could enhance the ability of the algorithm to explore multiple plausible pathways to extreme events, improving sampling efficiency and reducing estimator

variance. Second, while our current AI emulator focuses exclusively on atmospheric variables, surface temperature extremes are also strongly influenced by land surface processes, particularly soil moisture [47,54–56]. Developing a coupled land-atmosphere emulator remains a challenge, but incorporating soil moisture information, potentially through hybrid strategies that combine emulated forecasts with soil moisture state, could yield a more effective score function.

Beyond the specific case of midlatitude heat waves, the potential applications of AI + RES are broad. It could easily be applied to other types of weather extremes—such as tropical cyclones, blocking events, precipitation, wind, or compound events. AI + RES offers a powerful tool to generate targeted catalogs of rare events in poorly sampled regions of the state space, particularly where direct sampling of computationally expensive GCMs offers insufficient coverage. The framework could also be well suited to characterize the tails of distributions in subseasonal to seasonal ensemble forecasts, where large ensembles are needed. Moreover, by applying AI + RES to climate models running under different climate change scenarios (e.g., CMIP6 or the emerging global high-resolution runs), it would be possible to explore the evolution of the statistics of rare extremes in future climates. Note that the AI + RES can be readily applied to any state-of-the-art GCM; here, we focused the proof-of-concept on PlaSim as its computational efficiency allows for validation against a ground-truth dataset. However, this Letter does not address the question of how faithfully PlaSim rare-event statistics represent the real world.

Regarding the AI emulators themselves, the utility of AI + RES is twofold. As the climate modeling community moves toward building emulators for increasingly high-resolution GCMs, obtaining sufficient physically generated samples to accurately train and validate them with physically generated data becomes a key challenge for rare events. Using AI + RES to generate such catalogs of extremes will constitute precious training and validation datasets for the emulators.

Finally, while this Letter focused on weather extremes, our approach should be broadly applicable. AI + RES could be extended beyond climate science to any high-dimensional dynamical system for which an AI surrogate model can be constructed.

The code implementing AI + RES, along with model configurations, analysis scripts, and the data files used to generate the figures in this Letter, is available at [57]. For practical reasons, the full long PlaSim runs will not be publicly posted, but they can be made available upon request from the authors.

Acknowledgments—A. L. was funded in part by RTE France, the French transmission system operator, and benefited from the national agency for research and technology (ANRT) funding under the CIFRE Contract No. 2023/0506. A. W. is supported by the Eric and Wendy Schmidt AI in Science Fellowship, a program of this research received support through Schmidt Sciences, LLC. A. W. and A. L. are also supported by France-Chicago Center FACCTs Award No. 300004248. This work was also supported by the Institut des Mathématiques pour la Planète Terre (IMPT) and funded in part by the U.S. National Science Foundation through Awards No. NSF RISE–2425898, No. NSF-2531264, and No. NSF-2425899. This research is also partially funded by the Institute for Climate and Sustainable Growth at the University of Chicago. In addition, the authors thank Alessandro Lovo for fruitful discussions and insightful ideas on efficiently coupling RES and AI forecasts, which were invaluable to this project; Anna Asch for her work in making the diagnostic plots of the AI emulator; and Justin Finkel for engaging and insightful discussions. Computational resources were provided by NSF ACCESS (allocation ATM170020), NCAR's CISL (allocation UCHI0014), the University of Chicago Research Computing Center, and GENCI-IDRIS (Grant No. 2025-AD010116382).

In alphabetical order, D. S. A., F. B., P. H., A. L., C. L. P., J. W., and A. W. conceptualized the work. D. S. A., F. B., P. H., and J. W. supervised and managed the project. A. L. and A. W. curated the data. A. L. and A. W. developed the codes. A. L. conducted the experiments and did the analysis. A. W. led the development of the AI emulator. A. L. and A. W. wrote the original draft. D. S. A., F. B., L. D., P. H., A. L., C. L. P., J. W., and A. W. reviewed and edited the Letter.

Data availability—The data that support the findings of this article are not publicly available upon publication because it is not technically feasible and/or the cost of preparing, depositing, and hosting the data would be prohibitive within the terms of this research project. The data are available from the authors upon reasonable request.

