nature portfolio

# Peer Review File

# Extreme Event Aware (η-) Learning

Corresponding Author: Professor Themistoklis Sapsis

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors present an innovative and mathematically elegant optimization framework designed to predict statistically consistent extreme events in zero-extreme-data regimes. This work addresses a critical bottleneck in computational physics, climate science, and engineering. Furthermore, the proposed integration of optimal transport with deep generative models constitutes a practical contribution to the field of rare event modeling. However, before the manuscript can be recommended for publication, several significant theoretical fragilities and limitations concerning spatial degradation and prior misspecification should be addressed. Therefore, I recommend acceptance subject to major revisions, as detailed in the report below.

1) Manuscript Summary
The authors introduce Extreme Event Aware (η-) Learning, a machine learning optimization framework designed to predict statistically consistent and physically plausible rare events, even when the training dataset completely lacks extreme examples. The framework incorporates a priori statistical information about an extreme observable into the training process using Optimal Transport. Specifically, the authors rigorously quantify the conditions under which standard Empirical Risk Minimization (ERM) fails in data-scarce regimes, and they resolve this by augmenting the loss function with a 1-Wasserstein ($W_1$) distance penalty. The method demonstrates significant performance in prototype problems and real-world precipitation downscaling tasks.

2) Research Relevance
This work is highly relevant to computational physics, climate science, and engineering, where simulating extreme events is computationally prohibitive and real-world extreme data is naturally scarce. Compared to established literature, such as active learning or large deviation theory, which typically require partial access to governing equations or a small seed of extreme samples, η-learning operates in the zero-extreme-data regime. Relying on a predefined reference distribution ($v_0$) to guide the model is an innovative approach that constitutes a strong contribution to rare event modeling.

3) Methodology
Overall, the methodology is compelling, and the ERM failure quantification (Theorem 4) rigorously formalizes an intuitive problem. However, the manuscript should be revised to address theoretical fragilities and formulation constraints:

i) The Problem Statement relies heavily on advanced measure theory and topological terminology (e.g., σ-algebras, Polish spaces) without accompanying physical intuition. To better approach applied scientists who would benefit most from this framework, the authors could include a plain-English, high-level summary of the physical pipeline (Input → Complex State → 1D Observable) at the beginning of this section.

ii) The Problem Statement introduces two restrictive assumptions. First, it assumes independent and identically distributed (IID) samples, discarding the temporal trajectories and strong correlations that characterize complex dynamical systems (e.g., climate dynamics). Second, it defines extremes via a static scalar threshold, excluding compound or duration-based extremes. The authors should explicitly acknowledge these limitations and discuss whether the framework could be adapted for non-IID data or alternative extreme definitions.

iii) The proof for Theorem 4 relies on bounding parameters $\tilde{C}$ and $\hat{C}$, which the authors admit lie "beyond the scope of this work" to compute. To strengthen the practical applicability of this definition, the authors should provide at least one concrete analytical example of a simple function class where these constants can be explicitly bounded.

iv) The theoretical formulation assumes a strictly deterministic dataset and estimator. The authors should briefly discuss how the theoretical lower bounds might shift if the inherent stochastic variance of training modern deep learning models (e.g., random weight initialization, batching) is properly accounted for.

v) In the Supplementary Material, the authors acknowledge that gradients from the ERM loss and the $W_1$ penalty frequently point in opposite directions. Because this contradictory signaling can lead to severe training divergence, this gradient conflict should be explicitly discussed in the main text, and a discussion on how the balancing hyperparameter $\lambda$ mitigates it would be helpful.

vi) Theorem 6 relies on Assumption 5, which requires the model to make systematically biased errors without positive and negative errors canceling out. As this is hard to constrain during training, the authors should clarify how its failure impacts the point-wise reliability of the $\eta$-map.

vii) In the Optimality section (Assumptions 8a and 8b), the bounding parameters $\varepsilon_1$ and $\varepsilon_2$ are introduced to formalize L1 deviations in the non-extreme set. A brief physical intuition of these parameters should be provided to improve readability.

3) Results and Discussion
While the empirical results are compelling for the five distinct problems shown, some aspects regarding spatial structure and prior misspecification should be clarified:

i) The toy models inadvertently expose the fragility of using a scalar observable $g(u)$. In the 2D-to-1D problem, the physical location of the extreme shifts across initializations (SI Figure 2); a metric (e.g., ensemble spatial variance) should quantify this. Furthermore, in the 2D-to-2D problem, the scalar penalty causes notable spatial distortion via interference between $u_1$ and $u_2$ (Figure 2). This interference effect should be discussed as a limitation of the 1D Wasserstein penalty.

ii) For the first precipitation downscaling challenge, the degradation of the weighted coverage statistic at lower thresholds (SI Figures 8 and 9) indicates that penalizing the spatial maximum explicitly costs bulk spatial coherence. The authors should report standard spatial similarity metrics (e.g., SSIM or full-grid RMSE) alongside the extreme metrics to quantify how much spatial structure is sacrificed.

iii) The case study using a hypothesized GEVD prior (SI Figure 12) highlights that the model will faithfully hallucinate extremes to match a potentially flawed distribution. A quantitative sensitivity analysis (ablation study) showing how drastically physical fidelity degrades when $v_0$ is misspecified by a known margin would be valuable.

iv) The algorithm requires dynamic approximation of the 1-Wasserstein distance, yet a discussion on computational complexity is missing. The authors should include a brief quantitative analysis to set realistic expectations, particularly concerning the overhead of adding this to already expensive generative pipelines like Flow Matching.

v) While the Discussion highlights the framework's versatility, it lacks critical evaluation. The authors must add a "Limitations and Future Work" subsection summarizing the major caveats identified above. Acknowledging these will strengthen the manuscript and guide future research toward multi-dimensional observables.

vi) The empirical reproducibility of the work is overall high. The authors provide thorough details regarding the ERA5-Land dataset processing, hyperparameter choices, and algorithmic training loops.

(Remarks on code availability)
First, it should be noted that I reviewed the provided code repository without installing or running the scripts. The repository maps to the experiments in the manuscript. The authors provide separate notebooks for the different case studies (the 2D toy problems, ERA5-Land downscaling, and the generative model integration). While the code for the mathematical implementation is there, it would be very helpful if the authors added inline comments linking the core definitions directly to the numbered equations in the text. This would make it much easier for readers to follow the translation from theory to code.

Regarding usability, the README needs to be updated with proper environment setup instructions, ideally through a requirements.txt file. Please ensure that the exact versions of the core dependencies are pinned so the codebase does not break due to future library updates. Finally, a brief note on the expected hardware requirements (such as required GPU VRAM and approximate processing times) should be added. This is useful for anyone looking to reproduce these results.

Finally, for the real-world precipitation challenge, the code's usability relies on the availability of the ERA5-Land dataset. The repository could include a minimal, pre-processed subset of the data, or an automated script that downloads a small sample, so users can easily test the pipeline without needing to download and format massive climate datasets.

Reviewer #2

(Remarks to the Author)
In this paper, an Extreme Event Aware (e2a or eta) or $\eta$-learning algorithm is developed. It does not assume the existence of extreme events in the available data. It reduces uncertainty even in ‘uncharted’ extreme-event regions by enforcing the extreme-event statistics of an observable indicative of extremeness during training, which can be supported by qualitative arguments or estimated from unlabeled data.

Overall, the method is novel, and the paper is well-written. The topic is also a good fit for the scope of Nature Communications. That said, a few points are not very clear, and I believe they need to be explained and/or validated more systematically. I suggest a revision of this paper.

It looks like g is a very low-dimensional output function. In the paper, it says g belongs to R, which is one dimension. However, in practice, the observable may not always be a scalar. It can be an extreme pattern or a combination of several features. What is the scalability of the method?

The use of the Wasserstein Distance needs to be explained more systematically. While it's true that it serves as a useful statistical regularization, the advantages and disadvantages should be discussed. What will be the conditions that make this regularization particularly suitable for extreme events? What about K-L divergence and other statistical metrics? Even qualitative comparisons or discussions can be very helpful in supporting the use of the method proposed here.

Following the above question, the Wasserstein Distance is a statistical regularization, which is also the emphasis of this paper. However, how does the method, with such a statistical regularization, help learn the dynamics of extreme events? In principle, with only statistical information, the mechanisms of the extreme events generated can be very different from the truth. This is certainly not the case with the numerical results. Therefore, something behind the scenes must play some roles here. Adding more reasoning and even potential caveats will be very necessary.

Is there a way to automatically choose the hyperparameters to make the algorithm more objective? Otherwise, some additional prior knowledge is needed, which can sometimes complicate the problem.

One thing I feel very confused about is the toy model examples. I do see you get extreme events. However, the locations and amplitudes differ from the truth, e.g., in Fig 1. While this can still give you the right spatiotemporal statistics, I wonder what the target is here. I guess, in addition to sampling extreme events across the domain (in SI Fig 2), what's more important is to get the point-wise statistics, not the spatiotemporal average. It is also practically useful to predict the locations and times of extreme events. This relates to my previous question about the dynamical mechanisms that generate extreme events.

I also noticed that, for example, in Figure 6 and maybe Figure 5 as well, the estimated tail is above the true tail. Is this because of the over-regularization? Is there a general criterion to mitigate this, or does it need a few trials to get the optimal solution?

Did you consider the case with observational noise? Can you filter the noise at the same time, so you won't treat the noisy observation as the truth when applying your method?

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have thoroughly addressed all the concerns raised, so I recommend the manuscript for acceptance in its current form.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have addressed all the questions I asked. In particular, the authors provided a very comprehensive response letter. I do not have any further concerns. I also want to remark that I co-reviewed this paper with an early-career scientist. Thank you for this great work.

(Remarks on code availability)
The authors have improved the code, which is now clear.

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

# Point-to-Point Responses to Reviewers*

Kai Chang and Themistoklis P. Sapsis

## General Edits

In addition to the revisions made in direct response to the reviewers' comments, we have made several additional edits that we believe improve the clarity, organization, and readability of the manuscript. We summarize these changes here for completeness and will list the specific edits below.

1. We changed the theorem numbering in the SI from, for example, Theorem 1 to Theorem S1 to improve readability.
2. We changed the paragraph headings "Description of the results" in the Results section to "Analysis of the results" to more accurately reflect that these paragraphs interpret the numerical results rather than merely describe them.
3. The second data-consistency condition (Definition 3, eq. (9)) is relaxed from $\hat{C} < 1$ to $\hat{C} < \infty$:

$$\hat{C} := \inf_{C \geq 0} \left\{ \int_{\mathcal{X} \setminus E} |y - y_\xi| \, \mathrm{d}\mu \leq C \cdot \int_E |y - y_\xi| \, \mathrm{d}\mu \right\} < \infty.$$

   The proof strategies for the theorems remain the same.

## Responses to Reviewer #1

**General assessment.** *The authors present an innovative and mathematically elegant optimization framework designed to predict statistically consistent extreme events in zero-extreme-data regimes. This work addresses a critical bottleneck in computational physics, climate science, and engineering. Furthermore, the proposed integration of optimal transport with deep generative models constitutes a practical contribution to the field of rare event modeling. However, before the manuscript can be recommended for publication, several significant theoretical fragilities and limitations concerning spatial degradation and prior misspecification should be addressed. Therefore, I recommend acceptance subject to major revisions, as detailed in the report below.*

**Response.** We thank the reviewer for the positive assessment of the proposed framework and for identifying important points where the scope and limitations should be clarified.

*The reviewers' original comments are copied and pasted in italicized texts, with an identifier provided before each quote. Edits to the manuscript and SI are shown in this color.

**Manuscript summary.** *The authors introduce Extreme Event Aware ($\eta$-) Learning, a machine learning optimization framework designed to predict statistically consistent and physically plausible rare events, even when the training dataset completely lacks extreme examples. The framework incorporates a priori statistical information about an extreme observable into the training process using Optimal Transport. Specifically, the authors rigorously quantify the conditions under which standard Empirical Risk Minimization (ERM) fails in data-scarce regimes, and they resolve this by augmenting the loss function with a 1-Wasserstein ($W_1$) distance penalty. The method demonstrates significant performance in prototype problems and real-world precipitation downscaling tasks.*

**Response.** We appreciate this accurate summary.

**Research relevance.** *This work is highly relevant to computational physics, climate science, and engineering, where simulating extreme events is computationally prohibitive and real-world extreme data is naturally scarce. Compared to established literature, such as active learning or large deviation theory, which typically require partial access to governing equations or a small seed of extreme samples, $\eta$-learning operates in the zero-extreme-data regime. Relying on a predefined reference distribution ($\nu_0$) to guide the model is an innovative approach that constitutes a strong contribution to rare event modeling.*

**Response.** We thank the reviewer for recognizing the intended contribution.

**Methodology comment #i.** *The Problem Statement relies heavily on advanced measure theory and topological terminology, e.g. $\sigma$-algebras and Polish spaces, without accompanying physical intuition. To better approach applied scientists who would benefit most from this framework, the authors could include a plain-English, high-level summary of the pipeline (Input $\to$ Complex State $\to$ 1D Observable) at the beginning of this section.*

**Response.** We agree. We have added a plain-language paragraph at the beginning of the Problem Statement section at line 134. To avoid repetitions, we have also shortened the original problem formulation. The revised opening of the section reads as follows.

At a high level, the setting considered here consists of three objects. First, an input $x$ describes the uncertain physical driver of a system, such as an initial condition, forcing parameter, or low-fidelity data such as low-resolution fields. Second, the system maps this input to a high-dimensional state $u(x)$, such as a dynamic trajectory or high-fidelity data. Third, an application-dependent observable $g(u(x))$ summarizes the state into a scalar quantity whose large values signify the extreme event of interest. Thus, the pipeline is

$$x \longmapsto u(x) \longmapsto g(u(x)).$$

The learning task is to approximate either the full state map $u$ or the composite response $g \circ u$, even when the available training data contain few or no extreme events.

To formalize this setting, let $X : (\Omega, \Sigma, P) \to (\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be an input random variable, where $\mathcal{X} \subset \mathbb{R}^d$ is equipped with its relative Borel $\sigma$-algebra. We denote the law of $X$ by $\mu$, namely

$$\mu(A) = P(X^{-1}(A)), \qquad A \in \mathcal{B}(\mathcal{X}).$$

We write $X \sim \mu$, and use $x \sim \mu$ to denote a realization of the random input. Thus, $\mu$ describes the distribution of inputs encountered by the system. We assume that $\mu$ is either simple to sample from, such as a Gaussian distribution, or sufficiently well represented by existing samples.

Define the state map

$$u : \mathcal{X} \to \mathcal{U} \subset \mathbb{R}^m, \qquad x \mapsto u(x),$$

where $\mathcal{U}$ is the state space. For each input realization $x$, $u(x)$ denotes the corresponding system state. In the settings of interest, accurate evaluations of $u$ are computationally expensive, and available observations of $u$ may be incomplete.

Next, define the pre-determined observable

$$g : \mathcal{U} \to \mathcal{Y} \subset \mathbb{R}, \qquad u \mapsto g(u).$$

This map provides a scalarization of the state space used in the subsequent distributional formulation. Examples include a spatial maximum, a spatial average, an energy functional, or another physically meaningful summary of the state. Throughout this work, we focus on scalar-valued observables; remarks on extensions to multivariate settings are provided under "Limitations and Future Work".

Finally, let

$$y := g \circ u, \qquad Y := y(X).$$

Assuming $y$ is $\mathcal{B}(\mathcal{X})$-measurable with respect to $\mathcal{B}(\mathcal{Y})$, the scalar response $Y$ is a random variable with law given by the push-forward measure $y_{\#}\mu$:

$$y_{\#}\mu(B) = \mu(y^{-1}(B)), \qquad B \in \mathcal{B}(\mathcal{Y}),$$

where $y^{-1}(B) = \{x \in \mathcal{X} : y(x) \in B\}$. We assume that $y_{\#}\mu$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, so that its probability density function exists.

**Methodology comment #ii.** *The Problem Statement introduces two restrictive assumptions. First, it assumes independent and identically distributed (IID) samples, discarding the temporal trajectories and strong correlations that characterize complex dynamical systems, e.g. climate dynamics. Second, it defines extremes via a static scalar threshold, excluding compound or duration-based extremes. The authors should explicitly acknowledge these limitations and discuss whether the framework could be adapted for non-IID data or alternative extreme definitions.*

**Response.** We thank the reviewer for pointing out these limitations. We agree that the IID assumption does not represent temporally correlated trajectory data. In the present theory, however, independence is used only to obtain an explicit expression for the probability that the training set does not intersect the extreme set $E$. We have clarified this point in the Problem Statement and added the corresponding extension to dependent data in the SI. Specifically, we revised the Problem Statement at line 194 as follows:

We study the data-driven setting where we have access to a dataset $\mathcal{D} = \{(x_i, u_i, y_i)\}_{i=1}^n$, where $u_i = u(x_i)$ and $y_i = g(u_i)$. [...] For the theoretical results below, the inputs $x_1, \ldots, x_n$ are assumed to be independent and identically distributed according to $\mu$. The extension to dependent data is provided in the Supplementary Information (SI) section "Theoretical Extensions".

We also added subsection "Extension to Non-IID Data" under the section "Theoretical Extensions" in the SI:

## Extension to Non-IID Data

The IID assumption in Theorem 4 of the main text is used only to evaluate the probability that the training set does not intersect the extreme set $E$. Let $X_1, \ldots, X_n$ be a possibly dependent sequence

with common marginal law $\mu$, and define

$$q_n(E) := \mathbb{P}\{X_i \notin E,\ i = 1, \dots, n\}.$$

For data collected along a temporal trajectory, this probability can equivalently be written as

$$q_n(E) = \mathbb{P}\{\tau_E > n\}, \qquad \tau_E := \inf\{i \geq 1 : X_i \in E\}.$$

Since $\mu(E) \leq \delta$, the union bound gives the dependence-independent estimate

$$q_n(E) \geq 1 - \sum_{i=1}^{n} \mathbb{P}\{X_i \in E\} \geq 1 - n\delta.$$

Thus, $q_n(E) \geq p$ whenever

$$n \leq \frac{1-p}{\delta}.$$

Sharper estimates of $q_n(E)$ may be obtained under additional assumptions on the temporal process. For instance, when $(X_i)$ is a stationary Markov chain with invariant law $\mu$, $q_n(E) = \mathbb{P}_\mu(\tau_E > n)$ can be analyzed as the survival probability of the chain killed upon entering $E$, with mixing-time, spectral-gap, coupling, or hitting-large-set estimates yielding bounds in terms of $\mu(E)$ and the relaxation scale of the chain. For trajectories generated by small random perturbations of deterministic dynamics, related hitting- and exit-time estimates may also be derived from the Freidlin-Wentzell theory of randomly perturbed dynamical systems [4, 5].

On the event that the training set does not intersect $E$, the remainder of the proof of Theorem 4 is unchanged. Consequently, its lower bound holds with probability at least $q_n(E)$.

We also agree that certain applications may require more than a scalar observable. It is adopted because the one-dimensional quantile representation of $W_1$ underlies both the efficient implementation and the tail-error optimality results. The framework can nevertheless be extended to multivariate observables in a computationally scalable manner. There are three distinct settings to consider.

If a vector of features $G = (g_1, \dots, g_r)$ admits a physically meaningful scalar severity score $s : \mathbb{R}^r \to \mathbb{R}$, one can take $g = s \circ G$ and apply the present scalar framework directly. This includes pattern-based events whenever a suitable scalar score is available.

