

Response to reviewers of "Towards accurate extreme event likelihoods from diffusion model climate emulators" (egusphere-2026-2610)

We thank both reviewers for their careful and constructive comments, which have substantially improved the manuscript. Below we respond point by point, quoting each comment (RC1, RC2) followed by our response.

Note: During the revision we identified an opportunity to improve the core model that our importance sampling results rest upon. Specifically, we found that using exponential-moving-averages (EMA) of the model checkpoints greatly improves the TC climatology of the core model, and have modified quantitative results and figures in this revision using this new version of the model. This has allowed it to essentially close the gap between the TC frequencies of the model's climatology (control in Fig. 4) and ERA5. We confirmed via the cBottle paper's own AMIP climatology evaluation (spectra, seasonal cycle, precipitation PDF, variance maps, TC frequency) that the EMA weights are climatologically as good as or better than the net weights (this evaluation can be found under this [link](#), and these weights have been published). While this change does not modify our core conclusions, the end impact is more clear when using a better base model.

Reviewer 1

RC1.1 — Using diffusion models for computing sample probabilities is a very good idea... The presentation of the method and results is very good... The paper is honest about the computational cost... Computing odds ratio for guided vs unguided samples is a neat idea...

Thank you very much for the encouraging assessment.

RC1.2 — The weakness of the method is that it is a bit weird to only apply guidance for a given noise range (15, 20)... I would have liked to see how this affects the importance sampling estimate. Also... using a classifier trained on clean data produces low-quality gradients... but it does not seem to be the case here since the classifier is the same model as the denoiser (if I understood correctly paragraph 2.2), so it does not have this limitation and should work at all noise levels.

The classifier is in fact a separate checkpoint from the denoiser but the same architecture (clarified in §2.2); it is however noise-conditioned and trained jointly with a BCE head over the same noise distribution, so it does not suffer the clean-data-classifier gradient problem. We ran the requested experiment directly (Appendix, "Is the guidance window necessary for an accurate estimate?"): full-trajectory guidance, even at a strength matched to reproduce the [15,20]-window's mean classifier probability, gives $\log o$ spanning hundreds of nats, and an estimate that disagrees by orders of magnitude with both the windowed estimate and the control. Around 98% of this variance comes from the reverse-path terms, not from the guidance term itself. We conclude the window is required for accuracy, not only numerical stability.

RC1.3 — The main weakness of the paper is for me the absence of an analysis of the estimated uncertainty depending on the computational cost... I would like to see to what extent the 11x slowdown for ODE integration is needed.

Added (Appendix, "Cost-normalized estimator uncertainty"): a 19-setting ablation (base steps 18/36, extra steps in-window 0/5/25, Hutchinson probes $k=1/3/6$, plus one Heun point) on 500 rescored samples, using measured H100 GPU-seconds. Extra steps in the window are required to avoid biasing the estimate; increasing k helps up to $k = 3$ with diminishing returns beyond; an 18-step base grid biases p even when the mean log odds ratio looks similar to the 36-step grid. We did not find a cheaper setting that preserves accuracy at the paper's working point.

RC1.4 — it would have been nice to compute probabilities not clustered by detection threshold from the classifier, but rather by TC strength (which is probably correlated, but would have given a more interpretable result)... We expect that the uncertainty of the importance sampling estimate should decrease compared to the Monte Carlo estimate for more extreme events, which we begin to see on figure 4a) for 85% detection threshold but could be clearer.

We agree that ultimately we are interested in likelihoods for extremes measured in physical units. We did not guide directly on such variables, but on TC detection probability, since a TC is a complex 'semantic' feature spanning typical extreme values and spatial patterns across multiple variables — the classifier is the only summary metric we have available to guide on.

Added (Appendix, "Physical TC intensity"): the same exceedance-frequency comparison stratified by HPX3 maximum 10 m wind speed instead of classifier probability, plus a classifier-probability-vs-wind/pressure scatter. Importance sampling sits above control at moderate wind thresholds (12–18 m/s), because the guided proposal targets classifier probability rather than wind speed directly, and matches control again at 20–22 m/s. Overall, it is a noisier result than on the guidance variable itself, as is to be expected. We make this limitation clear in results section 3.3 and in the conclusions.

The reason we do not show even higher thresholds which would make the IS advantage clearer is that sample counts in the ERA5 and control sets become very low, already only 6 samples at the highest threshold. If we wanted to include higher thresholds, we would need additional ERA5 and control samples, but if we want to use the same SST boundary conditions also more

guided/IS samples, which would be computationally expensive. A full guided run for Fig. 4 is a relatively costly inference at 650 GPU-h.

RC1.5 — Given the importance of the TC classifier, it would have been nice to put a bit more details on it in the paper (how it was trained, its accuracy as function of noise, scale of its gradients..)

Added a new Appendix subsection, "Classifier skill and calibration versus noise level": architecture and training provenance (separate `cBottle-3d-tc` checkpoint, HPX3 classification head, IBTrACS-derived labels, BCE loss weight 0.1, trained over the same noise distribution as the denoiser), plus tropical-band AUC, as a function of σ on held-out September 2018 ERA5.