[1] R. S. Tol, A meta-analysis of the total economic impact of climate change, Energy Policy **185**, 113922 (2024).
[2] J. Anchen, V. B. Gonzalez, M. Chatterjee, R. Egloff, A. Felderer, A. Mejlerö, A. Vischer, and B. Wilke, Swiss Re SONAR: New emerging risk insights, Technical Report, Swiss Re Institute, Zurich, Switzerland, 2025.
[3] K. L. Ebi, J. Vanos, J. W. Baldwin, J. E. Bell, D. M. Hondula, N. A. Errett, K. Hayes, C. E. Reid, S. Saha, J. Spector *et al.*, Extreme weather and climate change: Population health and health system implications, Annual review of public health **42**, 293 (2021).
[4] C. C. Ummenhofer and G. A. Meehl, Extreme weather and climate events with ecological relevance: A review, Phil. Trans. R. Soc. B **372**, 20160135 (2017).
[5] A. C. Gonçalves, X. Costoya, R. Nieto, and M. L. Liberato, Extreme weather events on energy systems:

A comprehensive review on impacts, mitigation, and adaptation measures, Sustainable Energy Res. **11**, 4 (2024).

[6] IPCC, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (Cambridge University Press, Cambridge, England, 2021).

[7] IPCC, *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by H.-O. Pörtner, D. Roberts, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, and B. Rama (Cambridge University Press, Cambridge, England, 2022).

[8] J. Zeder, S. Sippel, O. C. Pasche, S. Engelke, and E. M. Fischer, The effect of a short observational record on the statistics of temperature extremes, Geophys. Res. Lett. **50**, e2023GL104090 (2023).

[9] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events: For Insurance and Finance*, Stochastic Modelling and Applied Probability (Springer, Berlin Heidelberg, 2013).

[10] W. K. Huang, M. L. Stein, D. J. McInerney, S. Sun, and E. J. Moyer, Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (GEV) distributions, Adv. Stat. Climatol. Meteorol. Oceanogr. **2**, 79 (2016).

[11] V. M. Gálfi, T. Bódai, and V. Lucarini, Convergence of extreme value statistics in a two-layer quasi-geostrophic atmospheric model, Complexity **2017**, 5340858 (2017).

[12] C. Le Priol, J. M. Monteiro, and F. Bouchet, Using rare event algorithms to understand the statistics and dynamics of extreme heatwave seasons in South Asia, Environ. Res. **3**, 045016 (2024).

[13] F. Ragone, J. Wouters, and F. Bouchet, Computation of extreme heat waves in climate models using a large deviation algorithm, Proc. Natl. Acad. Sci. U.S.A. **115**, 24 (2018).

[14] R. Webber, D. Plotkin, M. O'Neill, D. Abbot, and J. Weare, Practical rare event sampling for extreme mesoscale weather, Chaos **29**, 053109 (2019).

[15] F. Ragone and F. Bouchet, Rare event algorithm study of extreme warm summers and heatwaves over Europe, Geophys. Res. Lett. **48**, e2020GL091197 (2021).

[16] D. S. Abbot, R. J. Webber, S. Hadden, D. Seligman, and J. Weare, Rare event sampling improves Mercury instability statistics, Astrophys. J. **923**, 236 (2021).

[17] J. Finkel and P. A. O'Gorman, Bringing statistics to storylines: Rare event sampling for sudden, transient extreme events, J. Adv. Model. Earth Syst. **16**, e2024MS004264 (2024).

[18] R. Noyelle, A. Caubel, Y. Meurdesoif, P. Yiou, and D. Faranda, Statistical and dynamical aspects of extremely hot summers in Western Europe sampled with a rare events algorithm, J. Clim. **38**, 4763 (2025).

[19] C. Gessner, E. M. Fischer, U. Beyerle, and R. Knutti, Very Rare Heat Extremes: Quantifying and understanding using ensemble reinitialization, J. Clim. **34**, 6619 (2021).

[20] L. Bloin-Wibe, R. Noyelle, V. Humphrey, U. Beyerle, R. Knutti, and E. Fischer, Estimating return periods for extreme events in climate models through ensemble boosting, EGUsphere **2025**, 1 (2025).