When scalarization is not desired but calibration of the componentwise marginal laws is sufficient, a direct extension is

$$\sum_{j=1}^{r} W_1((g_j \circ \phi_\theta)_\# \mu, \nu_{0,j}),$$

where $\nu_{0,j}$ denotes the reference law for the $j$-th observable. A detailed complexity analysis for the single-obsevable case is provided in newly added SI subsection Computational Complexity (also see the response to the relevant comment below). We briefly restate the analysis here for assistance. Let $n_{\tilde{x}}$ be the number of auxiliary inputs used to estimate the push-forward distributions, $n_q$ the number of quantile levels per observable, and $C_\mathrm{f}$ and $C_\mathrm{b}$ the costs of one forward and backward model propagation, respectively. At each quantile-index refresh, the model is evaluated once on all $n_{\tilde{x}}$ auxiliary inputs, and these outputs are shared across the $r$ observables, giving a cost of $O(n_{\tilde{x}} C_\mathrm{f})$. Locating the quantile samples requires sorting $n_{\tilde{x}}$ values for each observable, with total cost $O(r n_{\tilde{x}} \log n_{\tilde{x}})$. The subsequent gradient computation uses at most $r n_q$ selected quantile samples and therefore costs $O(r n_q (C_\mathrm{f} + C_\mathrm{b}))$. Thus, under the worst-case assumption that the quantile

indices are refreshed at every iteration, the componentwise Wasserstein regularization has total cost

$$O(n_{\tilde{x}} C_{\mathrm{f}} + r n_{\tilde{x}} \log n_{\tilde{x}} + r n_q (C_{\mathrm{f}} + C_{\mathrm{b}})) .$$

The extension therefore scales linearly with the number of observables $r$, while the full inference pass over the auxiliary inputs is shared across all observables.

Third, if the joint dependence or correlation structure of $G$ is also important, then marginal matching is insufficient and a genuinely multivariate discrepancy is needed. Tools such as sliced Wasserstein distance and entropically regularized optimal transport can be integrated into the framework in place of the scalar $W_1$ term. Both have been widely adopted in machine learning as computationally tractable approaches to multivariate distribution comparison: sliced Wasserstein reduces the problem to one-dimensional projections, while entropic regularization enables efficient and parallelizable Sinkhorn iterations. We expect tail-enhanced versions of these discrepancies to be important when the goal is to correct tail bias in a multivariate extreme-event setting.

To clarify the scope of the formulation, we added the following sentence in the Problem Statement at line 168:

Throughout this work, we focus on scalar-valued observables; remarks on extensions to multivariate settings are provided under "Limitations and Future Work".

We also added the following discussion to the newly added "Limitations and Future Work" subsection at line 1067:

Finally, the present theory and algorithms are developed for scalar observables, for which $W_1$ enables efficient computation and underpins the tail optimality results. For a vector observable $G = (g_1, \ldots, g_r)$, one may either introduce a physically meaningful severity function $s : \mathbb{R}^r \to \mathbb{R}$ and apply the present framework to $g = s \circ G$, or use a sum of componentwise $W_1$ penalties when matching the marginal laws is sufficient. The latter scales linearly with $r$, but does not constrain the correlation structure among the components. When the joint law of $G$ is essential, genuinely multivariate discrepancies based on, for example, sliced Wasserstein distance or entropically regularized optimal transport are required [1, 2]. Extending the present tail-focused theory and scalable algorithms to these settings is another important direction.

**Methodology comment #iii.** *The proof for Theorem 4 relies on bounding parameters $\widetilde{C}$ and $\widehat{C}$, which the authors admit lie "beyond the scope of this work" to compute. To strengthen the practical applicability of this definition, the authors should provide at least one concrete analytical example of a simple function class where these constants can be explicitly bounded.*

**Response.** We thank the reviewer for this comment. We agree that an explicit analytical example would make the data-consistency condition more transparent and practically useful. We have added to the SI an extra section An Analytic Example for Data-Consistency, and have referenced to it in the main text at line 955:

While it is beyond the scope of this work to characterize these constants for general approximation classes, they can be computed explicitly in simple settings. We provide such an example in the SI.

The added section reads as follows.

## An Analytic Example for Data-Consistency

To strengthen the practical applicability of the data-consistent condition introduced in Definition 3 of the main text, we provide a simple, interpretable example where the constants $\widetilde{C}$ and $\widehat{C}$ can be explicitly bounded and the bound directly depends on the smoothness of the true composite map, the approximation class, and the extreme set, as introduced in the condition.

Let $\mathcal{X} = [0,1]$, let $\mu$ be the uniform measure on $[0,1]$ and $E = [1-\delta, 1]$. Consider the affine approximation class

$$\mathcal{F} = \{x \mapsto ax + b : \ a, b \in \mathbb{R}\}$$

and the true map

$$y(x) = a_\star x + b_\star + M\psi_r(x),$$

where $a_\star, b_\star \in \mathbb{R}$, $M \gg 1$, and

$$\psi_r(x) = \begin{cases} 0, & x < 1-\delta, \\ \left(\dfrac{x-(1-\delta)}{\delta}\right)^r, & x \in [1-\delta, 1], \end{cases} \qquad r \geq 2.$$

Here, $M$ controls the extremeness of the unresolved extreme feature while the exponent $r$ controls its smoothness. This construction gives $\psi_r \in C^{r-1}$ at $1-\delta$, while the $r$-th derivative has a jump. In particular, as $r$ increases, the bump becomes flatter near $1-\delta$, and thus smoother.

Assume now that the training data do not intersect $E$, and that there are at least 2 distinct training data points. Then, on the observed region $\mathcal{X} \setminus E = [0, 1-\delta)$, empirical-risk minimization over $\mathcal{F}$ recovers

$$y_\xi(x) = a_\star x + b_\star,$$

in a noise-free setting. Hence,

$$y(x) - y_\xi(x) = 0, \qquad x \in \mathcal{X} \setminus E,$$

while on $E$,

$$y(x) - y_\xi(x) = M\psi_r(x) \geq 0.$$

Therefore, a standard calculation gives

$$\int_E |y - y_\xi| \, d\mu = M \int_{1-\delta}^1 \left(\frac{x-(1-\delta)}{\delta}\right)^r dx = \frac{M\delta}{r+1}.$$

On the other hand,

$$\int_{\mathcal{X}\setminus E} |y - y_\xi| \, d\mu = 0, \qquad \left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \, d\mu \right| = 0.$$

Hence the smallest admissible constants in the data-consistency conditions are $\widetilde{C} = \widehat{C} = 0$.

We now consider a slightly perturbed version, representing the potential noise in the non-extreme region, where

$$\int_{\mathcal{X}\setminus E} |y - y_\xi| \, d\mu \leq \varepsilon_{\text{bulk}}.$$

By definition,

$$\widehat{C} := \inf \left\{ C \geq 0 : \int_{\mathcal{X}\setminus E} |y - y_\xi| \, d\mu \leq C \int_E |y - y_\xi| \, d\mu \right\}.$$

Note that for small $\delta$, by linearity of the non-extreme feature and the approximation class,

$$\int_E |y - y_\xi| \, d\mu = \frac{M\delta}{r+1} + O(\varepsilon_{\text{bulk}}\delta),$$

it follows that one admissible choice of $C$ is

$$C = \frac{\varepsilon_{\text{bulk}}}{\frac{M\delta}{r+1} + O(\varepsilon_{\text{bulk}}\delta)} = \frac{1}{\frac{M\delta}{(r+1)\varepsilon_{\text{bulk}}} + O(\delta)}.$$

When the noise in $\mathcal{X} \setminus E$ does not 'magically' recover the unobserved extreme, i.e. when $M/\varepsilon_{\text{bulk}}$ dominates, we may simply pick

$$C = \frac{\varepsilon_{\text{bulk}} \, (r+1)}{M\delta},$$

and therefore,

$$\widehat{C} \leq \frac{\varepsilon_{\text{bulk}}(r+1)}{M\delta}.$$

Similarly,

$$\widetilde{C} := \inf \left\{ C \geq 0 : \left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \, d\mu \right| \leq C \left| \int_E (y - y_\xi) \, d\mu \right| \right\}.$$

Using

$$\left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \, d\mu \right| \leq \int_{\mathcal{X}\setminus E} |y - y_\xi| \, d\mu \leq \varepsilon_{\text{bulk}},$$

we again obtain the upper bound

$$\widetilde{C} \leq \frac{\varepsilon_{\text{bulk}}(r+1)}{M\delta}.$$

Consequently, we see that $\widehat{C}, \widetilde{C} < 1$ when $M/\varepsilon_{\text{bulk}}$ dominates, i.e. when the error in the unobserved extreme set dominates that in the observed non-extreme region.

**Methodology comment #iv.** *The theoretical formulation assumes a strictly deterministic dataset and estimator. The authors should briefly discuss how the theoretical lower bounds might shift if the inherent stochastic variance of training modern deep learning models, e.g. random weight initialization and batching, is properly accounted for.*

**Response.** We agree that this point should be made explicit. The previously submitted SI already contained an extension of the theory to random estimators and random datasets in the section Theoretical Extensions, under the subsection Generalize to Random Estimator and Dataset. We have now further clarified this point in the main text at line 244 to read as follows.

First, both the dataset $\mathcal{D}$ and the estimator $y_\xi$ are treated as deterministic objects. This simplification is not an essential restriction of the argument. In the presence of training randomness of modern neural networks, the same argument applies conditioning on a realization of the randomness, while averaging over these sources gives the corresponding expected bound. Thus stochastic variance may change the realized constants or confidence level of the lower bound developed below, but it does not remove the missing-tail contribution caused by data scarcity.

We have also revised the concluding sentence of this assumptions paragraph at line 255 to explicitly direct readers to the relevant SI section:

These limitations and considerations are either relaxed or addressed in the SI section "Theoretical Extensions".

**Methodology comment #v.** *In the Supplementary Material, the authors acknowledge that gradients from the ERM loss and the $W_1$ penalty frequently point in opposite directions. Because this contradictory signaling can lead to severe training divergence, this gradient conflict should be explicitly discussed in the main text, and a discussion on how the balancing hyperparameter $\lambda$ mitigates it would be helpful.*

**Response.** We thank the reviewer for pointing this out. We agree that the gradient conflict between the supervised ERM term and the $W_1$ penalty should be discussed more explicitly in the main text. We have revised the paper and the SI to discuss this more extensively.

To understand how to resolve the gradient conflict, we first want to clarify the source of this issue. The central working premise of $\eta$-learning is that the ERM estimator captures the bulk input-output relation, but underestimates or omits the tail behavior due to limited data. The role of the $W_1$ penalty is to correct the learned map in regions that affect the observable tail.

This premise also explains why a gradient conflict can occur during optimization. The ERM loss is computed from the labeled pairs $(x_i, u_i)$, whereas the $W_1$ loss is computed from tail samples selected according to the current observable distribution $(g \circ \phi_\theta)_{\#}\mu$. More precisely, at each tail-sample-update step, auxiliary samples from $\mu$ are passed through the current $\eta$-map to form an empirical approximation of $(g\circ\phi_\theta)_{\#}\mu$, and the upper-tail samples are used to define the $W_1$ penalty. In the intended extreme-event-deficient regime, the supervised data, as well as the bulk samples well represented by the ERM estimator, are expected to lie primarily in the bulk of the observable distribution. However, if the combined ERM+$W_1$ objective is optimized before the ERM loss has been sufficiently minimized, the model-induced ordering of samples by their predicted observable values can be unreliable. As a result, samples that should be treated as bulk under the supervised input-output relation can be spuriously ranked among the upper-tail samples used to define the $W_1$ loss. When this happens, the two loss components will provide incompatible gradient signals: the ERM term anchors the learned map to the labeled input-output relation, while the $W_1$ term pulls spuriously selected bulk samples toward upper-tail quantiles of $\nu_0$, leading to oscillatory or unstable optimization.

Conversely, once the ERM loss has been sufficiently minimized, the model-induced ordering used by the $W_1$ term is anchored to a predictor that already respects the supervised input-output relation. Bulk samples are then assigned ERM-consistent observable values and are no longer systematically ranked among the upper-tail samples merely because of unreliable early predictions. Thus, the tail set used to define the $W_1$ penalty is no longer dominated by samples that should be treated as bulk under the ERM-consistent map. In this regime, the $W_1$ term does not systematically pull ERM-consistent bulk predictions toward extreme quantiles of $\nu_0$. Instead, it primarily adjusts the tail of the learned observable distribution, while the ERM term plays a mainly stabilizing role, anchoring the learned map near the low-ERM regime already reached.

Based on this reasoning, our main stabilization mechanism is ERM pre-training. Specifically, we first train the model using only the supervised ERM loss, and then initialize both the $\eta$-learning stage and the initial tail set used to approximate the $W_1$ penalty from this pre-trained estimator, as shown in Algorithm 1 in the SI. At this point, the predictor already captures the labeled input-output relation well and induces a reliable ranking of the auxiliary samples. Consequently, $W_1$

is estimated with tail candidates under an ERM-consistent observable distribution. The ERM and $W_1$ terms are therefore much less prone to the severe misalignment that occurs when the $W_1$ penalty is introduced before the supervised fit has been learned. During the subsequent $\eta$-learning iterations, the tail set is dynamically updated as the learned map evolves, keeping the samples used to estimate $W_1$ separated from the ERM-induced bulk. This procedure systematically mitigates the conflict caused by unreliable tail-sample selection and has empirically yielded consistently stable and robust training.

We next clarify the role of the balancing hyperparameter $\lambda$. The parameter $\lambda$ can mitigate residual instability by controlling the relative magnitude of the two gradient components in

$$\nabla_\theta L(\theta) = \nabla_\theta L_{\mathrm{ERM}}(\theta) + \lambda \nabla_\theta L_{W_1}(\theta).$$

One potentially beneficial choice is to select $\lambda$ so that the scaled $W_1$ gradient has a comparable Euclidean norm to the ERM gradient, that is

$$\lambda \|\nabla_\theta L_{W_1}(\theta)\|_2 = \|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2, \implies \lambda = \frac{\|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2}{\|\nabla_\theta L_{W_1}(\theta)\|_2 + \varepsilon},$$

with a small $\varepsilon > 0$. This balances the magnitudes of the two gradient components and prevents the $W_1$ correction from overwhelming the supervised anchor.

However, we emphasize that $\lambda$ does not by itself resolve the directional incompatibility between anti-aligned gradients. It rescales the $W_1$ contribution, but it does not change the fact that an unreliable tail-set selection can make the two losses prefer opposing parameter updates. For this reason, we view ERM pre-training and the associated tail-set update as the main mechanism for resolving the severe early-stage conflict, with $\lambda$ serving as a secondary balancing parameter during the subsequent $\eta$-learning stage.

This interpretation is also supported by the robustness results reported in SI Fig. 4. In that experiment, the learned tail remains close to $\nu_0$ as $\lambda$ varies from $10^{-4}$ to 10, a range spanning five orders of magnitude. This indicates that our algorithm does not require delicate tuning of $\lambda$; rather, $\lambda$ mainly controls the strength of the residual tail correction around an already well-anchored supervised solution.

We have expanded the subsection Effective Training Strategies to discuss this reasoning in detail:

### Effective Training Strategies

Accurately estimating tail quantiles at every iteration can be memory- and compute-intensive, because it naively requires evaluating the model on a large auxiliary sample set $\tilde{x} \sim \mu$ at each update. To mitigate this, we use an inference-informed continual training strategy: periodically, we run inference on the auxiliary set to identify the inputs whose current observable values realize the target quantiles in $\mathcal{Q}$, and then restrict subsequent $W_1$ estimations to these identified inputs until the next refresh. This preserves a reliable estimate of the tail loss while substantially reducing per-iteration cost. A detailed computational complexity analysis of the full procedure is provided in SI Subsection "Computational Complexity".

This tail-set selection procedure also creates a potential optimization issue. The $W_1$ term is evaluated on inputs selected according to the current model-induced observable distribution $(g \circ \phi_\theta)_{\#}\mu$. Early in training, before the supervised input-output relation has been learned, this model-induced ordering can be unreliable. As a result, samples that should correspond to bulk behavior under an ERM-consistent predictor can be spuriously selected as tail samples and pulled

toward upper quantiles of $\nu_0$, while the ERM loss pulls the model toward the labeled pairs. The two loss components therefore provide conflicting gradient signals, leading to oscillatory or unstable optimization.

We mitigate this issue using a staged training procedure. We first train the model using only the supervised ERM loss, and then use the ERM estimator to initialize both the subsequent $\eta$-learning stage as well as the initial tail set used to approximate the $W_1$ penalty. After this pre-training step, the ordering used to select tail samples is anchored to a predictor that already respects the observed input-output relation. Consequently, the $W_1$ penalty primarily acts as a tail correction to an ERM-consistent map, while the ERM term continues to stabilize the learned map around the supervised solution. During the subsequent $\eta$-learning iterations, the tail set is periodically refreshed as the learned map evolves, keeping the samples used to estimate $W_1$ separated from the ERM-induced bulk. This procedure systematically mitigates the conflict caused by unreliable tail-sample selection and has empirically yielded consistently stable and robust training.

The balancing parameter $\lambda$, on the other hand, controls the strength of this residual tail correction. In practice, choosing $\lambda$ to make the $W_1$ gradient comparable in magnitude to the ERM gradient can be beneficial. However, $\lambda$ only rescales the $W_1$ contribution; it does not by itself resolve the directional incompatibility that can arise from unreliable early-stage tail-set selection. Thus, the main stabilization mechanism is ERM pre-training together with tail-set refresh, while $\lambda$ serves as a secondary balancing parameter. This interpretation aligns with the robustness study in SI Fig. 4, where $y_{\eta\#}\mu$ remains stable over a broad range of $\lambda$. A more detailed discussion of the gradient conflict and the practical choice of $\lambda$ is provided in SI Subsection "Pre-training and Inference-Informed Continual Training".

The analysis of the robustness result has also been expanded at line 626:

Additionally, as shown in SI Fig. 4, $y_{\eta\#}\mu$ remains stable as the balancing hyperparameter $\lambda$ varies from $10^{-4}$ to 10, spanning five orders of magnitude, while all other training configurations are fixed. This indicates that our method is numerically robust to the choice of $\lambda$, a practically desirable property. This robustness is enabled by the staged training procedure described later in "Effective Training Strategies".

The SI subsection "Pre-training and Inference-Informed Continual Training" has been revised as follows.

### Pre-training and Inference-Informed Continual Training

Several difficulties arise in the actual optimization of (14). We assume $\mathcal{F}_\theta$ is a neural network optimized with a stochastic first-order optimization method $\mathcal{M}$. The update rule is denoted by

$$\phi_\theta^{(k+1)} \leftarrow \mathcal{M}\left(\phi_\theta^{(k)}, \nabla_\theta \mathcal{L}_\theta^{(k)}\right).$$

Our focus remains on (14), as the other formulation can be viewed as a special case.

***Memory Challenges and Tail-Set Selection in $W_1$ Estimation.*** A practical challenge stems from approximating the $W_1$ term, even with its tractable one-dimensional quantile representation. Accurate estimation requires evaluating $g \circ \phi_\theta$ over the auxiliary set $\mathcal{D}_{\tilde{x}}$, which is typically much larger than the labeled training set. If this full auxiliary evaluation were included in every gradient-carrying update, the memory cost would be substantial.