RC1.6 — In Figure 4a) we can see that for TC detection threshold = 85%, the guided model compensates for the under-representation of extremes in cBottle but barely increases upon the frequency in ERA5. If we increase guidance strength and guidance noise interval we would hope to be able to sample those extreme events a lot more. Is that all we can do?

Switching to EMA weights, reported throughout this revision, narrows (though does not eliminate) the ERA5 bias.

RC1.7 — First the writing suggests that the negative log likelihood bound is the formalization, whereas it's rather the decomposition (8) that is informative. Second, (8) is an average on the distribution, which does not tell you much about individual sample probabilities... It could turn out that this quantity is upper bounded by the KL under some conditions.

Agreed, and your guess is correct. We replaced the previous justification with a sample-wise bound,

$$E_{p_{data}} |\log p_{data} - \log p_{\theta}| \leq D_{KL} + \sqrt{2D_{KL}},$$

which only requires finite KL divergence and directly bounds the expected absolute log-probability error per sample.

RC1.8 — The discussion on adaptive guidance scale could probably be put after the results or in related work to streamline presentation of the method.

Moved to a new Appendix subsection ("Adaptive guidance scale"); the main text now states only the constant- γ choice used in this work.

RC1.9 — Typos: "research has has focused" (p.1); "with the the" (p.10)

Both fixed.

Reviewer 2

RC2.1 — ...the manuscript explains the workflow very clearly. A key positive result is that the reweighted guided samples reproduce the TC frequency estimated independently from unguided cbottle simulations... For the rarest threshold considered, the importance sampling estimate also begins to show lower relative sampling uncertainty than ordinary MC...

Thank you very much for the positive assessment.

RC2.2 — It would strengthen the climate interpretation considerably to evaluate the generated storms using an independent tracker or physical measures such as maximum wind speed or minimum pressure, rather than treating increasing classifier confidence as equivalent to increasing TC severity.

As for RC1.4 above, we added a physical-intensity analysis (wind speed, MSLP) in the Appendix. We note that since the TCs are many-variable features, where an individual feature like maximum wind is only correlated with TC classifier probability that was used to guide, we expect a higher variance in the estimates using the physical thresholds.

In addition, specifically for the independent-tracker request: we ran `TempestExtremes DetectNodes` (independent of the classifier, no track-stitching) on the same samples. Of classifier-flagged samples, 92.6% (guided), 89.1% (control), and 98.5% (ERA5) also register an independent `TempestExtremes` hit at $p > 0.5$, rising to 100% for all three at $p > 0.98$. We note this is a resolution-limited proxy at our ~ 100 km native resolution, not a best-track validation.

RC2.3 — It would therefore be useful to show how the estimated event probability and uncertainty change with ODE resolution, divergence estimator accuracy, and the chosen guidance interval... A cost normalized comparison... would also be more informative than comparing estimators at equal sample count...

This is a good question and was also raised by the other reviewer. We performed detailed analysis and ablations to quantify the impact of number of these settings. Please see the two new Appendix subsections ("Estimator uncertainty and computational cost" and "Is the guidance window necessary for an accurate estimate?"). On the 33x cost figure specifically: the sweep confirms this is largely necessary—cheaper settings bias the estimate rather than merely adding noise.

RC2.4 — The Antarctic likelihood experiment is less convincing than the tc analysis... I would therefore regard this portion as an exploratory demonstration of what diffusion likelihoods might eventually enable rather than evidence that the present method is suitable for event attribution.

We agree with this framing and have adopted it. The subsection states explicitly before presenting results that log-likelihood does not track temperature-anomaly extremeness in this application, and the unphysical Antarctic teacher-model response, previously only discussed alongside the appendix figure, is now flagged in the main text at the point the SST-perturbation result is introduced.

Added (sec. 3.4): “We flag upfront that the result of this section is largely exploratory and negative: as we show below, the diffusion-model log-likelihood does not, in this application, track the extremeness of the sampled temperature anomaly. The weak-but-correct-sign SST sensitivity in panel d) below is still informative about what these likelihoods can and cannot resolve, but does not demonstrate a working extreme-event-attribution tool.”

RC2.5 — Equation 17, the text refers to as the standard error of the importance sampling estimate, while the expression appears to give the variance of the individual weighted samples rather than the variance of their mean.

Correct, the plotted error bars were already computed correctly in code; only the printed equation was wrong (a factor of K , and it additionally referenced the wrong, unguided variable).

Fixed, and we now state explicitly that this is a standard error of the mean, while the ERA5 comparison uses a block bootstrap (4-day blocks, 1000 replicates).

RC2.6 — The manuscript defines the ratio as unguided p divided by guided p , so successful guidance should usually make the log ratio negative. However, section 3.1 says a positive log odds ratio is expected, which seems inconsistent.

Correct, Sections 3.1 and 3.2 were inconsistent; 3.2's convention is the one matching the code. We fixed Section 3.1 and added one explicit sign-convention statement after the definition of $o(x)$: guidance $\Rightarrow \log o(x) < 0$, with $\Delta_{latent} < 0$, $\Delta_{denoiser} < 0$, and the plotted guidance-divergence term equal to $-\Delta_{guidance}$ (positive as plotted). This is now consistent throughout, including the panel-by-panel discussion in Section 3.3.