[21] J. Finkel and P. A. O'Gorman, Boosting ensembles for statistics of tails at conditionally optimal advance split times, arXiv:2507.22310.

[22] J. Wouters, R. K. Schiemann, and L. C. Shaffrey, Rare event simulation of extreme European winter rainfall in an intermediate complexity climate model, J. Adv. Model. Earth Syst. **15**, e2022MS003537 (2023).

[23] D. Barriopedro, E. M. Fischer, J. Luterbacher, R. M. Trigo, and R. García-Herrera, The hot summer of 2010: Redrawing the temperature record map of Europe, Science **332**, 220 (2011).

[24] P. A. Stott, D. A. Stone, and M. R. Allen, Human contribution to the European heatwave of 2003, Nature (London) **432**, 610 (2004).

[25] R. H. White, S. Anderson, J. F. Booth, G. Braich, C. Draeger, C. Fei, C. D. Harley, S. B. Henderson, M. Jakob, C.-A. Lau *et al.*, The unprecedented Pacific Northwest heatwave of June 2021, Nat. Commun. **14**, 727 (2023).

[26] V. M. Galfi and V. Lucarini, Fingerprinting heatwaves and cold spells and assessing their response to climate change using large deviation theory, Phys. Rev. Lett. **127**, 058701 (2021).

[27] J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli *et al.*, FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators, arXiv:2202.11214.

[28] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu *et al.*, Learning skillful medium-range global weather forecasting, Science **382**, 1416 (2023).

[29] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, Accurate medium-range global weather forecasting with 3D neural networks, Nature (London) **619**, 533 (2023).

[30] I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia *et al.*, Probabilistic weather forecasting with machine learning, Nature (London) **637**, 84 (2025).

[31] Z. Ben Bouallegue, M. C. Clare, L. Magnusson, E. Gascon, M. Maier-Gerber, M. Janoušek, M. Rodwell, F. Pinault, J. S. Dramsch, S. T. Lang *et al.*, The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context, Bull. Am. Meteorol. Soc. **105**, E864 (2024).

[32] O. Watt-Meyer, B. Henn, J. McGibbon, S. K. Clark, A. Kwa, W. A. Perkins, E. Wu, L. Harris, and C. S. Bretherton, ACE2: Accurately learning subseasonal to decadal atmospheric variability and forced responses, arXiv:2411.11268.

[33] D. Kochkov, J. Yuval, I. Langmore, P. Norgaard, J. Smith, G. Mooers, M. Klöwer, J. Lottes, S. Rasp, P. Düben *et al.*,