To address this, we note that only $n_q$ specific inputs from $\mathcal{D}_{\tilde{x}}$ are ultimately needed to compute the empirical quantile loss, namely the inputs whose current observable values realize the

quantile levels in $\mathcal{Q}$. This motivates the Inference-Informed Continual Training (IICT) procedure. Periodically, we run inference on the full auxiliary set $\mathcal{D}_{\tilde{x}}$ and compute

$$\{g(\phi_\theta(\tilde{x}_j))\}_{j=1}^{n_{\tilde{x}}}.$$

We then identify the indices of the inputs corresponding to the prescribed quantile levels in $\mathcal{Q}$ and store them in a set $\mathcal{I}$. Between refreshes, the $W_1$ loss is evaluated only on the $\mathcal{I}$-indexed inputs. Thus, full inference on $\mathcal{D}_{\tilde{x}}$ is used only to identify the relevant tail samples, while the gradient-carrying $W_1$ update is restricted to the selected quantile samples. The refresh frequency is controlled by a hyperparameter $\omega$, with $\mathcal{I}$ updated every $\omega$ iterations. This procedure substantially reduces memory usage while preserving an accurate estimate of the tail contribution to the $W_1$ loss.

***Gradient Conflict and ERM Pre-training.*** The tail-set selection procedure also explains a potential source of optimization instability. The $W_1$ loss is evaluated on inputs selected according to the current model-induced observable distribution $(g \circ \phi_\theta)_{\#}\mu$. Early in training, before the supervised input-output relation has been learned, the ordering of the auxiliary observable values $g(\phi_\theta(\tilde{x}_j))$ can be unreliable. Consequently, inputs that should correspond to bulk behavior under an ERM-consistent predictor may be spuriously selected into $\mathcal{I}$ and treated as high-quantile samples. The $W_1$ term then pulls these samples toward upper quantiles of the reference distribution $\nu_0$, while the ERM loss pulls the model toward the labeled input-output relation. These two signals can be incompatible, leading to oscillatory or unstable optimization.

Our main stabilization mechanism is ERM pre-training. We first minimize the ERM loss alone and initialize $\phi_\theta$ with the resulting ERM estimator $u_\xi$ when solving (14). The initial index set $\mathcal{I}$ used by the $W_1$ penalty is also computed from this ERM-pretrained model. Thus, the first $W_1$ correction is applied to quantile samples selected under a predictor that already captures the observed input-output relation. This substantially reduces the chance that bulk samples are spuriously treated as tail samples because of unreliable early predictions. During $\eta$-learning, $\mathcal{I}$ is periodically refreshed as $\phi_\theta$ evolves, keeping the $W_1$ correction aligned with the current observable distribution.

***Choice of $\lambda$.*** The balancing parameter $\lambda$ controls the relative magnitude of the ERM and $W_1$ gradient components. Writing

$$\mathcal{L}_\theta = L_{\mathrm{ERM}}(\theta) + \lambda L_{W_1}(\theta),$$

we have

$$\nabla_\theta \mathcal{L}_\theta = \nabla_\theta L_{\mathrm{ERM}}(\theta) + \lambda \nabla_\theta L_{W_1}(\theta).$$

A practical scale for $\lambda$ is obtained by choosing it so that the scaled $W_1$ gradient has a comparable Euclidean norm to the ERM gradient during the early $\eta$-learning iterations:

$$\lambda \|\nabla_\theta L_{W_1}(\theta)\|_2 = \|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2 \implies \lambda = \frac{\|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2}{\|\nabla_\theta L_{W_1}(\theta)\|_2 + \varepsilon},$$

with a small $\varepsilon > 0$ for numerical stability. This prevents the $W_1$ correction from overwhelming the supervised anchor.

However, $\lambda$ only rescales the $W_1$ gradient; it does not change its direction. Therefore, while magnitude balancing could be potentially beneficial, it alone cannot resolve the directional incompatibility caused by unreliable tail-set selection. For this reason, ERM pre-training and periodic tail-set refresh are the primary mechanisms for stable optimization, while $\lambda$ serves as a secondary balancing parameter that controls the strength of the residual tail correction. This interpretation

is consistent with the robustness experiment in SI Fig. 4, where the learned observable distribution remains stable as $\lambda$ varies from $10^{-4}$ to 10 while all other training configurations are fixed.

**Methodology comment #vi.** *Theorem 6 relies on Assumption 5, which requires the model to make systematically biased errors without positive and negative errors canceling out. As this is hard to constrain during training, the authors should clarify how its failure impacts the point-wise reliability of the $\eta$-map.*

**Response.** We thank the reviewer for this comment. We agree that Assumption 5 being uniformly satisfied can be hard to impose and should not be interpreted as a property that is automatically enforced by the training objective. We have revised the discussion following Theorem 6 at line 443 to clarify its role and the consequence of its failure. That paragraph now reads as follows.

Theorem 6 reveals an important insight: effective recovery of extreme events is intrinsically linked to the distributional structure of the observable, justifying the $W_1$ term in $\eta$-learning as a distributional regularizer in limited-data regimes. However, the condition that Assumption 5 is uniformly satisfied as $W_1\left(y_{\xi\#}\mu, y_{\#}\mu\right) \to 0$ can be hard to impose during training without additional information. Therefore, if Assumption 5 cannot be assured, convergence in the observable distribution does not, by itself, imply point-wise reliability of the learned $\eta$-map. The learned map should then be interpreted as producing distributionally calibrated candidate extremes, rather than as guaranteeing recovery of their true location, time, or mechanism. Point-wise reliability requires additional identifying information, such as richer observables, spatial or temporal references, or structural and physical constraints during training.

**Methodology comment #vii.** *In the Optimality section, Assumptions 8a and 8b introduce the bounding parameters $\epsilon_1$ and $\epsilon_2$ to formalize $L^1$ deviations in the non-extreme set. A brief physical intuition of these parameters should be provided to improve readability.*

**Response.** We thank the reviewer for the suggestion. We have added this explanation in the Optimality section, immediately after Assumptions 8a and 8b, at line 501. We have also revised the surrounding text to clarify the rationale behind these two assumptions and to improve readability and flow. The revised passage reads as follows.

Intuitively, $\varepsilon_1$ controls the residual error of the ordinary supervised estimator in the non-extreme region, where labeled data are abundant. It therefore represents the baseline bulk approximation error. The parameter $\varepsilon_2$ controls the deviation of the $\eta$-estimator deviates from this supervised estimator in the same non-extreme region. Hence, Assumption 8a requires the baseline estimator to be reasonably accurate in the non-extreme region, while Assumption 8b further requires the discrepancy between the estimators $y_\xi$ and $y_\eta$ to be controlled there. These are natural conditions because both estimators are constrained by the same supervised-learning objective on the data-rich non-extreme region. With these assumptions in place, we are prepared to derive bounds expressed in terms of the exact approximation errors.

**Results comment #i.** *The toy models inadvertently expose the fragility of using a scalar observable $g(u)$. In the 2D-to-1D problem, the physical location of the extreme shifts across initializations (SI Figure 2); a metric, e.g. ensemble spatial variance, should quantify this. Furthermore, in the 2D-to-2D problem, the scalar penalty causes notable spatial distortion via interference between $u_1$*

*and $u_2$ (Figure 2). This interference effect should be discussed as a limitation of the 1D Wasserstein penalty.*

**Response.** We thank the reviewer for this helpful suggestion. We have revised the toy-example discussion to make this limitation explicit, while keeping the main-text discussion concise and moving the quantitative diagnostics to the SI.

For the 2D-to-1D problem, we now quantify the ensemble spatial variability of the inferred peak location. Specifically, using $K = 20$ independently initialized $\eta$-estimators with the same training configurations, we compute

$$\widehat{x}_k = \arg\max_{x \in \mathcal{X}} y_\eta^{(k)}(x), \qquad \bar{x} = \frac{1}{K} \sum_{k=1}^{K} \widehat{x}_k,$$

and

$$\sigma_{\text{loc}}^2 = \frac{1}{K} \sum_{k=1}^{K} \|\widehat{x}_k - \bar{x}\|_2^2.$$

For this problem, the ensemble mean inferred location is

$$\bar{x} = (-0.113, -0.461),$$

and the ensemble spatial variance is

$$\sigma_{\text{loc}}^2 = 15.935, \qquad \sigma_{\text{loc}} = 3.992.$$

These diagnostics show that independently initialized $\eta$-estimators can place the inferred high-output region at substantially different spatial locations. They are now reported in the SI subsection "The 2D-to-1D Toy Problem":

**Ensemble spatial variance.** To quantify the spatial variability of the inferred high-output region across neural-network initializations, we train $K = 20$ independent $\eta$-estimators while keeping the rest of the training configurations fixed. For realization $k$, define

$$\widehat{x}_k = \arg\max_{x \in \mathcal{X}} y_\eta^{(k)}(x).$$

We report the ensemble mean location

$$\bar{x} = \frac{1}{K} \sum_{k=1}^{K} \widehat{x}_k$$

and the ensemble spatial variance

$$\sigma_{\text{loc}}^2 = \frac{1}{K} \sum_{k=1}^{K} \|\widehat{x}_k - \bar{x}\|_2^2.$$

For the 2D-to-1D toy problem, the true grid maximizer is

$$x_\star = (2.020, -2.020),$$

whereas the ensemble mean inferred location is

$$\bar{x} = (-0.113, -0.461).$$

The ensemble spatial variance is

$$\sigma_{\text{loc}}^2 = 15.935, \qquad \sigma_{\text{loc}} = 3.992.$$

Thus, independently initialized $\eta$-estimators can place the inferred high-output region at substantially different spatial locations, quantifying the non-uniqueness presented in SI Fig. 3.

In the main text, we have rephrased the original analysis to improve readability, and added the new diagnostic at line 570:

***Analysis of the results.*** Fig. 1 shows the true map, the MSE map, and one realization of the $\eta$-map, while SI Fig. 1 compares the induced distributions $y_{\#}\mu$, $y_{\eta\#}\mu$, and $y_{\xi\#}\mu$. Additional $\eta$-map realizations and their corresponding PDF comparisons are shown in SI Figs. 2 and 3. The MSE map accurately fits the data-rich region but fails to infer the unobserved target extreme region, illustrating the data-consistent setting of Definition 3 and the fundamental limitation of MSE-based learning under tail data scarcity. In contrast, the $\eta$-map does not exactly recover the true extreme structure, but it introduces high-output mass in the tail while preserving the prediction in the data-rich region. This behavior is expected by theory: because the assumptions of Theorem 6 are not enforced as constraints during training, exact pointwise recovery is not guaranteed, and the available guarantee is convergence in $L^1(\mu)$. Distributionally, however, SI Fig. 1 shows that $y_{\eta\#}\mu$ is substantially closer to the reference distribution, which is the truth in this example, than the MSE-induced law, especially in the tail. Thus, $\eta$-learning reduces the epistemic uncertainty caused by missing extreme samples by producing statistically plausible extremes, rather than by uniquely identifying their exact spatial configuration.

The additional realizations in SI Fig. 2 show that distributional convergence does not uniquely determine the spatial location of the inferred high-output region. We quantify this variability in the SI by training $K = 20$ independently initialized $\eta$-estimators under the same configuration. The resulting maximizer locations have ensemble spatial standard deviation $\sigma_{\text{loc}} = 3.992$, while the 2-norm of the true maximizer is 2.86. Thus, independently initialized $\eta$-estimators can place the inferred high-output region at substantially different spatial locations.

The ensemble analysis highlights a broader implication of $\eta$-learning: enforcing scalar observable tail statistics generates candidate extremes, but does not necessarily identify the underlying physical structure, such as the peak location shown in this example. This variability reflects the residual physical uncertainty that remains after enforcing the observable tail statistics, rather than a failure to recover a uniquely identifiable hidden event. Nevertheless, two aspects are stable across realizations: predictions remain accurate in the data-rich region, and, as shown in SI Fig. 3, the induced output PDFs remain consistent with the reference distribution. In the Discussion section, we further describe how this uncertainty can be systematically characterized and used in downstream applications.

For the 2D-to-2D problem, rather than attributing the spatial distortion to noisy information in $u_2$, we now describe it as a structural limitation of using a scalar observable. We have also rephrased the previous analysis to improve readability. The result analysis now reads as follows at line 640:

***Analysis of the results.*** Fig. 2 shows the true map, the MSE map, and one realization of the $\eta$-map, while SI Fig. 5 compares the induced observable distributions $y_{\#}\mu$, $y_{\eta\#}\mu$, and $y_{\xi\#}\mu$. Additional $\eta$-map realizations and their corresponding PDF comparisons are shown in SI Figs. 6

and 7. As in the previous example, the MSE estimator accurately captures the map only in the data-rich region, but fails to detect the unobserved target extreme region. In contrast, the $\eta$-map uses the reference observable distribution to generate high-output responses in the missing tail region. Although the reconstructed extremes are not pointwise exact, the induced observable PDF remains closely aligned with the reference distribution, as shown in SI Fig. 5.

This example also exposes a limitation of constraining a scalar observable for high-dimensional states. Since $g(u) = 2|u_1| + \frac{1}{2}|u_2|$ aggregates the two state components into a single scalar quantity, the $W_1$ penalty constrains the observable law, but not how the extreme response decomposes between the two components. Compared with the 2D-to-1D case, this introduces an additional ambiguity: distortion in one state variable can be partially compensated by changes in the other while preserving similar scalar observable statistics. This scalarization-induced interference explains the more pronounced spatial distortion visible in Fig. 2, especially in $u_1$.

The additional realizations in SI Fig. 6 further show variability in the inferred extreme-region structure across neural-network initializations. Nevertheless, the induced observable PDFs remain consistently aligned with the reference distribution, as shown in SI Fig. 7. Overall, these results demonstrate that $\eta$-learning can generate plausible tail events under missing extreme data, but scalar-observable matching alone may leave residual uncertainty in the component-wise structure of high-dimensional states.

**Results comment #ii.** *For the first precipitation downscaling challenge, the degradation of the weighted coverage statistic at lower thresholds (SI Figures 8 and 9) indicates that penalizing the spatial maximum explicitly costs bulk spatial coherence. The authors should report standard spatial similarity metrics, e.g. SSIM or full-grid RMSE, alongside the extreme metrics to quantify how much spatial structure is sacrificed.*

**Response.** We thank the reviewer for this suggestion. We have added full-field spatial-fidelity diagnostics for the architecture-matched MSE and $\eta$ downscalers. Specifically, we now report full-grid RMSE as well as mean and standard deviation of SSIM. These metrics are computed over the complete HR precipitation fields.

The added diagnostics are evaluated over all $N = 9{,}044$ retained paired fields and over subsets defined by the true HR observable distribution. The bulk subsets are defined by $g(u) \leq Q_{\text{true}}(q)$ for $q \in \{0.7, 0.8, 0.9\}$, and the tail subsets are defined by $g(u) \geq Q_{\text{true}}(q)$ for $q \in \{0.95, 0.975, 0.99\}$.

The results confirm that emphasizing the spatial maximum incurs a measurable cost in pointwise fidelity, compared with the MSE-map. Over the full evaluation set, the full-grid RMSE increases from 2.877611 mm for the MSE downscaler to 3.111615 mm for the $\eta$ downscaler, an 8.13% relative increase. In the bulk subsets, the relative RMSE increases are smaller, ranging from 6.20% to 6.92%. In the tail subsets, the RMSE cost is larger, ranging from 9.21% to 10.30%, reflecting the greater difficulty of matching pointwise intensities on more extreme fields.

The SSIM diagnostics indicate that this pointwise cost is not accompanied by a comparable degradation of spatial structure. On the full evaluation set, $\overline{\text{SSIM}}$ changes from 0.947003 to 0.944207, a 0.30% relative decrease. Across all reported subsets, the largest relative $\overline{\text{SSIM}}$ decrease is 0.48%, occurring in the $q \geq 0.95$ tail regime. $\sigma_{\text{SSIM}}$ is slightly larger for the $\eta$ downscaler in each stratum, but the increase is small. These results refine the interpretation of SI Figs. 8 and 9: the lower-threshold weighted-coverage degradation reflects a real change in the magnitude-weighted allocation of precipitation above moderate thresholds, but it does not indicate a broad collapse of

spatial morphology.

We have revised the result analysis in the main text at line 718:

We further evaluate two statistics that are not directly enforced during training: the conditional mean $m(u_0; t, n_g)$ and the weighted coverage $c(u_0; t, n_g)$, where $u_0$ is a sample field, $t$ is a threshold, and $n_g$ is the number of spatial grid cells. The conditional mean measures the average magnitude above a threshold, while the weighted coverage measures the fraction of total field magnitude contributed by entries exceeding that threshold. Formal definitions are given in the SI. For $m$, we compute the PDFs across a range of thresholds from 154 to 214, and similarly for $c$. As shown in SI Fig. 8, the $u_\eta$-mapped samples consistently outperform the $u_\xi$-mapped samples with respect to the conditional mean statistic, further confirming the effectiveness of $\eta$-learning in capturing extreme values. SI Fig. 9 shows a trade-off at lower thresholds: the $u_\eta$-mapped weighted-coverage distribution agrees less well with the truth than the $u_\xi$-mapped distribution, especially in the bulk. This is expected because the training objective constrains the spatial maximum, not the magnitude-weighted exceedance fraction. However, as the threshold increases, the MSE-PDF progressively collapses into a Dirac delta, reflecting its failure to represent extremes, whereas the $\eta$-PDF continues to provide a valid proxy of the true distribution.

To further quantify the associated full-field cost, we provide RMSE and structural similarity index measure (SSIM) diagnostics in SI Table 1. Over all 9044 paired fields, the full-grid RMSE increases from 2.878 for $u_\xi$ to 3.112 for $u_\eta$, an 8.13% relative increase. Average SSIM ($\overline{\text{SSIM}}$) changes only from 0.947 to 0.944, a 0.30% relative decrease. We also evaluate subsets stratified by $\nu_0$: define $\mathcal{I}_{\le q} = \{1 \le j \le 9044 : \max(u_j) \le F_{\nu_0}^{-1}(q)\}$ and $\mathcal{I}_{\ge q}$ similarly, where $F_{\nu_0}^{-1}(q)$ denotes the $q$-quantile of $\nu_0$. For bulk subsets $\mathcal{I}_{\le 0.7}$, $\mathcal{I}_{\le 0.8}$, and $\mathcal{I}_{\le 0.9}$, the $u_\eta$-RMSE increase ranges from 6.20% to 6.92%, while the $u_\eta$-$\overline{\text{SSIM}}$ decrease remains below 0.27%. For tail subsets $\mathcal{I}_{\ge 0.95}$, $\mathcal{I}_{\ge 0.975}$, and $\mathcal{I}_{\ge 0.99}$, the $u_\eta$-RMSE increase ranges from 9.21% to 10.30%, while the $u_\eta$-$\overline{\text{SSIM}}$ decrease remains below 0.48%. These results show that emphasizing the scalar observable introduces a modest pointwise-intensity cost, with a larger effect on high-maximum fields, but does not produce a significant degradation in the local spatial structure measured by SSIM. Additional details of the spatial-fidelity analysis are provided in the SI.

Taken together, these diagnostics clarify the trade-off. On one hand, $\eta$-learning captures the prescribed extreme-observable statistics and transfers this improvement to related high-threshold statistics that are not explicitly enforced. On the other hand, this improvement comes with a modest spatial fidelity cost, primarily through pointwise amplitudes rather than a broad degradation of spatial morphology.