Neural general circulation models for weather and climate, Nature (London) **632**, 1060 (2024).
[34] W. E. Chapman, J. S. Schreck, Y. Sha, D. J. Gagne II, D. Kimpara, L. Zanna, K. J. Mayer, and J. Berner, CAMulator: Fast emulation of the community atmosphere model, arXiv:2504.06007.
[35] A. Mahesh, W. D. Collins, B. Bonev, N. Brenowitz, Y. Cohen, J. Elms, P. Harrington, K. Kashinath, T. Kurth, J. North *et al.*, Huge ensembles–Part 1: Design of ensemble weather forecasts using spherical Fourier neural operators, Geosci. Model Dev. **18**, 5575 (2025).
[36] A. Mahesh, W. D Collins, B. Bonev, N. Brenowitz, Y. Cohen, P. Harrington, K. Kashinath, T. Kurth, J. North, T. A. O'Brien *et al.*, Huge ensembles–Part 2: Properties of a huge ensemble of hindcasts generated with spherical Fourier neural operators, Geosci. Model Dev. **18**, 5605 (2025).
[37] Y. Q. Sun, P. Hassanzadeh, M. Zand, A. Chattopadhyay, J. Weare, and D. S. Abbot, Can AI weather models predict out-of-distribution gray swan tropical cyclones?, Proc. Natl. Acad. Sci. U.S.A. **122**, e2420914122 (2025).
[38] Y. Q. Sun, P. Hassanzadeh, T. Shaw, and H. A. Pahlavan, Predicting Beyond Training Data via Extrapolation versus Translocation: AI Weather Models and Dubai's Unprecedented 2024 Rainfall, arXiv (2025).
[39] Z. Zhang, E. Fischer, J. Zscheischler, and S. Engelke, Numerical models outperform AI weather forecasts of record-breaking extremes, arXiv:2508.15724.
[40] A. Wikner, A. Lancelin, T. Arcomano, D. P. Karan Jakhar, F. Bouchet, and P. Hassanzadeh, Can AI Climate Emulators Quantify the Statistics of the Rarest Unseen Weather Extremes?, (AGU Annual Meeting 2025, New Orleans, LA, 2025), Abstract NG24A-06.
[41] F. Falasca, Probing forced responses and causality in data-driven climate emulators: Conceptual limitations and the role of reduced-order models, Phys. Rev. Res. **7**, 043314 (2025).
[42] K. Fraedrich, H. Jansen, E. Kirk, U. Luksch, and F. Lunkeit, The planet simulator: Towards a user friendly model, Meteorol. Z. **14**, 299 (2005).
[43] Q. Zhao, Y. Guo, T. Ye, A. Gasparrini, S. Tong, A. Overcenco, A. Urban, A. Schneider, A. Entezari, A. M. Vicedo-Cabrera *et al.*, Global, regional, and national burden of mortality associated with non-optimal ambient temperatures from 2000 to 2019: A three-stage modelling study, Lancet Planet. Health **5**, e415 (2021).
[44] R. Newman and I. Noy, The global costs of extreme weather that are attributable to climate change, Nat. Commun. **14**, 6103 (2023).
[45] V. Thompson, D. Mitchell, G. C. Hegerl, M. Collins, N. J. Leach, and J. M. Slingo, The most at-risk regions in the world for high-impact heatwaves, Nat. Commun. **14**, 2152 (2023).
[46] See Supplemental Material at http://link.aps.org/supplemental/10.1103/b1gc-9c2q for details on the PlaSim GCM configuration, AI emulator architecture and training, RES algorithm implementation and hyperparameters, additional baselines (including PFS + RES), speed-up metrics, and supplementary figures.
[47] G. Miloshevich, B. Cozian, P. Abry, P. Borgnat, and F. Bouchet, Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data, Phys. Rev. Fluids **8**, 040501 (2023).
[48] H. Kantz, E. G. Altmann, S. Hallerberg, D. Holstein, and A. Riegert, Dynamical interpretation of extreme events: Predictability and predictions, in *Extreme Events in Nature and Society* (Springer, New York, 2006), pp. 69–93.
[49] M. Ghil, P. Yiou, S. Hallegatte, B. D. Malamud, P. Naveau, A. Soloviev, P. Friederichs, V. Keilis-Borok, D. Kondrashov, V. Kossobokov, O. Mestre, C. Nicolis, H. W. Rust, P. Shebalin, M. Vrac, A. Witt, and I. Zaliapin, Extreme events: Dynamics, statistics and prediction, Nonlinear Processes Geophys. **18**, 295 (2011).
[50] F. Alet, I. Price, A. El-Kadi, D. Masters, S. Markou, T. R. Andersson, J. Stott, R. Lam, M. Willson, A. Sanchez-Gonzalez *et al.*, Skillful joint probabilistic weather forecasting from marginals, arXiv:2506.10772.
[51] S. Lang, M. Alexe, M. C. Clare, C. Roberts, R. Adewoyin, Z. B. Bouallègue, M. Chantry, J. Dramsch, P. D. Dueben, S. Hahner *et al.*, AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score, arXiv:2412.15832.
[52] G. Couairon, R. Singh, A. Charantonis, C. Lessig, and C. Monteleoni, ArchesWeather & ArchesWeatherGen: A deterministic and generative model for efficient ML weather forecasting, arXiv:2412.12971.
[53] A. Zhou, A. Wikner, A. Lancelin, P. Hassanzadeh, and A. B. Farimani, Reframing generative models for physical systems using stochastic interpolants, arXiv:2509.26282.
[54] F. D'Andrea, A. Provenzale, R. Vautard, and N. De Noblet-Decoudré, Hot and cool summers: Multiple equilibria of the continental water cycle, Geophys. Res. Lett. **33**, L24807 (2006).
[55] E. M. Fischer, S. I. Seneviratne, P. L. Vidale, D. Lüthi, and C. Schär, Soil Moisture–Atmosphere Interactions during the 2003 European Summer Heat Wave, J. Clim. **20**, 5081 (2007).
[56] R. Vautard, P. Yiou, F. D'Andrea, N. de Noblet, N. Viovy, C. Cassou, J. Polcher, P. Ciais, M. Kageyama, and Y. Fan, Summertime European heat and drought waves induced by wintertime Mediterranean rainfall deficit, Geophys. Res. Lett. **34**, L07711 (2007).
[57] https://github.com/amaurylancelin/AI-RES-public.
[58] P. Del Moral, *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications* (Springer, New York, 2004).
[59] J.-C. Deville and Y. Tille, Unequal probability sampling without replacement through a splitting method, Biometrika **85**, 89 (1998).
[60] A. Chattopadhyay, E. Nabizadeh, and P. Hassanzadeh, Analog forecasting of extreme-causing weather patterns using deep learning, J. Adv. Model. Earth Syst. **12**, e2019MS001958 (2020).
[61] J. Finkel, R. J. Webber, E. P. Gerber, D. S. Abbot, and J. Weare, Learning forecasts of rare stratospheric transitions from short simulations, Mon. Weather Rev. **149**, 3647 (2021).

[62] J. Finkel, E. P. Gerber, D. S. Abbot, and J. Weare, Revealing the statistics of extreme events hidden in short weather forecast data, AGU Adv. **4**, e2023AV000881 (2023).
[63] V. Mascolo, A. Lovo, C. Herbert, and F. Bouchet, Gaussian framework and optimal projection of weather fields for prediction of extreme events, J. Adv. Model. Earth Syst. **17**, e2024MS004487 (2025).
[64] A. Lovo, A. Lancelin, C. Herbert, and F. Bouchet, Tackling the accuracy–interpretability trade-off in a hierarchy of machine learning models for the prediction of extreme heatwaves, Artif. Intell. Earth Syst. **4**, 240094 (2025).
[65] J. Finkel, D. S. Abbot, and J. Weare, Path properties of atmospheric transitions: Illustration with a low-order sudden stratospheric warming model, J. Atmos. Sci. **77**, 2327 (2020).
[66] F. Cérou, Genetic genealogical models in rare event analysis, Ph.D. thesis, INRIA, 2006.

Correction: The allocation number for NCAR's CISL in the Acknowledgments section contained an error and has been fixed.

End Matter

Definition of the heat wave index—Following previous studies [13,47], we define the heat wave index $A_L(t)$ as a spatiotemporal average of 2-meter air temperature, T_{2m}, as in the following:

$$A_L(t) \coloneqq \frac{1}{L}\int_t^{t+L}\left(\frac{1}{\mathcal{R}}\int_{\mathcal{R}} T_{2m}(\vec{r},s)d\vec{r}\right)ds. \quad \text{(A1)}$$

We want to estimate the return times of the events $A_L(t_f) > a$ for large thresholds a. We typically use $L = 7$ days, a window size small enough to capture the peak of midlatitude heat waves yet large enough to have significant societal impacts. In this Letter, we consider two regions, $\mathcal{R}$, each defined as a 3×3 grid point domain centered over France and the U.S. Midwest (SM Fig. S7 [46]). We initialize simulations at $t_0 =$ July 2 and integrate forward the PlaSim GCM until $t_f + L$, with $t_f =$ August 1, chosen to coincide with the climatological peak of $A_L(t_f)$.

AI + RES *with diffusion Monte Carlo*—For the RES component of AI + RES, we use a version of the *diffusion Monte Carlo* (DMC) algorithm that closely follows the formulations proposed in [14] and [16]. It relies on a one-dimensional *score function* (or *reaction coordinate*), θ: $\mathbb{R}^d \to \mathbb{R}$, which assigns higher values to regions of phase space associated with the rare event of interest. The algorithm enhances sampling in regions where θ is high (via duplication or "splitting") and suppresses sampling where it is low (via termination or "killing").