In SI subection *The Real-World Precipitation Downscaling Challenge*, immediately after the definitions of the conditional mean and weighted coverage statistics at line 1210, we add:

**Full-field spatial-fidelity metrics.** The conditional mean and weighted coverage characterize thresholded aggregate statistics of each precipitation field. They do not directly measure pointwise reconstruction accuracy or local spatial structure. We therefore supplement them with full-grid RMSE and structural similarity index measure (SSIM).

Let $u_j$ and $\widehat{u}_j$ denote the true and predicted HR fields for sample $j$, and let $n_g = 80 \cdot 160$ denote the number of HR grid cells. We compute

$$\text{RMSE} = \left[ \frac{1}{N n_g} \sum_{j=1}^{N} \|\widehat{u}_j - u_j\|_2^2 \right]^{1/2}.$$

Here, $N$ depends on the selected subset. We also compute the SSIM for each prediction-truth pair on the full HR field. Let $\mathcal{W}$ denote the set of $7 \times 7$ local windows. For a window $w \in \mathcal{W}$, define

$$\mathrm{SSIM}_w(\widehat{u}_j, u_j) = \frac{(2m_{\widehat{u}_{j,w}} m_{u_j,w} + C_1)(2\sigma_{\widehat{u}_j u_j,w} + C_2)}{(m^2_{\widehat{u}_{j,w}} + m^2_{u_j,w} + C_1)(\sigma^2_{\widehat{u}_{j,w}} + \sigma^2_{u_j,w} + C_2)},$$

where $m_{\widehat{u}_{j,w}}$ and $m_{u_j,w}$ are the local means, $\sigma^2_{\widehat{u}_{j,w}}$ and $\sigma^2_{u_j,w}$ are the local variances, and $\sigma_{\widehat{u}_j u_j,w}$ is the local covariance. The stabilizing constants are

$$C_1 = (0.01L)^2, \qquad C_2 = (0.03L)^2,$$

with the common intensity range

$$L = \max_{j,z} u_j(z) - \min_{j,z} u_j(z),$$

where the maximum and minimum are taken over the true HR set. The SSIM is then

$$\mathrm{SSIM}(\widehat{u}_j, u_j) = \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \mathrm{SSIM}_w(\widehat{u}_j, u_j).$$

Numerically, SSIM is a normalized local similarity score whose maximum value is 1, attained when the predicted and true fields agree within the local window up to numerical precision. Values close to 1 indicate that the two fields have similar local mean intensity, contrast, and spatial covariation, even if their pointwise amplitudes are not identical. Thus, SSIM complements RMSE: RMSE measures pointwise amplitude error in physical units, whereas SSIM measures whether the local spatial pattern is preserved after normalization by the local means and variances. Using a common intensity range $L$ for all samples makes the SSIM values comparable across subsets. Using a common intensity range $L$ for all samples makes the SSIM values comparable across subsets. We report the sample mean $\overline{\mathrm{SSIM}}$ and sample standard deviation $\sigma_{\mathrm{SSIM}}$ for each selected subset.

For the $\eta$-map, we also report the relative RMSE increase compared to the MSE-map,

$$r_{\eta/\mathrm{MSE}} = \frac{\mathrm{RMSE}_\eta - \mathrm{RMSE}_{\mathrm{MSE}}}{\mathrm{RMSE}_{\mathrm{MSE}}}.$$

The bulk subsets are defined as

$$\mathcal{I}_{\le q} := \{1 \le j \le 9044 : \max(u_j) \le F^{-1}_{\nu_0}(q)\},$$

and the tail subsets are defined as

$$\mathcal{I}_{\ge q} := \{1 \le j \le 9044 : \max(u_j) \ge F^{-1}_{\nu_0}(q)\}.$$

Over the full evaluation set with 9044 samples, the $\eta$ downscaler has an 8.13% larger full-grid RMSE than the MSE downscaler. The $\overline{\mathrm{SSIM}}$ decrease is 0.30%, and $\sigma_{\mathrm{SSIM}}$ increases by 5.36%. Thus, the main full-field cost is pointwise intensity error rather than a large loss of local spatial structure.

The bulk subsets show the same pattern with a smaller RMSE cost. For $\mathcal{I}_{\le 0.7}$, $\mathcal{I}_{\le 0.8}$, and $\mathcal{I}_{\le 0.9}$, the relative RMSE increases are 6.20%, 6.55%, and 6.92%, respectively. The corresponding $\overline{\mathrm{SSIM}}$ decreases are 0.19%, 0.23%, and 0.27%. Hence, in the bulk and moderate regimes, the $\eta$-map sacrifices a small amount of pointwise fidelity while retaining high structural similarity.

| Subset | Method | $N$ | RMSE | $\overline{\text{SSIM}}$ | $\sigma_{\text{SSIM}}$ | $r_{\eta/\text{MSE}}$ |
|---|---|---|---|---|---|---|
| Full | MSE | 9044 | 2.877611 | 0.947003 | 0.043672 | – |
| | $\eta$ | | 3.111615 | 0.944207 | 0.046011 | 8.13% |
| $\mathcal{I}_{\leq 0.7}$ | MSE | 6331 | 1.739108 | 0.965508 | 0.029943 | – |
| | $\eta$ | | 1.846915 | 0.963688 | 0.031964 | 6.20% |
| $\mathcal{I}_{\leq 0.8}$ | MSE | 7235 | 2.008767 | 0.960048 | 0.033681 | – |
| | $\eta$ | | 2.140332 | 0.957853 | 0.035877 | 6.55% |
| $\mathcal{I}_{\leq 0.9}$ | MSE | 8139 | 2.344438 | 0.954014 | 0.037959 | – |
| | $\eta$ | | 2.506664 | 0.951470 | 0.040278 | 6.92% |
| $\mathcal{I}_{\geq 0.95}$ | MSE | 453 | 6.469183 | 0.877379 | 0.040998 | – |
| | $\eta$ | | 7.130227 | 0.873164 | 0.042823 | 10.22% |
| $\mathcal{I}_{\geq 0.975}$ | MSE | 227 | 7.191779 | 0.875916 | 0.040929 | – |
| | $\eta$ | | 7.932229 | 0.872659 | 0.042427 | 10.30% |
| $\mathcal{I}_{\geq 0.99}$ | MSE | 91 | 7.904075 | 0.870943 | 0.038264 | – |
| | $\eta$ | | 8.631702 | 0.869387 | 0.039693 | 9.21% |

Table 1: HR spatial-fidelity metrics for the precipitation downscalers. The subsets are selected by the true HR observable distribution: $\mathcal{I}_{\leq q}$ consists of sample indices whose corresponding true HR spatial maximum falls below the $q$-quantile, and $\mathcal{I}_{\geq q}$ is similarly defined. Here $r_{\eta/\text{MSE}} = (\text{RMSE}_\eta - \text{RMSE}_{\text{MSE}})/\text{RMSE}_{\text{MSE}}$ denotes the relative RMSE increase of the $\eta$-downscaler over the MSE-downscaler.

In the tail subsets, both downscalers have larger RMSE because the selected fields are more extreme. The $\eta$-downscaler has relative RMSE increases of 10.22%, 10.30%, and 9.21% for $\mathcal{I}_{\geq 0.95}$, $\mathcal{I}_{\geq 0.975}$, and $\mathcal{I}_{\geq 0.99}$ respectively. However, $\overline{\text{SSIM}}$ remains close to the MSE baseline, with relative decreases of 0.48%, 0.37%, and 0.18%. These results show that the spatial-maximum penalty induces a measurable but limited full-field cost. Together with SI Fig. 9, they indicate that the lower-threshold weighted-coverage degradation reflects a change in the magnitude-weighted allocation of precipitation above moderate thresholds, not a broad collapse of spatial morphology.

**Results comment #iii.** *The case study using a hypothesized GEVD prior (SI Figure 12) highlights that the model will faithfully hallucinate extremes to match a potentially flawed distribution. A quantitative sensitivity analysis, i.e. ablation study, showing how drastically physical fidelity degrades when $\nu_0$ is misspecified by a known margin would be valuable.*

**Response.** We thank the reviewer for this suggestion. We have added a controlled misspecification analysis for the hypothesized-GEVD experiment. The goal is to quantify how errors in the supplied scalar tail information affect the fidelity of the resulting HR precipitation fields.

We perturb the supplied reference values directly at the selected quantile levels $\mathcal{Q} = \{q_i\}_{i=1}^{n_q}$. Let $Q_{\text{true}}$ and $Q_{\text{GEVD}}$ denote the empirical HR and hypothesized GEVD quantile functions of the spatial-maximum observable. We define

$$Q_\alpha(q_i) = Q_{\text{true}}(q_i) + \alpha\,[Q_{\text{GEVD}}(q_i) - Q_{\text{true}}(q_i)]\,, \qquad q_i \in \mathcal{Q}.$$

Here, $\alpha = 0$ gives the empirical HR tail target and $\alpha = 1$ gives the hypothesized GEVD target.

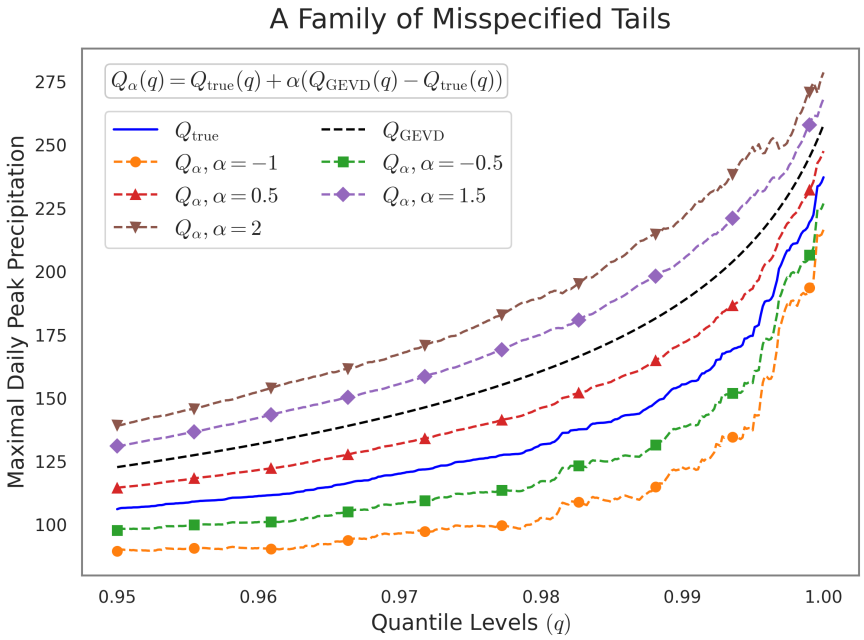



Figure 1: Sensitivity analysis for misspecified reference tail. Tail-target profiles obtained by shifting the true HR quantiles using the parameter $\alpha$.

Negative $\alpha$ shifts the target in the lighter-tail direction, while $\alpha > 1$ extrapolates beyond the original GEVD target. Defining

$$\Delta_\alpha = \frac{1}{n_q} \sum_{i=1}^{n_q} |Q_\alpha(q_i) - Q_{\text{true}}(q_i)|,$$

we have

$$\Delta_\alpha = |\alpha| \Delta_1.$$

Thus, $|\alpha|$ measures the size of the perturbation relative to the discrepancy between the hypothesized GEVD and empirical HR tail targets. We evaluated five misspecified-tail models with

$$\alpha \in \{-1, -0.5, 0.5, 1.5, 2\},$$

while holding the remaining training configurations fixed. The family of tail-shifted $\nu_0$ is provided in Fig. 1, and the corresponding $\eta$-estimated HR observable distributions are presented in Fig. 2.

The results show clear sensitivity of the pointwise spatial fidelity to heavier-tail misspecification. The full-grid RMSE increases from 2.926038 at $\alpha = -1$ to 3.527282 at $\alpha = 2$, a 20.55% increase. The intermediate values are ordered by tail strength: $\alpha = -0.5$ remains close to $\alpha = -1$, $\alpha = 0.5$ gives a moderate increase, and $\alpha = 1.5$ is close to the largest-error end of the sweep. At the same perturbation magnitude, the heavier-tail direction is more damaging than the lighter-tail direction: $\alpha = 0.5$ gives RMSE 3.193944, compared with 2.967890 for $\alpha = -0.5$. Within the positive heavy-tail sweep, increasing the perturbation from $0.5\Delta_1$ to $2\Delta_1$ increases RMSE from 3.193944 to 3.527282, a 10.43% increase.

$\overline{\text{SSIM}}$ remains high across the sweep, between 0.940689 and 0.947280. It is not strictly monotone: the $\alpha = 2$ model has the largest RMSE but a slightly higher $\overline{\text{SSIM}}$ than the $\alpha = 1.5$ model. $\sigma_{\text{SSIM}}$

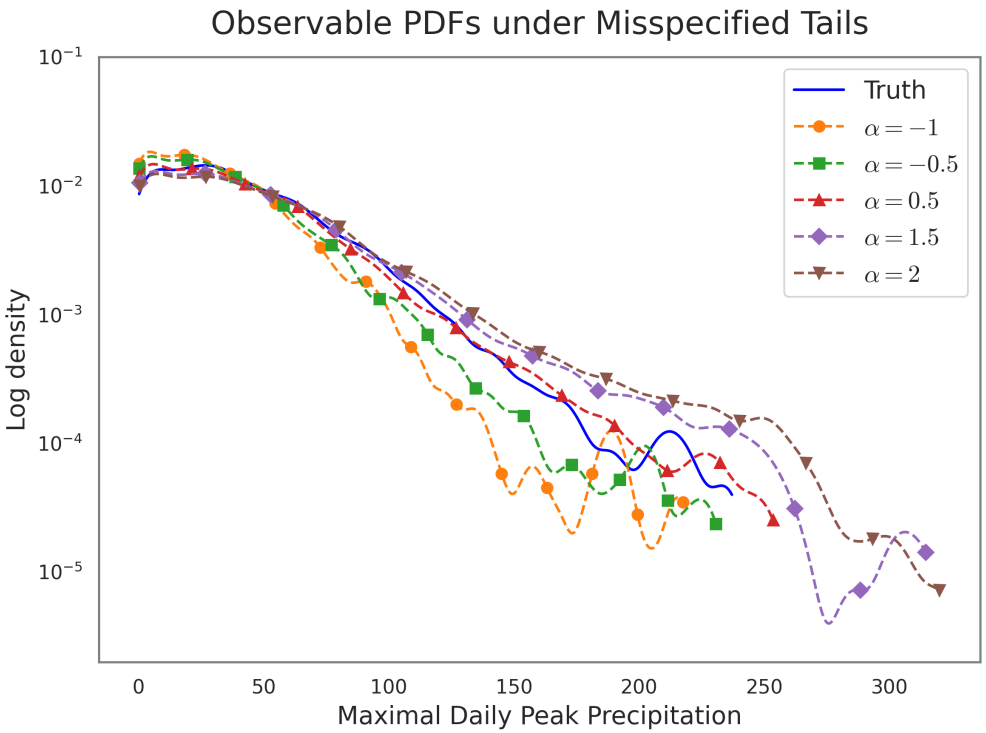



Figure 2: Sensitivity analysis for misspecified reference tail. Resulting observable distributions from $\eta$-downscalers trained with the corresponding shifted $\nu_0$ shown in Fig. 1.

increases mildly as the tail is made heavier, from 0.043792 at $\alpha = -1$ to 0.047701 at $\alpha = 2$. These results indicate that the misspecification primarily affects pointwise fidelity, while the aggregate spatial structure measured by SSIM remains comparatively stable.

We have revised the result analysis in the main text at line 917:

***Analysis of the results.*** The visualizations of the resulting downscaled fields and the corresponding observable PDFs are presented in SI Fig. 12 and Fig. 6. The observable law of the $\eta$-downscaled fields is shifted toward the hypothesized GEVD and therefore has a heavier tail than the true HR law. This experiment shows that the framework can generate fields conditional on tail information extending beyond the observed record, thereby broadening our understanding of potential extreme scenarios by leveraging only hypothetical tail information—a critical advantage in risk-sensitive applications.

At the same time, the experiment does not validate the hypothesized $\nu_0$ or the amplified precipitation values produced under that hypothesis. This distinction is important: since the distributional regularizer is designed to enforce the supplied tail statistics, if the tail is misspecified, the learned map can generate correspondingly misspecified extreme amplitudes. We quantify this effect through the controlled sensitivity analysis reported in SI Table 2, with details provided in the same SI subsection. The supplied tail quantiles are perturbed by a scalar parameter $\alpha$, for which the misspecification magnitude is proportional to $|\alpha|$. As the target is moved in the heavier-tail direction from $\alpha = -1$ to $\alpha = 2$, the full-grid $u_\eta$-RMSE increases from 2.926 to 3.527, a 20.55% increase. $u_\eta$-$\overline{\text{SSIM}}$ remains between 0.941 and 0.947, indicating that the main degradation is in

amplitude-sensitive pointwise fidelity rather than in the spatial morphology measured by SSIM.

Accordingly, SI Fig. 12 should be interpreted as showing scenarios conditional on the hypothesized GEVD. The fields illustrate how the learned downscaling map responds when an assumed heavier tail is imposed, rather than providing evidence that the amplified extremes will occur or that their magnitudes are physically validated. The experiments therefore demonstrate both the scenario-generation capability of the framework and its dependence on the accuracy of the supplied reference information. Such conditional scenarios can help inform risk assessment and decision-making when the reference tail is treated as an explicit modeling hypothesis and accompanied by proper sensitivity analysis.

In SI subsection "$\eta$-Downscaling with Hypothesized Heavier-Tailed Distributions", we add:

**Sensitivity to misspecified tails.** We examine how errors in the supplied tail information propagate to the downscaled precipitation fields. The $W_1$ estimator uses $\nu_0$ through its quantiles at the selected probability levels

$$\mathcal{Q} = \{q_i\}_{i=1}^{n_q}, \qquad n_q = 350, \qquad q_i \in [0.95, 1].$$

We thus perturb $\nu_0$ directly at $\mathcal{Q}$. Let $Q_{\text{true}}(q)$ denote the true HR observable quantile, and let $Q_{\text{GEVD}}(q)$ denote the corresponding quantile of the hypothesized truncated GEVD. We define

$$Q_\alpha(q_i) = Q_{\text{true}}(q_i) + \alpha \left[ Q_{\text{GEVD}}(q_i) - Q_{\text{true}}(q_i) \right], \qquad q_i \in \mathcal{Q}.$$

Here, $\alpha = 0$ gives the empirical HR target, while $\alpha = 1$ gives the hypothesized GEVD target. Negative $\alpha$ moves the supplied tail information in the lighter-tail direction, whereas $\alpha > 1$ extrapolates beyond the GEVD target.

The average magnitude of the imposed tail perturbation is

$$\Delta_\alpha = \frac{1}{n_q} \sum_{i=1}^{n_q} \left| Q_\alpha(q_i) - Q_{\text{true}}(q_i) \right|.$$

Since

$$Q_\alpha(q_i) - Q_{\text{true}}(q_i) = \alpha \left[ Q_{\text{GEVD}}(q_i) - Q_{\text{true}}(q_i) \right],$$

we have

$$\Delta_\alpha = |\alpha| \Delta_1.$$

Thus, $|\alpha|$ specifies the misspecification magnitude relative to the discrepancy between the hypothesized GEVD tail and the true HR tail. For each value of

$$\alpha \in \{-1, -0.5, 0.5, 1.5, 2\},$$

we use the same training configuration except that $\nu_0$ varies across the perturbed family. We evaluate the complete predicted HR fields using RMSE together with $\overline{\text{SSIM}}$ and $\sigma_{\text{SSIM}}$. All metrics are computed over the $N = 9{,}044$ HR fields.