The mathematical properties of DMC have been rigorously studied—see, for instance, [58]—with results establishing its convergence and asymptotic behavior as the ensemble size N tends to infinity. Under mild integrability assumptions, DMC provides unbiased estimators that converge in the large-N limit. These convergence results are quite general and remain valid even for systems of arbitrarily high dimension d. Any statistical quantity that can be computed via direct sampling can, in principle, also be estimated using DMC. This includes observables that depend on the full trajectory of the process from the initial time t_0 up to a given time t_f. Although those convergence properties are very general, the choice of the score function can greatly influence the variances of the estimators obtained with DMC, highlighting the importance of this Letter.

We denote by $(X_t^n)_{1\le n\le N}$ the N realizations generated during the simulation and refer to each sample X_t^n as a *walker*. The DMC-based AI + RES splitting algorithm shown in Fig. 1(a) proceeds as follows: we simulate N walkers in parallel with the physics-based model (the PlaSim GCM, in our case), starting at time $t_0 = 0$ with small initial condition perturbations. An iterative duplication process is performed at fixed times $t_1 < \cdots < t_K = t_f$, called *resampling times* and marked by dashed vertical lines in Fig. 1. These may be spaced either uniformly or nonuniformly; we chose here to perform resampling at uniform time intervals of length τ, where τ is a tunable parameter. For notational convenience, we set $X_k^i \coloneqq X_{t_k}^i$. The resampling steps are governed by a sequence of splitting functions V_k, which are themselves functions of the score function θ (itself obtained from the AI emulator forecasts, in the case of AI + RES; see next section). Walkers with large scores are more likely to be duplicated, while walkers with small scores are more likely to be terminated. Following [16], we choose a splitting function of the form $V_k = C_k\theta(X_k)$, where $C_k > 0$ is a constant that can vary at each resampling step. In practice, the score function is rescaled before computing the splitting function; see SM Sec. S4.1 [46] for more details. The choice of C_k is critical as it controls the strength of the selection process; we discuss it in SM Sec. S4.2 [46]. After resampling, we continue to simulate the selected walkers until the next resampling time, when the process is repeated, until the walkers reach the final time horizon $t_f = t_K$. Each walker is accompanied by a weight w_k^i that allows for *unbiased* estimation of rare event statistics of interest.

The procedure is described in Algorithm 1.

One should note that the DMC algorithm applies a selection procedure before the first simulation step. In practice, if

Algorithm 1. Diffusion Monte Carlo (DMC) algorithm.

Input: N walkers $(X_0^n)_{1\le n\le N}$ starting from an initial condition.
Initialize: Choose a sequence of resampling times $0 = t_0 < t_1 < \cdots < t_K$ and a family of splitting functions $(V_k)_k$ depending on a score function θ.
for $k = 0$ to K **do**
  (i) Reweighting: For each walker i
  if $k = 0$ **then**
    Define initial weights: $w_0^i = \exp(V_0(X_0^i))$
  else
    Define weights:
  $w_k^i = \bar{w}_{k-1} \exp\left(V_k(X_k^i) - V_{k-1}(\hat{X}_{k-1}^i)\right)$
  end if
  Compute average weight: $\bar{w}_k = (1/N)\sum_{i=1}^N w_k^i$
  (ii) Resampling: Create an updated ensemble of walkers $(\hat{X}_k^i)_{1\le i\le N}$ by sampling N_k^i copies of each X_k^i such that $\sum_i N_k^i = N$ and $\mathbb{E}[N_k^i] = (w_k^i/\bar{w}_k)$.
  (iii) Simulation: Integrate the model from t_k to t_{k+1}: $\hat{X}_k^i \to X_{k+1}^i$.
end for
Note: Walkers for which $N_k^i = 0$ are terminated and replaced by copies of walkers with $N_k^j \ge 2$.

one does not want to apply this first selection step, it can be avoided by choosing $V_0 = 0$. We do this in this Letter.

We perform the resampling step (ii) using the *pivotal sampling* scheme [59]. For a stochastic model, duplicated walkers will separate naturally after resampling. For the deterministic chaotic system that we work with, it is necessary to perturb the model immediately after each resampling step. Following the approach of [13], we apply perturbations to the spherical harmonics of the logarithm of the surface pressure field. In our implementation, we use a perturbation amplitude of 3×10^{-3}.