The full-grid RMSE increases as the supplied target is shifted in the heavier-tail direction. The lowest RMSE occurs at $\alpha = -1$, with RMSE $= 2.926038$, while the largest occurs at $\alpha = 2$, with RMSE $= 3.527282$. This is a 20.55% increase relative to the $\alpha = -1$ case. The intermediate values follow the same ordering by tail strength: $\alpha = -0.5$ remains close to $\alpha = -1$, $\alpha = 0.5$ is moderately larger, and $\alpha = 1.5$ is close to the largest-error end of the sweep. At equal misspecification magnitude, the heavier-tail direction is more damaging: $\alpha = 0.5$ gives RMSE 3.193944, whereas

| $\alpha$ | RMSE | $\overline{\text{SSIM}}$ | $\sigma_{\text{SSIM}}$ |
|---|---|---|---|
| $-1.0$ | 2.926038 | 0.947280 | 0.043792 |
| $-0.5$ | 2.967890 | 0.946510 | 0.044193 |
| 0.5 | 3.193944 | 0.944542 | 0.045382 |
| 1.5 | 3.423614 | 0.940689 | 0.048021 |
| 2.0 | 3.527282 | 0.941217 | 0.047701 |

Table 2: Sensitivity of the full-field fidelity of the $\eta$-mapped fields to misspecification of the supplied reference tail. All metrics are computed over the full $N = 9044$ paired HR fields.

$\alpha = -0.5$ gives RMSE 2.967890. Within the positive heavy-tail sweep, increasing the target discrepancy from $0.5\Delta_1$ to $2\Delta_1$ increases RMSE by 10.43%.

The SSIM behavior is more stable. $\overline{\text{SSIM}}$ remains between 0.940689 and 0.947280 for all misspecified-tail models. It is not strictly monotone in $\alpha$: the $\alpha = 2$ model has the largest RMSE but a slightly higher $\overline{\text{SSIM}}$ than the $\alpha = 1.5$ model. $\sigma_{\text{SSIM}}$ increases mildly from 0.043792 at $\alpha = -1$ to 0.047701 at $\alpha = 2$. Thus, heavier-tail misspecification mainly degrades amplitude-sensitive pointwise fidelity, while the aggregate spatial structure measured by SSIM remains relatively close to the truth.

This experiment confirms that the pointwise fidelity of the $\eta$-generated fields is sensitive to the supplied reference tail. A heavier-than-correct reference profile can induce mismatched pointwise precipitation amplitudes even when the fields retain high SSIM. The GEVD-based outputs should therefore be interpreted conditionally on the hypothesized tail law, rather than as physically validated predictions.

**Results comment #iv.** *The algorithm requires dynamic approximation of the 1-Wasserstein distance, yet a discussion on computational complexity is missing. The authors should include a brief quantitative analysis to set realistic expectations, particularly concerning the overhead of adding this to already expensive generative pipelines like Flow Matching.*

**Response.** We thank the reviewer for raising this concern. We have added a detailed computational complexity analysis to the SI and referenced it in the main text at line 1140:

A detailed computational complexity analysis of the full procedure is provided in SI subsection "Computational Complexity".

We emphasize that, in the generative-modeling experiment, $\eta$-learning does not modify the training or sampling procedure of the Flow Matching (FM) model. The low-resolution (LR) FM model is trained and used to generate LR samples exactly as when no $\eta$-Learning is involved. The $\eta$-map is trained separately and applied only after FM sampling. This requires one additional forward pass per generated LR sample. In particular, FM sampling involves no $W_1$ computation or backward propagation through the $\eta$-map. Therefore, aside from the separate one-time training of the $\eta$-map, the only generation-time overhead relative to the baseline FM pipeline is the forward propagation of the generated LR samples through the trained $\eta$-map.

The complexity analysis subsection added to the SI at line 942 is attached below.

### Computational Complexity

Here, a detailed complexity analysis of the full $\eta$-Learning algorithm, i.e. Algorithm 1, is provided. We distinguish the inference-only computation used to identify the relevant tail samples from the gradient-carrying computation used to update the neural network. Let $n_{\tilde{x}} = |\mathcal{D}_{\tilde{x}}|$ be the number of inputs used to estimate the push-forward distribution, $n_q = |\mathcal{Q}|$ the number of quantiles for estimating $W_1$, and $C_\mathrm{f}$ and $C_\mathrm{b}$ the per-sample forward- and backward-propagation costs through $g \circ \phi_\theta$, respectively. Repeated input indices are allowed when different quantile levels correspond to the same empirical sample; hence, the gradient-carrying $W_1$ loss always contains $n_q$ samples.

We consider the conservative case $\omega = 1$, in which case the quantile indices are updated at every optimization iteration. First, $g \circ \phi_\theta$ is evaluated on all $n_{\tilde{x}}$ inputs in inference mode, and the resulting scalar values are sorted to identify the prescribed quantiles. This step has cost

$$O\left(n_{\tilde{x}} C_\mathrm{f} + n_{\tilde{x}} \log n_{\tilde{x}}\right).$$

Because this full-sample evaluation is forward-only, it does not require retaining a computational graph. The $W_1$ loss is then computed and back-propagated using only on the $n_q$ quantile-indexed samples, with cost

$$O\left(n_q (C_\mathrm{f} + C_\mathrm{b})\right).$$

Therefore, the additional per-iteration cost of the $W_1$ regularization is

$$O\left(n_{\tilde{x}} C_\mathrm{f} + n_{\tilde{x}} \log n_{\tilde{x}} + n_q (C_\mathrm{f} + C_\mathrm{b})\right).$$

The main practical distinction is that forward propagation is performed on the full input pool, whereas the more expensive backward propagation is restricted to the $n_q$ dynamically selected tail samples. Note that $n_{\tilde{x}}$ and $n_q$ can be mini-batched if the memory cap is of concern.

In the precipitation experiment, $n_{\tilde{x}} = 9044$, and the matched tail begins at approximately the 0.975 quantile. The quantile grid therefore contains approximately

$$n_q \simeq (1 - 0.975) n_{\tilde{x}} \simeq 226$$

terms. Thus, although all 9044 inputs are used in the inference-only forward pass, only approximately 2.5% as many samples are used in the gradient-carrying $W_1$ update.

For the Flow Matching experiment, the LR Flow Matching model is trained on the LR data and used to generate the 25-year-equivalent LR samples exactly as it would be without $\eta$-learning. The trained $\eta$-map is then applied once to each generated LR sample to obtain a tail-corrected HR sample. For fair comparison, $n_{\tilde{x}}$ LR samples are generated, and the additional $\eta$-evaluation cost is therefore

$$O\left(n_{\tilde{x}} C_\mathrm{f}\right).$$

This is a single inference-only post-processing pass per sample. It does not alter the training cost of the Flow Matching model, does not add computations to its numerical sampling steps, and requires no $W_1$ evaluation or backward propagation. The other additional expense is the separate, one-time training of the $\eta$-map, whose per-iteration complexity is given above.

**Results comment #v.** *While the Discussion highlights the framework's versatility, it lacks critical evaluation. The authors must add a "Limitations and Future Work" subsection summarizing the major caveats identified above. Acknowledging these will strengthen the manuscript and guide future research toward multi-dimensional observables.*

**Response.** We thank the reviewer for the suggestion. We have added a dedicated subsection titled "Limitations and Future Work" at line 1033 to acknowledge the major caveats and discuss important future directions. The subsection reads as follows.

### Limitations and Future Work

While $\eta$-learning offers a promising approach for modeling extreme events under data scarcity, several important limitations remain. Here, we identify key limitations and corresponding directions for future work.

A central limitation is that matching the distribution of a prescribed observable does not identify the dynamical mechanism that generates extreme events. The $W_1$ regularizer constrains observable statistics, but does not by itself determine the temporal evolution, spatial organization, or broader physical coherence of the generated extremes. To the extent that these properties are recovered, they are inherited from the supervised input-output data rather than supplied by the $W_1$ term. Consequently, when distinct mechanisms or state configurations induce the same observable law, they are not distinguishable under the present objective. An important direction is therefore to extend the current framework with temporal and spatial constraints, conservation laws, or other physical priors.

This non-identifiability also arises at the level of individual events: the prescribed observable law does not generally determine the precise location, occurrence time, or full spatial structure of an unobserved extreme. Variability in these features across independently trained $\eta$-maps should therefore be interpreted as residual uncertainty. Importantly, this uncertainty can be exploited systematically. Ensembles of $\eta$-maps can generate diverse yet statistically consistent candidate extremes, which can then be organized according to physically meaningful attributes such as event location, occurrence time, or spatial pattern. Representative scenarios can guide subsequent rounds of adaptive data acquisition, such as targeted high-fidelity simulations or experiments, allowing the corresponding hypotheses to be tested, refined, or rejected while progressively reducing uncertainty in data-scarce regions.

Another limitation is the reliance on an accurate reference law $\nu_0$. Because the tail-emphasized $W_1$ estimation is designed to match the supplied tail, misspecification of $\nu_0$, or a shift in the target tail relative to $\nu_0$, can be transferred directly to the generated extremes. A promising extension is a distributionally robust formulation that replaces a single reference law with an ambiguity set of plausible laws, with particular emphasis on uncertainty and shifts in the tail. Worst-case or uncertainty-weighted tail objectives could reduce overcommitment to any single hypothesized distribution while preserving awareness of extreme behavior.

Finally, the present theory and algorithms are developed for scalar observables, for which $W_1$ enables efficient computation and underpins the tail optimality results. For a vector observable $G = (g_1, \ldots, g_r)$, one may either introduce a physically meaningful severity function $s : \mathbb{R}^r \to \mathbb{R}$ and apply the present framework to $g = s \circ G$, or use a sum of componentwise $W_1$ penalties when matching the marginal laws is sufficient. The latter scales linearly with $r$, but does not constrain the correlation structure among the components. When the joint law of $G$ is essential, genuinely multivariate discrepancies based on, for example, sliced Wasserstein distance or entropically regularized optimal transport are required [1, 2]. Extending the present tail-focused theory and scalable algorithms to these settings is another important direction.

**Results comment #vi.** *The empirical reproducibility of the work is overall high. The authors provide thorough details regarding the ERA5-Land dataset processing, hyperparameter choices, and*

*algorithmic training loops.*

**Response.** We appreciate this assessment. We have further improved reproducibility by expanding the code documentation, pinning dependencies, adding hardware/runtime notes, and providing a minimal ERA5-Land test pipeline as described below.

**Code availability.** *First, it should be noted that I reviewed the provided code repository without installing or running the scripts. The repository maps to the experiments in the manuscript. The authors provide separate notebooks for the different case studies, the 2D toy problems, ERA5-Land downscaling, and the generative model integration. While the code for the mathematical implementation is there, it would be very helpful if the authors added inline comments linking the core definitions directly to the numbered equations in the text. This would make it much easier for readers to follow the translation from theory to code.*

*Regarding usability, the README needs to be updated with proper environment setup instructions, ideally through a* `requirements.txt` *file. Please ensure that the exact versions of the core dependencies are pinned so the codebase does not break due to future library updates. Finally, a brief note on the expected hardware requirements, such as required GPU VRAM and approximate processing times, should be added. This is useful for anyone looking to reproduce these results.*

*Finally, for the real-world precipitation challenge, the code's usability relies on the availability of the ERA5-Land dataset. The repository could include a minimal, pre-processed subset of the data, or an automated script that downloads a small sample, so users can easily test the pipeline without needing to download and format massive climate datasets.*

**Response.** We thank the reviewer for examining the code repository. We have revised the implementation and documentation to address all of the requested reproducibility components. The core training and evaluation routines now contain inline comments linking the implementation to the corresponding equations and algorithms in the manuscript. The README also includes an equation-to-code map identifying the implementations of the supervised and $W_1$-regularized objectives, the tail-quantile approximation, the quantile-index update, and the inference-informed continual training algorithm.

We have further added a pinned software environment and complete setup instructions, including the tested Python, PyTorch, CUDA, and supporting-library versions. The README reports the tested hardware configuration, data dimensions, principal hyperparameters, and approximate runtime and peak GPU-memory requirements for the toy problems, ERA5-Land downscaling, Flow Matching training and sampling, and GEVD-prior downscaling. Finally, we have provided a minimal ERA5-Land reproducibility workflow, together with the associated preprocessing and smoke-test instructions, so that users can verify data loading, loss evaluation, and model execution without reproducing the complete full-scale dataset pipeline.

## Response to Reviewer #2

**General assessment.** *In this paper, an Extreme Event Aware (e2a or eta) or $\eta$-learning algorithm is developed. It does not assume the existence of extreme events in the available data. It reduces uncertainty even in "uncharted" extreme-event regions by enforcing the extreme-event statistics of an observable indicative of extremeness during training, which can be supported by qualitative arguments or estimated from unlabeled data.*

*Overall, the method is novel, and the paper is well-written. The topic is also a good fit for the scope of Nature Communications. That said, a few points are not very clear, and I believe they need to be explained and/or validated more systematically. I suggest a revision of this paper.*

**Response.** We thank the reviewer for the positive assessment and for identifying points where the method should be explained more systematically.

**Comment #1.** *It looks like $g$ is a very low-dimensional output function. In the paper, it says $g$ belongs to $\mathbb{R}$, which is one dimension. However, in practice, the observable may not always be a scalar. It can be an extreme pattern or a combination of several features. What is the scalability of the method?*

**Response.** We agree that certain applications may require more than a scalar observable. The present work formulates $g$ as scalar because the one-dimensional quantile representation of $W_1$ underlies both the efficient implementation and the tail-error optimality results. The framework can nevertheless be extended to multivariate observables in a computationally scalable manner. There are three distinct settings to consider.

If a vector of features $G = (g_1, \dots, g_r)$ admits a physically meaningful scalar severity score $s : \mathbb{R}^r \to \mathbb{R}$, one can take $g = s \circ G$ and apply the present scalar framework directly. This includes pattern-based events whenever a suitable scalar score is available.

When scalarization is not desired and calibration of the componentwise marginal laws is sufficient, a direct extension is

$$\sum_{j=1}^{r} W_1((g_j \circ \phi_\theta)_{\#}\mu, \nu_{0,j}) \,,$$

where $\nu_{0,j}$ denotes the reference law for the $j$-th observable. A detailed complexity analysis for the single-obsevable case is provided in newly added SI subsection Computational Complexity. We briefly restate the analysis here for assistance. Let $n_{\tilde{x}}$ be the number of auxiliary inputs used to estimate the push-forward distributions, $n_q$ the number of quantile levels per observable, and $C_\mathrm{f}$ and $C_\mathrm{b}$ the costs of one forward and backward model propagation, respectively. At each quantile-index refresh, the model is evaluated once on all $n_{\tilde{x}}$ auxiliary inputs, and these outputs are shared across the $r$ observables, giving a cost of $O(n_{\tilde{x}} C_\mathrm{f})$. Locating the quantile samples requires sorting $n_{\tilde{x}}$ values for each observable, with total cost $O(r n_{\tilde{x}} \log n_{\tilde{x}})$. The subsequent gradient computation uses at most $r n_q$ selected quantile samples and therefore costs $O(r n_q (C_\mathrm{f} + C_\mathrm{b}))$. Thus, under the worst-case assumption that the quantile indices are refreshed at every iteration, the componentwise Wasserstein regularization has total cost

$$O(n_{\tilde{x}} C_\mathrm{f} + r n_{\tilde{x}} \log n_{\tilde{x}} + r n_q (C_\mathrm{f} + C_\mathrm{b})) \,.$$

The extension therefore scales linearly with the number of observables $r$, while the full inference

pass over the auxiliary inputs is shared across all observables.

Third, if the joint dependence or correlation structure of $G$ is also important, then marginal matching is insufficient and a genuinely multivariate discrepancy is needed. Tools such as sliced Wasserstein distance and entropically regularized optimal transport can be integrated into the framework in place of the scalar $W_1$ term. Both have been widely adopted in machine learning as computationally tractable approaches to multivariate distribution comparison: sliced Wasserstein reduces the problem to one-dimensional projections, while entropic regularization enables efficient and parallelizable Sinkhorn iterations. We expect tail-enhanced versions of these discrepancies to be important when the goal is to correct tail bias in a multivariate extreme-event setting.

We have further emphasized the scalar-observable setup and pointed out the extensions to multivariate observables in the Problem Statement at line 168:

Throughout this work, we focus on scalar-valued observables; remarks on extensions to multivariate settings are provided under "Limitations and Future Work".

We have also added the following discussion to the newly added "Limitations and Future Work" under the section "Discussion" at line 1067:

Finally, the present theory and algorithms are developed for scalar observables, for which $W_1$ enables efficient computation and underpins the tail optimality results. For a vector observable $G = (g_1, \ldots, g_r)$, one may either introduce a physically meaningful severity function $s : \mathbb{R}^r \to \mathbb{R}$ and apply the present framework to $g = s \circ G$, or use a sum of componentwise $W_1$ penalties when matching the marginal laws is sufficient. The latter scales linearly with $r$, but does not constrain the correlation structure among the components. When the joint law of $G$ is essential, genuinely multivariate discrepancies based on, for example, sliced Wasserstein distance or entropically regularized optimal transport are required [1, 2]. Extending the present tail-focused theory and scalable algorithms to these settings is another important direction.

**Comment #2.** *The use of the Wasserstein Distance needs to be explained more systematically. While it is true that it serves as a useful statistical regularization, the advantages and disadvantages should be discussed. What will be the conditions that make this regularization particularly suitable for extreme events? What about K-L divergence and other statistical metrics? Even qualitative comparisons or discussions can be very helpful in supporting the use of the method proposed here.*

**Response.** We thank the reviewer for this helpful suggestion. We agree that the comparison with other divergences/metrics could be made more systematic. We have revised the Discussion subsection "Practical Advantages of $W_1$" to clarify the comparison specifically with KL divergence and the $L^1$-log metric, which are the most relevant alternatives for the present setting.

In particular, for the usual forward KL form

$$D_{\mathrm{KL}}(p\|q) = \int p(y) \log \frac{p(y)}{q(y)}\, dy,$$

tail discrepancies are weighted by the small probability density $p(y)$. Therefore, even large discrepancies in low-probability regions may contribute weakly to the objective. The $L^1$-log metric, defined as

$$\int_{\mathcal{Y}} |\log p(y) - \log q(y)|\, dy.$$

mitigates this issue by comparing log densities directly and therefore gives stronger emphasis to tail discrepancies. However, this advantage comes with a practical limitation: the metric requires

reliable density estimates and becomes numerically unstable when either density is zero or extremely small. In particular, when the supports of the learned observable distribution and the reference distribution do not fully overlap, the log-density discrepancy diverges. This issue is especially common in the data-scarce regime considered here, where the learned model may initially assign little or no probability mass to extreme values. In such cases, KL-based objectives and the $L^1$-log metric require additional numerical devices to remain well defined.

In contrast, $W_1$ avoids this support-mismatch obstruction: for probability measures with finite first moments, it gives a finite and stable discrepancy even when the learned and reference observable laws have poorly overlapping tails. A $W_1$-based regularizer thus provides a stable training signal in this regime. The practical usefulness of $W_1$ also comes from the fact that, in one dimension, $W_1$ admits the quantile representation

$$W_1(\nu_1, \nu_2) = \int_0^1 \left| F_{\nu_1}^{-1}(q) - F_{\nu_2}^{-1}(q) \right| \, dq, \tag{1}$$

which makes the discrepancy straightforward to estimate from samples. More importantly, this representation naturally allows the numerical approximation to be concentrated on the tail; see Methods for details. As a result, $W_1$ allows for a numerically robust mechanism for making the regularization tail-aware.

The revised subsection now reads as follows.

### Practical Advantages of $W_1$

Aside from its theoretical optimality for controlling tail errors, $W_1$ offers several practical advantages over alternatives such as KL divergence and the $L^1$-log metric [3]. The key issue with the usual forward KL divergence,

$$D_{\mathrm{KL}}(p\|q) = \int p(y) \log \frac{p(y)}{q(y)} \, dy,$$

is that discrepancies in the tail are weighted by $p(y)$, which is small precisely in rare-event regions. Consequently, even substantial disagreement between $p$ and $q$ in the tail can contribute weakly to the total KL objective. In the present setting, this is undesirable because the main goal to recover statistically meaningful behavior in low-probability extreme regions.

The $L^1$-log metric mitigates this issue by comparing log densities directly. It is therefore more explicitly tail-aware and could in principle serve as a regularizer for $\eta$-learning. However, this advantage comes with a practical limitation: the metric requires reliable density estimates and becomes numerically unstable when either density is zero or extremely small. In particular, when the supports of the learned observable distribution and the reference distribution do not fully overlap, the log-density discrepancy diverges. This issue is especially common in the data-scarce regime considered here, where the learned model may initially assign little or no probability mass to extreme values. In such cases, KL-based objectives and the $L^1$-log metric require additional numerical devices to remain well defined.

In contrast, $W_1$ avoids this support-mismatch obstruction: for probability measures with finite first moments, it gives a finite and stable discrepancy even when the learned and reference observable laws have poorly overlapping tails. A $W_1$-based regularizer thus provides a stable training signal in this regime. The practical usefulness of $W_1$ also comes from the fact that, in one dimension, $W_1$ admits the quantile representation

$$W_1(\nu_1, \nu_2) = \int_0^1 \left| F_{\nu_1}^{-1}(q) - F_{\nu_2}^{-1}(q) \right| \, dq, \tag{2}$$

which makes the discrepancy straightforward to estimate from samples. More importantly, this representation naturally allows the numerical approximation to be concentrated on the tail; see Methods for details. As a result, $W_1$ allows for a numerically robust mechanism for making the regularization tail-aware.

**Comment #3.** *Following the above question, the Wasserstein Distance is a statistical regularization, which is also the emphasis of this paper. However, how does the method, with such a statistical regularization, help learn the dynamics of extreme events? In principle, with only statistical information, the mechanisms of the extreme events generated can be very different from the truth. This is certainly not the case with the numerical results. Therefore, something behind the scenes must play some roles here. Adding more reasoning and even potential caveats will be very necessary.*

**Response.** We thank the reviewer for raising this important point. We agree that the $W_1$ penalty is a statistical regularization and, by itself, does not identify or learn the dynamical mechanism that generates extreme events.

In the precipitation example, the dynamical and structural information enters through the supervised component of the learning problem. The dataset contains low-resolution (LR) precipitation snapshots over the full period $t \in [0, T]$, with $T = 25$ years, together with paired high-resolution (HR) snapshots over the first 0.5 years. Each LR snapshot provides a coarse representation of the corresponding physical state, while the LR–HR pairs provide labeled spatial correspondence for learning the downscaling map. The $W_1$ term does not add new mechanistic information; rather, it constrains this supervised map so that its induced observable statistics match the prescribed tail law.

Therefore, the method should be interpreted as a distributionally constrained supervised-learning framework, not as a mechanism-discovering method based on statistical information alone. If distinct mechanisms induce the same scalar observable law, the present formulation cannot distinguish them without additional identifying information, such as temporal or spatial structures and physics-informed constraints. In fact, one promising future direction is to develop learning frameworks that jointly infer the governing mechanisms while enforcing consistency with the prescribed extreme-event statistics.

We have incorporated an explicity caveat in the result analysis for the vanilla precipitation downscaling example. Line 706 now reads:

It is also crucial to note that the dynamical structure in the generated extreme events does not come from the Wasserstein regularization alone. Rather, it is inherited from the supervised problem: each LR snapshot provides a coarse representation of the corresponding physical state, while the paired LR–HR snapshots over the first 0.5 years provide labeled spatial correspondence.

The limitation that $\eta$-Learning does not identify the dynamical mechanism of extreme events and the correpsonding future direction are discussed in detail in the newly added subsection "Limitations and Future Work" at line 1038:

A central limitation is that matching the distribution of a prescribed observable does not identify the dynamical mechanism that generates extreme events. The $W_1$ regularizer constrains observable statistics, but does not by itself determine the temporal evolution, spatial organization, or broader physical coherence of the generated extremes. To the extent that these properties are recovered, they are inherited from the supervised input-output data rather than supplied by the $W_1$ term.

Consequently, when distinct mechanisms or state configurations induce the same observable law, they are not distinguishable under the present objective. An important direction is therefore to extend the current framework with temporal and spatial constraints, conservation laws, or other physical priors.

**Comment #4.** *Is there a way to automatically choose the hyperparameters to make the algorithm more objective? Otherwise, some additional prior knowledge is needed, which can sometimes complicate the problem.*

**Response.** We thank the reviewer for raising this important point. In the present framework, there are two types of hyperparameters. Some parameters, such as the quantile set $\mathcal{Q}$ or the cutoff $\tau$, specify which part of the observable distribution is regarded as the tail region of interest. These necessarily encode the target rare-event probability level and are therefore tied to the scientific or engineering objective; in other words, they are more application-specific. The balancing parameter $\lambda$, however, plays a different role: it controls the relative strength of the supervised ERM loss and the $W_1$ tail correction. We agree with the reviewer that it will be very desirable for $\lambda$ not to require delicate problem-specific tuning.

To address this concern, we have added a more explicit discussion of the robustness of our method with respect to $\lambda$. As reported in SI Fig. 4, the learned observable distribution remains stable as $\lambda$ varies from $10^{-4}$ to 10, while all other training configurations are fixed. This indicates that the method is numerically robust to the choice of $\lambda$ over a wide range. Thus, at least in the tested settings, $\lambda$ does not need to be finely tuned in order for the method to recover the target tail behavior.

This robustness is closely tied to the effective training strategies used in the algorithm. In particular, we do not introduce the $W_1$ correction from an untrained model. Instead, we first train the model using only the supervised ERM loss, and then initialize both the $\eta$-learning stage and the initial tail set used to approximate the $W_1$ penalty from the pre-trained supervised estimator. After this step, the model-induced ordering used to select tail samples is anchored to a predictor that already captures the observed input-output relation. Consequently, the $W_1$ term primarily acts as a residual tail correction around an ERM-consistent solution, rather than determining the entire learned map from scratch. The periodic tail-set refresh further keeps the $W_1$ correction aligned with the evolving observable distribution. This is why the method is much less sensitive to the precise value of $\lambda$ than a direct, unstaged optimization of the combined ERM+$W_1$ objective would be.

Nevertheless, to make the selection of $\lambda$ more objective in practice, we have added a gradient-scale balancing rule. Since

$$\nabla_\theta L(\theta) = \nabla_\theta L_{\mathrm{ERM}}(\theta) + \lambda \nabla_\theta L_{W_1}(\theta),$$

a potentially beneficial choice is to select $\lambda$ so that the scaled $W_1$ gradient has comparable Euclidean norm to the ERM gradient, namely

$$\lambda \|\nabla_\theta L_{W_1}(\theta)\|_2 = \|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2 \implies \lambda = \frac{\|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2}{\|\nabla_\theta L_{W_1}(\theta)\|_2 + \varepsilon},$$

with a small $\varepsilon > 0$. This provides an objective scale for $\lambda$, while the robustness study shows that the method does not require delicate tuning once the staged training procedure is used.

We have revised the main text to make these points explicit. In particular, we now explain in

subsection "Effective Training Strategies" that $\lambda$ is a secondary balancing parameter, while ERM pre-training and tail-set refresh are the primary mechanisms responsible for stable optimization:

### Effective Training Strategies

Accurately estimating tail quantiles at every iteration can be memory- and compute-intensive, because it naively requires evaluating the model on a large auxiliary sample set $\tilde{x} \sim \mu$ at each update. To mitigate this, we use an inference-informed continual training strategy: periodically, we run inference on the auxiliary set to identify the inputs whose current observable values realize the target quantiles in $\mathcal{Q}$, and then restrict subsequent $W_1$ estimations to these identified inputs until the next refresh. This preserves a reliable estimate of the tail loss while substantially reducing per-iteration cost. A detailed computational complexity analysis of the full procedure is provided in SI Subsection "Computational Complexity".

This tail-set selection procedure also creates a potential optimization issue. The $W_1$ term is evaluated on inputs selected according to the current model-induced observable distribution $(g \circ \phi_\theta)_{\#}\mu$. Early in training, before the supervised input-output relation has been learned, this model-induced ordering can be unreliable. As a result, samples that should correspond to bulk behavior under an ERM-consistent predictor can be spuriously selected as tail samples and pulled toward upper quantiles of $\nu_0$, while the ERM loss pulls the model toward the labeled pairs. The two loss components therefore provide conflicting gradient signals, leading to oscillatory or unstable optimization.

We mitigate this issue using a staged training procedure. We first train the model using only the supervised ERM loss, and then use the ERM estimator to initialize both the subsequent $\eta$-learning stage as well as the initial tail set used to approximate the $W_1$ penalty. After this pre-training step, the ordering used to select tail samples is anchored to a predictor that already respects the observed input-output relation. Consequently, the $W_1$ penalty primarily acts as a tail correction to an ERM-consistent map, while the ERM term continues to stabilize the learned map around the supervised solution. During the subsequent $\eta$-learning iterations, the tail set is periodically refreshed as the learned map evolves, keeping the samples used to estimate $W_1$ separated from the ERM-induced bulk. This procedure systematically mitigates the conflict caused by unreliable tail-sample selection and has empirically yielded consistently stable and robust training.

The balancing parameter $\lambda$, on the other hand, controls the strength of this residual tail correction. In practice, choosing $\lambda$ to make the $W_1$ gradient comparable in magnitude to the ERM gradient can be beneficial. However, $\lambda$ only rescales the $W_1$ contribution; it does not by itself resolve the directional incompatibility that can arise from unreliable early-stage tail-set selection. Thus, the main stabilization mechanism is ERM pre-training together with tail-set refresh, while $\lambda$ serves as a secondary balancing parameter. This interpretation aligns with the robustness study in SI Fig. 4, where $y_{\eta\#}\mu$ remains stable over a broad range of $\lambda$. A more detailed discussion of the gradient conflict and the practical choice of $\lambda$ is provided in SI Subsection "Pre-training and Inference-Informed Continual Training".

We have also revised the SI subsection "Pre-training and Inference-Informed Continual Training" to give a more detailed account of the gradient conflict, the role of tail-set selection, and the practical choice of $\lambda$:

### Pre-training and Inference-Informed Continual Training

Several difficulties arise in the actual optimization of (14). We assume $\mathcal{F}_\theta$ is a neural network optimized with a stochastic first-order optimization method $\mathcal{M}$. The update rule is denoted by

$$\phi_\theta^{(k+1)} \leftarrow \mathcal{M}\left(\phi_\theta^{(k)}, \nabla_\theta \mathcal{L}_\theta^{(k)}\right).$$

Our focus remains on (14), as the other formulation can be viewed as a special case.

***Memory Challenges and Tail-Set Selection in $W_1$ Estimation.*** A practical challenge stems from approximating the $W_1$ term, even with its tractable one-dimensional quantile representation. Accurate estimation requires evaluating $g \circ \phi_\theta$ over the auxiliary set $\mathcal{D}_{\tilde{x}}$, which is typically much larger than the labeled training set. If this full auxiliary evaluation were included in every gradient-carrying update, the memory cost would be substantial.

To address this, we note that only $n_q$ specific inputs from $\mathcal{D}_{\tilde{x}}$ are ultimately needed to compute the empirical quantile loss, namely the inputs whose current observable values realize the quantile levels in $\mathcal{Q}$. This motivates the Inference-Informed Continual Training (IICT) procedure. Periodically, we run inference on the full auxiliary set $\mathcal{D}_{\tilde{x}}$ and compute

$$\{g(\phi_\theta(\tilde{x}_j))\}_{j=1}^{n_{\tilde{x}}}.$$

We then identify the indices of the inputs corresponding to the prescribed quantile levels in $\mathcal{Q}$ and store them in a set $\mathcal{I}$. Between refreshes, the $W_1$ loss is evaluated only on the $\mathcal{I}$-indexed inputs. Thus, full inference on $\mathcal{D}_{\tilde{x}}$ is used only to identify the relevant tail samples, while the gradient-carrying $W_1$ update is restricted to the selected quantile samples. The refresh frequency is controlled by a hyperparameter $\omega$, with $\mathcal{I}$ updated every $\omega$ iterations. This procedure substantially reduces memory usage while preserving an accurate estimate of the tail contribution to the $W_1$ loss.

***Gradient Conflict and ERM Pre-training.*** The tail-set selection procedure also explains a potential source of optimization instability. The $W_1$ loss is evaluated on inputs selected according to the current model-induced observable distribution $(g \circ \phi_\theta)_{\#}\mu$. Early in training, before the supervised input-output relation has been learned, the ordering of the auxiliary observable values $g(\phi_\theta(\tilde{x}_j))$ can be unreliable. Consequently, inputs that should correspond to bulk behavior under an ERM-consistent predictor may be spuriously selected into $\mathcal{I}$ and treated as high-quantile samples. The $W_1$ term then pulls these samples toward upper quantiles of the reference distribution $\nu_0$, while the ERM loss pulls the model toward the labeled input-output relation. These two signals can be incompatible, leading to oscillatory or unstable optimization.

Our main stabilization mechanism is ERM pre-training. We first minimize the ERM loss alone and initialize $\phi_\theta$ with the resulting ERM estimator $u_\xi$ when solving (14). The initial index set $\mathcal{I}$ used by the $W_1$ penalty is also computed from this ERM-pretrained model. Thus, the first $W_1$ correction is applied to quantile samples selected under a predictor that already captures the observed input-output relation. This substantially reduces the chance that bulk samples are spuriously treated as tail samples because of unreliable early predictions. During $\eta$-learning, $\mathcal{I}$ is periodically refreshed as $\phi_\theta$ evolves, keeping the $W_1$ correction aligned with the current observable distribution.

***Choice of $\lambda$.*** The balancing parameter $\lambda$ controls the relative magnitude of the ERM and $W_1$ gradient components. Writing

$$\mathcal{L}_\theta = L_{\mathrm{ERM}}(\theta) + \lambda L_{W_1}(\theta),$$

we have

$$\nabla_\theta \mathcal{L}_\theta = \nabla_\theta L_{\mathrm{ERM}}(\theta) + \lambda \nabla_\theta L_{W_1}(\theta).$$

A practical scale for $\lambda$ is obtained by choosing it so that the scaled $W_1$ gradient has a comparable Euclidean norm to the ERM gradient during the early $\eta$-learning iterations:

$$\lambda \|\nabla_\theta L_{W_1}(\theta)\|_2 = \|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2 \implies \lambda = \frac{\|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2}{\|\nabla_\theta L_{W_1}(\theta)\|_2 + \varepsilon},$$

with a small $\varepsilon > 0$ for numerical stability. This prevents the $W_1$ correction from overwhelming the supervised anchor.

However, $\lambda$ only rescales the $W_1$ gradient; it does not change its direction. Therefore, while magnitude balancing could be potentially beneficial, it alone cannot resolve the directional incompatibility caused by unreliable tail-set selection. For this reason, ERM pre-training and periodic tail-set refresh are the primary mechanisms for stable optimization, while $\lambda$ serves as a secondary balancing parameter that controls the strength of the residual tail correction. This interpretation is consistent with the robustness experiment in SI Fig. 4, where the learned observable distribution remains stable as $\lambda$ varies from $10^{-4}$ to 10 while all other training configurations are fixed.

**Comment #5 .** *One thing I feel very confused about is the toy model examples. I do see you get extreme events. However, the locations and amplitudes differ from the truth, e.g. in Fig 1. While this can still give you the right spatiotemporal statistics, I wonder what the target is here. I guess, in addition to sampling extreme events across the domain, in SI Fig 2, what is more important is to get the point-wise statistics, not the spatiotemporal average. It is also practically useful to predict the locations and times of extreme events. This relates to my previous question about the dynamical mechanisms that generate extreme events.*

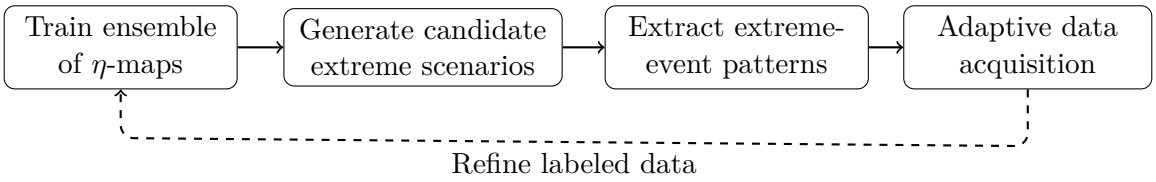
**Response.** We thank the reviewer for this question and agree that this point should be further clarified. The target of the toy examples, and of $\eta$-Learning in general, is not exact recovery of the *unobserved* extreme event location, amplitude, or time from training data that contain no such events. Such recovery cannot be expected from conventional supervised methods without extreme-event data, as shown in Theorem 4, and is non-identifiable from the scalar observable distribution alone.[1] Rather, the target is to learn maps that remain accurate in the data-rich region while matching the prescribed tail law, thereby generating *plausible candidate extremes* that are statistically consistent with the observable tail.

This distinction is important for interpreting Fig. 1 and SI Fig. 2. In these examples, the different peak locations obtained across independent $\eta$-maps should not be interpreted as failures to recover a hidden truth. They instead represent the residual spatial uncertainty that remains after enforcing the scalar tail statistics. If point-wise statistics are the intended target, then the corresponding point-wise or spatially conditioned observables should be included in the statistical constraint. With only a scalar extreme observable, $\eta$-Learning constrains the tail behavior of that observable, but it does not uniquely determine the full spatial arrangement of the corresponding extreme field.

One promising downstream use of the generated scenarios is precisely to expose this residual uncertainty in a systematic and algorithmic way. In applications where additional simulations or experiments can be performed, one can train an ensemble of $\eta$-maps, generate candidate extreme

[1] An exception is when Assumption 5 holds, which then would imply Theorem 6. However, enforcing Assumption 5 will likely require more information than the observable distribution. See line 443 for more discussion on this point.

scenarios, extract common patterns according to extreme-event time, location, or other physically meaningful features, and then use the resulting representatives to guide adaptive data acquisition. The pipeline is summarized in the following illustration.




The role of $\eta$-Learning in this workflow is therefore not to replace mechanistic validation, but to *propose* statistically consistent extreme-event hypotheses in regions where the original data provide no direct evidence. Subsequent high-fidelity simulations, experiments, or additional observations can then test, refine, or reject these candidate modes.