For any function ψ of the state (or history) of the system, the DMC algorithm yields the following unbiased estimator:

$$\mathbb{E}[\psi(X_k)] \approx \frac{\bar{w}_{k-1}}{N} \sum_{i=1}^{N} \psi(X_k^i) e^{-V_{k-1}(\hat{X}_{k-1}^i)}, \tag{B1}$$

where the expectation on the left is with respect to the distribution of the original process without RES. Using this formula with the choice $\psi = \mathbb{1}_{A\ge a}$ for a large $a \in \mathbb{R}$ and an observable A of the trajectory of X, one then has access to the probability of A taking extreme values ($\mathbb{E}[\mathbb{1}_{A\ge a}] = \mathbb{P}[A \ge a]$). In our case, this observable is $A_L(t_f)$ defined as the L-day average of the 2m temperature over the region $\mathcal{R}$ in Eq. (A1).

Choice of the AI emulator-based score function—The AI emulator that we use in AI + RES is trained to predict the state of the simulated physical system (the PlaSim GCM) a short time Δt in the future given the state at time t and then cycled autoregressively for longer predictions. As shown in Fig. 1(b), at each resampling time and for each walker i, we perform an M-member ensemble forecast using the AI emulator from the final state of the walker until the target period $[t_f, t_f + L]$. This yields an ensemble of M forecasts $\hat{A}_L^{i,j}(t_f|X_k^i)_{j=1}^M$ of the observable $A_L^i(t_f)$ [see Eq. (A1)] starting from the initial condition X_k^i. To the extent that the AI emulator faithfully reproduces the GCM's dynamics, the prediction furnished by the AI emulator ensemble is ideally suited for use in constructing the score function. Our results show that the AI emulator's weather forecast skill is sufficient to provide a very effective score function for AI + RES. More details on our AI emulator implementation are given in SM Sec. S3 [46].

We combine these ensemble forecasts into the score function by taking the ensemble mean of the forecast observable $\hat{A}_L^{i,j}(t_f|X_k^i)$:

$$\theta(X_k^i) = \frac{1}{M} \sum_{j=1}^{M} \hat{A}_L^{i,j}(t_f|X_k^i). \tag{C1}$$

We set the number of ensemble members per forecast to $M = 100$. Since our score function is defined using the ensemble mean, we do not expect substantial performance gains from increasing M. This parameter could, however, play a more significant role under alternative formulations of the score function.

We note that $A_L^i(t_f)$ could be predicted by other methods. We choose to leverage the AI emulator because of its proven forecasting skill and numerical efficiency, but other methods could be used. In particular, it would be interesting to test simpler methods based on directly learning the observable $A_L(t_f)$ (or its distribution) from an initial condition, such as in [47,60–64]. As a baseline, we also test the best possible forecast system, using ensembles of the PlaSim model itself to generate the score function as described in the *Ground truth and baselines* section.

TABLE I. Parameters used in the rare-event algorithm (Algorithm 1). N is the number of walkers, M is the forecast ensemble size, K is the number of resampling steps, τ is the resampling interval, C_k is the splitting constants, and L is the target duration. Simulations start on July 2 and run until $t_f + L$ with $t_f =$ August 1. The C_k were tuned empirically, and steps with $C_k = 0$ ensure that the walkers are well separated in phase space when selection begins.

Params	N	M	K	τ	C_k	L
Values	400	100	6	5 days	(0, 0, 0, 1.6, 1.8, 2.0)	7 days

We note that one could use a different statistic calculated from the ensemble forecast as the score function. The ideal score function to estimate the probability of achieving the rare event $A_L(t_f) > a$ for large a should anticipate walker paths leading to the rare event. An alternative score function appropriate for this goal is the conditional probability function $\theta(x) = P(A_L(t_f) > a|X_{t_k} = x)$, which is referred to as the *committor function* [65]. This score function has been proven to be optimal for a splitting algorithm similar to quantile DMC [66]. However, in our case there is no clear threshold, a, since we aim to characterize the entire tail of the distribution of $A_L(t_f)$. Our choice of score function, namely, the mean of the ensemble forecasts, favors the duplication of walkers for which the conditional expectation $\mathbb{E}[A_L(t_f)|X_{t_k} = x]$ is large. This approach is more conservative than selecting a fixed threshold *a priori*.

Choice of RES parameters—The RES hyperparameters used in all experiments are summarized in Table I.