We have revised the result analysis for the 2D-to-1D toy example in the main text at line 595 to read:

The ensemble analysis highlights a broader implication of $\eta$-learning: enforcing scalar observable tail statistics generates candidate extremes, but does not necessarily identify the underlying physical structure, such as the peak location shown in this example. This variability reflects the residual physical uncertainty that remains after enforcing the observable tail statistics, rather than a failure to recover a uniquely identifiable hidden event. [...] In the Discussion section, we further describe how this uncertainty can be systematically characterized and used in downstream applications.

Moreover, we have added a detailed account of the downstream algorithmic pipeline outlined above in the newly added subsection "Limitations and Future Work" at line 1048:

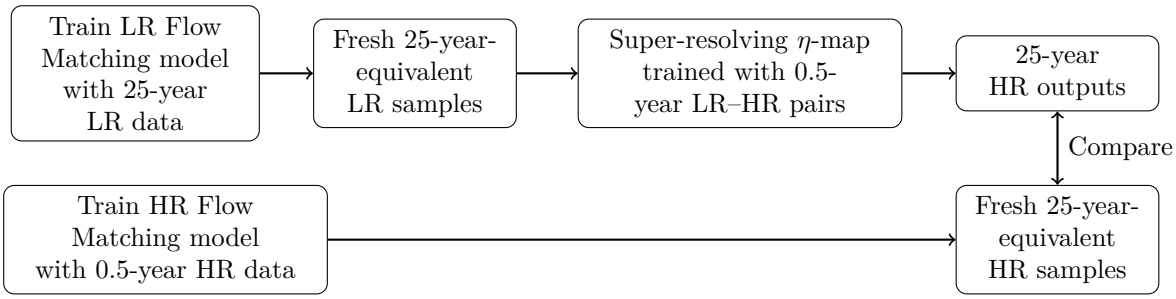
This non-identifiability also arises at the level of individual events: the prescribed observable law does not generally determine the precise location, occurrence time, or full spatial structure of an unobserved extreme. Variability in these features across independently trained $\eta$-maps should therefore be interpreted as residual uncertainty. Importantly, this uncertainty can be exploited systematically. Ensembles of $\eta$-maps can generate diverse yet statistically consistent candidate extremes, which can then be organized according to physically meaningful attributes such as event location, occurrence time, or spatial pattern. Representative scenarios can guide subsequent rounds of adaptive data acquisition, such as targeted high-fidelity simulations or experiments, allowing the corresponding hypotheses to be tested, refined, or rejected while progressively reducing uncertainty in data-scarce regions.

**Comment #6 .** *I also noticed that, for example, in Figure 6 and maybe Figure 5 as well, the estimated tail is above the true tail. Is this because of the over-regularization? Is there a general criterion to mitigate this, or does it need a few trials to get the optimal solution?*

**Response.** We thank the reviewer for raising this concern. We clarify two distinct cases.

For Fig. 5, the curve labeled "HR $\eta$-Generated Samples (0.5-Yr Training)" is produced by a two-stage generative pipeline: we train an LR Flow Matching model using 25 years of LR data, draw a fresh 25-year set of LR samples from the trained model, pass those LR samples through the super-resolving $\eta$-map trained with 0.5 years of LR–HR pairs (i.e., the map that produces Fig. 4), and compute the observable distribution of the resulting HR outputs. The workflow is summarized

schematically as



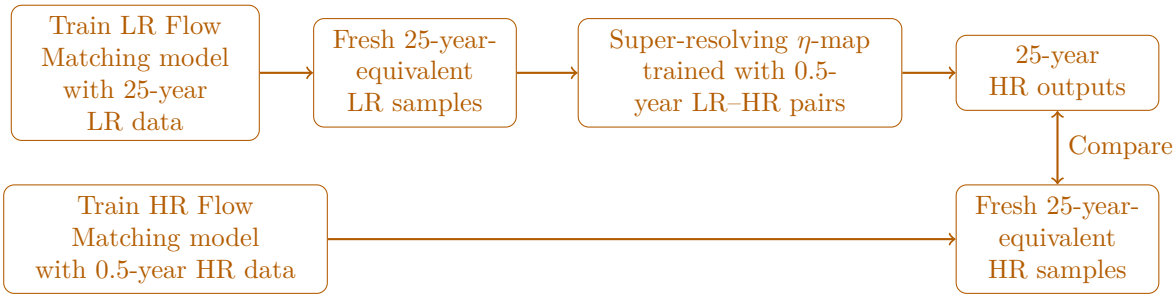

Thus, the star curve in Fig. 5 is not obtained by applying $u_\eta$ to the fixed empirical 25-year LR dataset. It is the observable distribution obtained from a finite set of samples generated by the LR Flow Matching model and subsequently post-processed by $u_\eta$. Therefore, the apparent tail elevation in Fig. 5 should be interpreted as uncertainty propagated through the low-resolution generative pipeline instead of over-regularization of $\eta$-learning.

We have added the following clarification in the main text at line 827:

The tail elevation relative to the ground truth reflects uncertainty from drawing a fresh finite 25-year LR sample from the LR FM model, rather than direct evidence of over-regularization of the $\eta$-map.

To faciliate the exposition, we have also added the schematic description to the SI in section "Experimental Setups and Details" under subsection "Correcting Statistical Biases of Deep Generative Models":

**Experimental Setup.** The experimental setup is summarized schematically below. We first train a low-resolution (LR) Flow Matching (FM) model using the full 25-year LR dataset. We then draw a fresh 25-year-equivalent collection of LR samples from this trained LR FM model and pass these samples through the super-resolving $\eta$-map trained with only 0.5 years of paired LR–HR data. As a baseline, we also train a high-resolution (HR) FM model using only the available 0.5-year HR dataset and draw a fresh 25-year-equivalent collection of HR samples from it. The observable distributions of these generated samples are then compared against the empirical 25-year HR observable distribution.




This workflow is also referenced in the main text at line 816:

A schematic description of this workflow is provided in the SI.

For Fig. 6, the interpretation is different. In the hypothesized GEVD experiment, the prescribed reference law $\nu_0$ is intentionally chosen to be heavier-tailed than the empirical 25-year HR observable

distribution. Therefore, the resulting heavier-tailed $y_{\eta\#}\mu$ is expected: the $\eta$-map is designed to be faithful to the prescribed reference law. This experiment serves as a prior-conditioned stress test rather than as an estimate of the true climatological observable distribution. Practically, this setting can represent, for example, an hypothetical adverse climate-tail scenario in climate change, where the goal is to generate statistically consistent extreme precipitation events for downstream analysis under a specified tail hypothesis.

We have revised the clarification in the main text at line 913:

We emphasize that here, $\nu_0$ is a deliberately heavier-tailed reference law used to explore what HR precipitation fields are statistically consistent with the paired LR-HR training data under a specified adverse-tail hypothesis.

**Comment #7 .** *Did you consider the case with observational noise? Can you filter the noise at the same time, so you will not treat the noisy observation as the truth when applying your method?*

**Response.** We thank the reviewer for raising this point. The present formulation is a supervised-learning framework rather than a filtering or data-assimilation framework. In particular, the setup does not assume an explicit latent dynamical model, observation operator, or measurement-noise model. Therefore, Kalman-type filters, particle filters, smoothing methods, and related data-assimilation procedures are not generic components of the proposed algorithm. If such additional structure is available in a particular application, these methods, or bias-correction procedures more generally, can be used as a *data-preprocessing* step to construct denoised training targets before applying $\eta$-Learning. The precipitation example is consistent with this viewpoint: we use reanalysis precipitation fields as the data source, so bias correction belongs to the construction of the training data rather than to $\eta$-Learning itself.

We have clarified the role of observational noise in Problem Statement at line 195:

We study the data-driven setting where we have access to a dataset $\mathcal{D} = \{(x_i, u_i, y_i)\}_{i=1}^n$, where $u_i = u(x_i)$ and $y_i = g(u_i)$. When the raw measurements used to construct $u_i$ may be noisy, $u_i$ should be understood as the processed training target obtained after any application-specific data-assimilation or bias-correction step. Such preprocessing requires additional assumptions, such as a measurement-noise model or an observation operator, and is therefore not a generic component of our setup.

A separate issue, more specific to $\eta$-learning, is the noise or misspecification in the reference law $\nu_0$. Since the $W_1$ term is designed to enforce agreement with the supplied $\nu_0$, distributional uncertainty in $\nu_0$, especially in the tail, can be transferred to the generated extremes. We have therefore added a sensitivity analysis for shifted reference laws and discussed distributionally robust extensions in which a single reference law is replaced by an ambiguity set of plausible tail laws in 'Limitations and Future Work". It reads as follows.

Another limitation is the reliance on an accurate reference law $\nu_0$. Because the tail-emphasized $W_1$ estimation is designed to match the supplied tail, misspecification of $\nu_0$, or a shift in the target tail relative to $\nu_0$, can be transferred directly to the generated extremes. A promising extension is a distributionally robust formulation that replaces a single reference law with an ambiguity set of plausible laws, with particular emphasis on uncertainty and shifts in the tail. Worst-case or uncertainty-weighted tail objectives could reduce overcommitment to any single hypothesized distribution while preserving awareness of extreme behavior.

## References


[1] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. doi:10.1007/s10851-014-0506-3.

[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

[3] Mustafa A. Mohamad and Themistoklis P. Sapsis. Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 115(44):11138–11143, 2018. doi:10.1073/pnas.1813263115.

[4] D. A. Levin and Y. Peres. *Markov Chains and Mixing Times*. 2nd ed. American Mathematical Society, Providence, RI, 2017. With contributions by E. L. Wilmer. doi:10.1090/mbk/107.

[5] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wissenschaften, Vol. 260. Springer, Berlin, Heidelberg, 3rd revised and enlarged edition, 2012. doi:10.1007/978-3-642-25847-3.

# Final Point-to-Point Responses to Reviewers*

Kai Chang and Themistoklis P. Sapsis

We sincerely thank the reviewers and co-reviewers for their careful assessment of our manuscript and for their constructive feedback throughout the review process.

## Response to Reviewer #1

**Remarks to the Author.** *The authors have thoroughly addressed all the concerns raised, so I recommend the manuscript for acceptance in its current form.*

**Response.** We sincerely thank the reviewer for their careful evaluation and for recommending acceptance of our manuscript.

## Responses to Reviewer #2

**Remarks to the Author.** *The authors have addressed all the questions I asked. In particular, the authors provided a very comprehensive response letter. I do not have any further concerns. I also want to remark that I co-reviewed this paper with an early-career scientist. Thank you for this great work.*

**Response.** We sincerely thank the reviewer and the early-career co-reviewer for their careful assessment, thoughtful comments, and kind words about our work and response letter.

**Remarks on code availability.** *The authors have improved the code, which is now clear.*

**Response.** We thank the reviewer for confirming that the revised code is now clear.

## Response to Reviewer #3

**Remarks to the Author.** *I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.*

**Response.** We sincerely thank the reviewer and co-reviewer for their time and contribution to the review process. We also appreciate the Nature Communications initiative to support and recognize early-career researchers in peer review.

*The reviewers' original comments are copied and pasted in italicized text, with an identifier provided before each quote.

# Point-to-Point Responses to Reviewers*

Kai Chang and Themistoklis P. Sapsis

## General Edits

In addition to the revisions made in direct response to the reviewers' comments, we have made several additional edits that we believe improve the clarity, organization, and readability of the manuscript. We summarize these changes here for completeness and will list the specific edits below.

1. We changed the theorem numbering in the SI from, for example, Theorem 1 to Theorem S1 to improve readability.
2. We changed the paragraph headings "Description of the results" in the Results section to "Analysis of the results" to more accurately reflect that these paragraphs interpret the numerical results rather than merely describe them.
3. The second data-consistency condition (Definition 3, eq. (9)) is relaxed from $\hat{C} < 1$ to $\hat{C} < \infty$:

$$\hat{C} := \inf_{C \geq 0} \left\{ \int_{\mathcal{X} \setminus E} |y - y_\xi| \, \mathrm{d}\mu \leq C \cdot \int_E |y - y_\xi| \, \mathrm{d}\mu \right\} < \infty.$$

   The proof strategies for the theorems remain the same.

## Responses to Reviewer #1

**General assessment.** *The authors present an innovative and mathematically elegant optimization framework designed to predict statistically consistent extreme events in zero-extreme-data regimes. This work addresses a critical bottleneck in computational physics, climate science, and engineering. Furthermore, the proposed integration of optimal transport with deep generative models constitutes a practical contribution to the field of rare event modeling. However, before the manuscript can be recommended for publication, several significant theoretical fragilities and limitations concerning spatial degradation and prior misspecification should be addressed. Therefore, I recommend acceptance subject to major revisions, as detailed in the report below.*

**Response.** We thank the reviewer for the positive assessment of the proposed framework and for identifying important points where the scope and limitations should be clarified.

*The reviewers' original comments are copied and pasted in italicized texts, with an identifier provided before each quote. Edits to the manuscript and SI are shown in this color.

**Manuscript summary.** *The authors introduce Extreme Event Aware ($\eta$-) Learning, a machine learning optimization framework designed to predict statistically consistent and physically plausible rare events, even when the training dataset completely lacks extreme examples. The framework incorporates a priori statistical information about an extreme observable into the training process using Optimal Transport. Specifically, the authors rigorously quantify the conditions under which standard Empirical Risk Minimization (ERM) fails in data-scarce regimes, and they resolve this by augmenting the loss function with a 1-Wasserstein ($W_1$) distance penalty. The method demonstrates significant performance in prototype problems and real-world precipitation downscaling tasks.*

**Response.** We appreciate this accurate summary.

**Research relevance.** *This work is highly relevant to computational physics, climate science, and engineering, where simulating extreme events is computationally prohibitive and real-world extreme data is naturally scarce. Compared to established literature, such as active learning or large deviation theory, which typically require partial access to governing equations or a small seed of extreme samples, $\eta$-learning operates in the zero-extreme-data regime. Relying on a predefined reference distribution ($\nu_0$) to guide the model is an innovative approach that constitutes a strong contribution to rare event modeling.*

**Response.** We thank the reviewer for recognizing the intended contribution.

**Methodology comment #i.** *The Problem Statement relies heavily on advanced measure theory and topological terminology, e.g. $\sigma$-algebras and Polish spaces, without accompanying physical intuition. To better approach applied scientists who would benefit most from this framework, the authors could include a plain-English, high-level summary of the pipeline (Input $\to$ Complex State $\to$ 1D Observable) at the beginning of this section.*

**Response.** We agree. We have added a plain-language paragraph at the beginning of the Problem Statement section at line 134. To avoid repetitions, we have also shortened the original problem formulation. The revised opening of the section reads as follows.

At a high level, the setting considered here consists of three objects. First, an input $x$ describes the uncertain physical driver of a system, such as an initial condition, forcing parameter, or low-fidelity data such as low-resolution fields. Second, the system maps this input to a high-dimensional state $u(x)$, such as a dynamic trajectory or high-fidelity data. Third, an application-dependent observable $g(u(x))$ summarizes the state into a scalar quantity whose large values signify the extreme event of interest. Thus, the pipeline is

$$x \longmapsto u(x) \longmapsto g(u(x)).$$

The learning task is to approximate either the full state map $u$ or the composite response $g \circ u$, even when the available training data contain few or no extreme events.

To formalize this setting, let $X : (\Omega, \Sigma, P) \to (\mathcal{X}, \mathcal{B}(\mathcal{X}))$ be an input random variable, where $\mathcal{X} \subset \mathbb{R}^d$ is equipped with its relative Borel $\sigma$-algebra. We denote the law of $X$ by $\mu$, namely

$$\mu(A) = P(X^{-1}(A)), \qquad A \in \mathcal{B}(\mathcal{X}).$$

We write $X \sim \mu$, and use $x \sim \mu$ to denote a realization of the random input. Thus, $\mu$ describes the distribution of inputs encountered by the system. We assume that $\mu$ is either simple to sample from, such as a Gaussian distribution, or sufficiently well represented by existing samples.

Define the state map

$$u : \mathcal{X} \to \mathcal{U} \subset \mathbb{R}^m, \qquad x \mapsto u(x),$$

where $\mathcal{U}$ is the state space. For each input realization $x$, $u(x)$ denotes the corresponding system state. In the settings of interest, accurate evaluations of $u$ are computationally expensive, and available observations of $u$ may be incomplete.

Next, define the pre-determined observable

$$g : \mathcal{U} \to \mathcal{Y} \subset \mathbb{R}, \qquad u \mapsto g(u).$$

This map provides a scalarization of the state space used in the subsequent distributional formulation. Examples include a spatial maximum, a spatial average, an energy functional, or another physically meaningful summary of the state. Throughout this work, we focus on scalar-valued observables; remarks on extensions to multivariate settings are provided under "Limitations and Future Work".

Finally, let

$$y := g \circ u, \qquad Y := y(X).$$

Assuming $y$ is $\mathcal{B}(\mathcal{X})$-measurable with respect to $\mathcal{B}(\mathcal{Y})$, the scalar response $Y$ is a random variable with law given by the push-forward measure $y_{\#}\mu$:

$$y_{\#}\mu(B) = \mu(y^{-1}(B)), \qquad B \in \mathcal{B}(\mathcal{Y}),$$

where $y^{-1}(B) = \{x \in \mathcal{X} : y(x) \in B\}$. We assume that $y_{\#}\mu$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, so that its probability density function exists.

**Methodology comment #ii.** *The Problem Statement introduces two restrictive assumptions. First, it assumes independent and identically distributed (IID) samples, discarding the temporal trajectories and strong correlations that characterize complex dynamical systems, e.g. climate dynamics. Second, it defines extremes via a static scalar threshold, excluding compound or duration-based extremes. The authors should explicitly acknowledge these limitations and discuss whether the framework could be adapted for non-IID data or alternative extreme definitions.*

**Response.** We thank the reviewer for pointing out these limitations. We agree that the IID assumption does not represent temporally correlated trajectory data. In the present theory, however, independence is used only to obtain an explicit expression for the probability that the training set does not intersect the extreme set $E$. We have clarified this point in the Problem Statement and added the corresponding extension to dependent data in the SI. Specifically, we revised the Problem Statement at line 194 as follows:

We study the data-driven setting where we have access to a dataset $\mathcal{D} = \{(x_i, u_i, y_i)\}_{i=1}^n$, where $u_i = u(x_i)$ and $y_i = g(u_i)$. [...] For the theoretical results below, the inputs $x_1, \ldots, x_n$ are assumed to be independent and identically distributed according to $\mu$. The extension to dependent data is provided in the Supplementary Information (SI) section "Theoretical Extensions".

We also added subsection "Extension to Non-IID Data" under the section "Theoretical Extensions" in the SI:

## Extension to Non-IID Data

The IID assumption in Theorem 4 of the main text is used only to evaluate the probability that the training set does not intersect the extreme set $E$. Let $X_1, \ldots, X_n$ be a possibly dependent sequence

with common marginal law $\mu$, and define

$$q_n(E) := \mathbb{P}\{X_i \notin E,\ i = 1, \dots, n\}.$$

For data collected along a temporal trajectory, this probability can equivalently be written as

$$q_n(E) = \mathbb{P}\{\tau_E > n\}, \qquad \tau_E := \inf\{i \geq 1 : X_i \in E\}.$$

Since $\mu(E) \leq \delta$, the union bound gives the dependence-independent estimate

$$q_n(E) \geq 1 - \sum_{i=1}^{n} \mathbb{P}\{X_i \in E\} \geq 1 - n\delta.$$

Thus, $q_n(E) \geq p$ whenever

$$n \leq \frac{1-p}{\delta}.$$

Sharper estimates of $q_n(E)$ may be obtained under additional assumptions on the temporal process. For instance, when $(X_i)$ is a stationary Markov chain with invariant law $\mu$, $q_n(E) = \mathbb{P}_\mu(\tau_E > n)$ can be analyzed as the survival probability of the chain killed upon entering $E$, with mixing-time, spectral-gap, coupling, or hitting-large-set estimates yielding bounds in terms of $\mu(E)$ and the relaxation scale of the chain. For trajectories generated by small random perturbations of deterministic dynamics, related hitting- and exit-time estimates may also be derived from the Freidlin-Wentzell theory of randomly perturbed dynamical systems [4, 5].

On the event that the training set does not intersect $E$, the remainder of the proof of Theorem 4 is unchanged. Consequently, its lower bound holds with probability at least $q_n(E)$.

We also agree that certain applications may require more than a scalar observable. It is adopted because the one-dimensional quantile representation of $W_1$ underlies both the efficient implementation and the tail-error optimality results. The framework can nevertheless be extended to multivariate observables in a computationally scalable manner. There are three distinct settings to consider.

If a vector of features $G = (g_1, \dots, g_r)$ admits a physically meaningful scalar severity score $s : \mathbb{R}^r \to \mathbb{R}$, one can take $g = s \circ G$ and apply the present scalar framework directly. This includes pattern-based events whenever a suitable scalar score is available.

When scalarization is not desired but calibration of the componentwise marginal laws is sufficient, a direct extension is

$$\sum_{j=1}^{r} W_1((g_j \circ \phi_\theta)_\# \mu, \nu_{0,j}),$$

where $\nu_{0,j}$ denotes the reference law for the $j$-th observable. A detailed complexity analysis for the single-obsevable case is provided in newly added SI subsection Computational Complexity (also see the response to the relevant comment below). We briefly restate the analysis here for assistance. Let $n_{\tilde{x}}$ be the number of auxiliary inputs used to estimate the push-forward distributions, $n_q$ the number of quantile levels per observable, and $C_\mathrm{f}$ and $C_\mathrm{b}$ the costs of one forward and backward model propagation, respectively. At each quantile-index refresh, the model is evaluated once on all $n_{\tilde{x}}$ auxiliary inputs, and these outputs are shared across the $r$ observables, giving a cost of $O(n_{\tilde{x}} C_\mathrm{f})$. Locating the quantile samples requires sorting $n_{\tilde{x}}$ values for each observable, with total cost $O(r n_{\tilde{x}} \log n_{\tilde{x}})$. The subsequent gradient computation uses at most $r n_q$ selected quantile samples and therefore costs $O(r n_q (C_\mathrm{f} + C_\mathrm{b}))$. Thus, under the worst-case assumption that the quantile

indices are refreshed at every iteration, the componentwise Wasserstein regularization has total cost

$$O(n_{\tilde{x}} C_{\mathrm{f}} + r n_{\tilde{x}} \log n_{\tilde{x}} + r n_q (C_{\mathrm{f}} + C_{\mathrm{b}})) .$$

The extension therefore scales linearly with the number of observables $r$, while the full inference pass over the auxiliary inputs is shared across all observables.

Third, if the joint dependence or correlation structure of $G$ is also important, then marginal matching is insufficient and a genuinely multivariate discrepancy is needed. Tools such as sliced Wasserstein distance and entropically regularized optimal transport can be integrated into the framework in place of the scalar $W_1$ term. Both have been widely adopted in machine learning as computationally tractable approaches to multivariate distribution comparison: sliced Wasserstein reduces the problem to one-dimensional projections, while entropic regularization enables efficient and parallelizable Sinkhorn iterations. We expect tail-enhanced versions of these discrepancies to be important when the goal is to correct tail bias in a multivariate extreme-event setting.

To clarify the scope of the formulation, we added the following sentence in the Problem Statement at line 168:

Throughout this work, we focus on scalar-valued observables; remarks on extensions to multivariate settings are provided under "Limitations and Future Work".

We also added the following discussion to the newly added "Limitations and Future Work" subsection at line 1067:

Finally, the present theory and algorithms are developed for scalar observables, for which $W_1$ enables efficient computation and underpins the tail optimality results. For a vector observable $G = (g_1, \ldots, g_r)$, one may either introduce a physically meaningful severity function $s : \mathbb{R}^r \to \mathbb{R}$ and apply the present framework to $g = s \circ G$, or use a sum of componentwise $W_1$ penalties when matching the marginal laws is sufficient. The latter scales linearly with $r$, but does not constrain the correlation structure among the components. When the joint law of $G$ is essential, genuinely multivariate discrepancies based on, for example, sliced Wasserstein distance or entropically regularized optimal transport are required [1, 2]. Extending the present tail-focused theory and scalable algorithms to these settings is another important direction.

**Methodology comment #iii.** *The proof for Theorem 4 relies on bounding parameters $\widetilde{C}$ and $\widehat{C}$, which the authors admit lie "beyond the scope of this work" to compute. To strengthen the practical applicability of this definition, the authors should provide at least one concrete analytical example of a simple function class where these constants can be explicitly bounded.*

**Response.** We thank the reviewer for this comment. We agree that an explicit analytical example would make the data-consistency condition more transparent and practically useful. We have added to the SI an extra section An Analytic Example for Data-Consistency, and have referenced to it in the main text at line 955:

While it is beyond the scope of this work to characterize these constants for general approximation classes, they can be computed explicitly in simple settings. We provide such an example in the SI.

The added section reads as follows.

## An Analytic Example for Data-Consistency

To strengthen the practical applicability of the data-consistent condition introduced in Definition 3 of the main text, we provide a simple, interpretable example where the constants $\widetilde{C}$ and $\widehat{C}$ can be explicitly bounded and the bound directly depends on the smoothness of the true composite map, the approximation class, and the extreme set, as introduced in the condition.

Let $\mathcal{X} = [0,1]$, let $\mu$ be the uniform measure on $[0,1]$ and $E = [1-\delta, 1]$. Consider the affine approximation class

$$\mathcal{F} = \{x \mapsto ax + b : \ a, b \in \mathbb{R}\}$$

and the true map

$$y(x) = a_\star x + b_\star + M\psi_r(x),$$

where $a_\star, b_\star \in \mathbb{R}$, $M \gg 1$, and

$$\psi_r(x) = \begin{cases} 0, & x < 1-\delta, \\ \left(\dfrac{x-(1-\delta)}{\delta}\right)^r, & x \in [1-\delta, 1], \end{cases} \qquad r \geq 2.$$

Here, $M$ controls the extremeness of the unresolved extreme feature while the exponent $r$ controls its smoothness. This construction gives $\psi_r \in C^{r-1}$ at $1-\delta$, while the $r$-th derivative has a jump. In particular, as $r$ increases, the bump becomes flatter near $1-\delta$, and thus smoother.

Assume now that the training data do not intersect $E$, and that there are at least 2 distinct training data points. Then, on the observed region $\mathcal{X} \setminus E = [0, 1-\delta)$, empirical-risk minimization over $\mathcal{F}$ recovers

$$y_\xi(x) = a_\star x + b_\star,$$

in a noise-free setting. Hence,

$$y(x) - y_\xi(x) = 0, \qquad x \in \mathcal{X} \setminus E,$$

while on $E$,

$$y(x) - y_\xi(x) = M\psi_r(x) \geq 0.$$

Therefore, a standard calculation gives

$$\int_E |y - y_\xi| \, d\mu = M \int_{1-\delta}^1 \left(\frac{x-(1-\delta)}{\delta}\right)^r dx = \frac{M\delta}{r+1}.$$

On the other hand,

$$\int_{\mathcal{X}\setminus E} |y - y_\xi| \, d\mu = 0, \qquad \left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \, d\mu \right| = 0.$$

Hence the smallest admissible constants in the data-consistency conditions are $\widetilde{C} = \widehat{C} = 0$.

We now consider a slightly perturbed version, representing the potential noise in the non-extreme region, where

$$\int_{\mathcal{X}\setminus E} |y - y_\xi| \, d\mu \leq \varepsilon_{\text{bulk}}.$$

By definition,

$$\widehat{C} := \inf \left\{ C \geq 0 : \int_{\mathcal{X}\setminus E} |y - y_\xi| \, d\mu \leq C \int_E |y - y_\xi| \, d\mu \right\}.$$

Note that for small $\delta$, by linearity of the non-extreme feature and the approximation class,

$$\int_E |y - y_\xi| \, d\mu = \frac{M\delta}{r+1} + O(\varepsilon_{\text{bulk}}\delta),$$

it follows that one admissible choice of $C$ is

$$C = \frac{\varepsilon_{\text{bulk}}}{\frac{M\delta}{r+1} + O(\varepsilon_{\text{bulk}}\delta)} = \frac{1}{\frac{M\delta}{(r+1)\varepsilon_{\text{bulk}}} + O(\delta)}.$$

When the noise in $\mathcal{X} \setminus E$ does not 'magically' recover the unobserved extreme, i.e. when $M/\varepsilon_{\text{bulk}}$ dominates, we may simply pick

$$C = \frac{\varepsilon_{\text{bulk}} \, (r+1)}{M\delta},$$

and therefore,

$$\widehat{C} \leq \frac{\varepsilon_{\text{bulk}}(r+1)}{M\delta}.$$

Similarly,

$$\widetilde{C} := \inf \left\{ C \geq 0 : \left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \, d\mu \right| \leq C \left| \int_E (y - y_\xi) \, d\mu \right| \right\}.$$

Using

$$\left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \, d\mu \right| \leq \int_{\mathcal{X}\setminus E} |y - y_\xi| \, d\mu \leq \varepsilon_{\text{bulk}},$$

we again obtain the upper bound

$$\widetilde{C} \leq \frac{\varepsilon_{\text{bulk}}(r+1)}{M\delta}.$$

Consequently, we see that $\widehat{C}, \widetilde{C} < 1$ when $M/\varepsilon_{\text{bulk}}$ dominates, i.e. when the error in the unobserved extreme set dominates that in the observed non-extreme region.

**Methodology comment #iv.** *The theoretical formulation assumes a strictly deterministic dataset and estimator. The authors should briefly discuss how the theoretical lower bounds might shift if the inherent stochastic variance of training modern deep learning models, e.g. random weight initialization and batching, is properly accounted for.*

**Response.** We agree that this point should be made explicit. The previously submitted SI already contained an extension of the theory to random estimators and random datasets in the section Theoretical Extensions, under the subsection Generalize to Random Estimator and Dataset. We have now further clarified this point in the main text at line 244 to read as follows.

First, both the dataset $\mathcal{D}$ and the estimator $y_\xi$ are treated as deterministic objects. This simplification is not an essential restriction of the argument. In the presence of training randomness of modern neural networks, the same argument applies conditioning on a realization of the randomness, while averaging over these sources gives the corresponding expected bound. Thus stochastic variance may change the realized constants or confidence level of the lower bound developed below, but it does not remove the missing-tail contribution caused by data scarcity.

We have also revised the concluding sentence of this assumptions paragraph at line 255 to explicitly direct readers to the relevant SI section:

These limitations and considerations are either relaxed or addressed in the SI section "Theoretical Extensions".

**Methodology comment #v.** *In the Supplementary Material, the authors acknowledge that gradients from the ERM loss and the $W_1$ penalty frequently point in opposite directions. Because this contradictory signaling can lead to severe training divergence, this gradient conflict should be explicitly discussed in the main text, and a discussion on how the balancing hyperparameter $\lambda$ mitigates it would be helpful.*

**Response.** We thank the reviewer for pointing this out. We agree that the gradient conflict between the supervised ERM term and the $W_1$ penalty should be discussed more explicitly in the main text. We have revised the paper and the SI to discuss this more extensively.

To understand how to resolve the gradient conflict, we first want to clarify the source of this issue. The central working premise of $\eta$-learning is that the ERM estimator captures the bulk input-output relation, but underestimates or omits the tail behavior due to limited data. The role of the $W_1$ penalty is to correct the learned map in regions that affect the observable tail.

This premise also explains why a gradient conflict can occur during optimization. The ERM loss is computed from the labeled pairs $(x_i, u_i)$, whereas the $W_1$ loss is computed from tail samples selected according to the current observable distribution $(g \circ \phi_\theta)_{\#}\mu$. More precisely, at each tail-sample-update step, auxiliary samples from $\mu$ are passed through the current $\eta$-map to form an empirical approximation of $(g\circ\phi_\theta)_{\#}\mu$, and the upper-tail samples are used to define the $W_1$ penalty. In the intended extreme-event-deficient regime, the supervised data, as well as the bulk samples well represented by the ERM estimator, are expected to lie primarily in the bulk of the observable distribution. However, if the combined ERM+$W_1$ objective is optimized before the ERM loss has been sufficiently minimized, the model-induced ordering of samples by their predicted observable values can be unreliable. As a result, samples that should be treated as bulk under the supervised input-output relation can be spuriously ranked among the upper-tail samples used to define the $W_1$ loss. When this happens, the two loss components will provide incompatible gradient signals: the ERM term anchors the learned map to the labeled input-output relation, while the $W_1$ term pulls spuriously selected bulk samples toward upper-tail quantiles of $\nu_0$, leading to oscillatory or unstable optimization.

Conversely, once the ERM loss has been sufficiently minimized, the model-induced ordering used by the $W_1$ term is anchored to a predictor that already respects the supervised input-output relation. Bulk samples are then assigned ERM-consistent observable values and are no longer systematically ranked among the upper-tail samples merely because of unreliable early predictions. Thus, the tail set used to define the $W_1$ penalty is no longer dominated by samples that should be treated as bulk under the ERM-consistent map. In this regime, the $W_1$ term does not systematically pull ERM-consistent bulk predictions toward extreme quantiles of $\nu_0$. Instead, it primarily adjusts the tail of the learned observable distribution, while the ERM term plays a mainly stabilizing role, anchoring the learned map near the low-ERM regime already reached.

Based on this reasoning, our main stabilization mechanism is ERM pre-training. Specifically, we first train the model using only the supervised ERM loss, and then initialize both the $\eta$-learning stage and the initial tail set used to approximate the $W_1$ penalty from this pre-trained estimator, as shown in Algorithm 1 in the SI. At this point, the predictor already captures the labeled input-output relation well and induces a reliable ranking of the auxiliary samples. Consequently, $W_1$

is estimated with tail candidates under an ERM-consistent observable distribution. The ERM and $W_1$ terms are therefore much less prone to the severe misalignment that occurs when the $W_1$ penalty is introduced before the supervised fit has been learned. During the subsequent $\eta$-learning iterations, the tail set is dynamically updated as the learned map evolves, keeping the samples used to estimate $W_1$ separated from the ERM-induced bulk. This procedure systematically mitigates the conflict caused by unreliable tail-sample selection and has empirically yielded consistently stable and robust training.

We next clarify the role of the balancing hyperparameter $\lambda$. The parameter $\lambda$ can mitigate residual instability by controlling the relative magnitude of the two gradient components in

$$\nabla_\theta L(\theta) = \nabla_\theta L_{\mathrm{ERM}}(\theta) + \lambda \nabla_\theta L_{W_1}(\theta).$$

One potentially beneficial choice is to select $\lambda$ so that the scaled $W_1$ gradient has a comparable Euclidean norm to the ERM gradient, that is

$$\lambda \|\nabla_\theta L_{W_1}(\theta)\|_2 = \|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2, \implies \lambda = \frac{\|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2}{\|\nabla_\theta L_{W_1}(\theta)\|_2 + \varepsilon},$$

with a small $\varepsilon > 0$. This balances the magnitudes of the two gradient components and prevents the $W_1$ correction from overwhelming the supervised anchor.

However, we emphasize that $\lambda$ does not by itself resolve the directional incompatibility between anti-aligned gradients. It rescales the $W_1$ contribution, but it does not change the fact that an unreliable tail-set selection can make the two losses prefer opposing parameter updates. For this reason, we view ERM pre-training and the associated tail-set update as the main mechanism for resolving the severe early-stage conflict, with $\lambda$ serving as a secondary balancing parameter during the subsequent $\eta$-learning stage.

This interpretation is also supported by the robustness results reported in SI Fig. 4. In that experiment, the learned tail remains close to $\nu_0$ as $\lambda$ varies from $10^{-4}$ to 10, a range spanning five orders of magnitude. This indicates that our algorithm does not require delicate tuning of $\lambda$; rather, $\lambda$ mainly controls the strength of the residual tail correction around an already well-anchored supervised solution.

We have expanded the subsection Effective Training Strategies to discuss this reasoning in detail:

### Effective Training Strategies

Accurately estimating tail quantiles at every iteration can be memory- and compute-intensive, because it naively requires evaluating the model on a large auxiliary sample set $\tilde{x} \sim \mu$ at each update. To mitigate this, we use an inference-informed continual training strategy: periodically, we run inference on the auxiliary set to identify the inputs whose current observable values realize the target quantiles in $\mathcal{Q}$, and then restrict subsequent $W_1$ estimations to these identified inputs until the next refresh. This preserves a reliable estimate of the tail loss while substantially reducing per-iteration cost. A detailed computational complexity analysis of the full procedure is provided in SI Subsection "Computational Complexity".

This tail-set selection procedure also creates a potential optimization issue. The $W_1$ term is evaluated on inputs selected according to the current model-induced observable distribution $(g \circ \phi_\theta)_{\#}\mu$. Early in training, before the supervised input-output relation has been learned, this model-induced ordering can be unreliable. As a result, samples that should correspond to bulk behavior under an ERM-consistent predictor can be spuriously selected as tail samples and pulled

toward upper quantiles of $\nu_0$, while the ERM loss pulls the model toward the labeled pairs. The two loss components therefore provide conflicting gradient signals, leading to oscillatory or unstable optimization.

We mitigate this issue using a staged training procedure. We first train the model using only the supervised ERM loss, and then use the ERM estimator to initialize both the subsequent $\eta$-learning stage as well as the initial tail set used to approximate the $W_1$ penalty. After this pre-training step, the ordering used to select tail samples is anchored to a predictor that already respects the observed input-output relation. Consequently, the $W_1$ penalty primarily acts as a tail correction to an ERM-consistent map, while the ERM term continues to stabilize the learned map around the supervised solution. During the subsequent $\eta$-learning iterations, the tail set is periodically refreshed as the learned map evolves, keeping the samples used to estimate $W_1$ separated from the ERM-induced bulk. This procedure systematically mitigates the conflict caused by unreliable tail-sample selection and has empirically yielded consistently stable and robust training.

The balancing parameter $\lambda$, on the other hand, controls the strength of this residual tail correction. In practice, choosing $\lambda$ to make the $W_1$ gradient comparable in magnitude to the ERM gradient can be beneficial. However, $\lambda$ only rescales the $W_1$ contribution; it does not by itself resolve the directional incompatibility that can arise from unreliable early-stage tail-set selection. Thus, the main stabilization mechanism is ERM pre-training together with tail-set refresh, while $\lambda$ serves as a secondary balancing parameter. This interpretation aligns with the robustness study in SI Fig. 4, where $y_{\eta\#}\mu$ remains stable over a broad range of $\lambda$. A more detailed discussion of the gradient conflict and the practical choice of $\lambda$ is provided in SI Subsection "Pre-training and Inference-Informed Continual Training".

The analysis of the robustness result has also been expanded at line 626:

Additionally, as shown in SI Fig. 4, $y_{\eta\#}\mu$ remains stable as the balancing hyperparameter $\lambda$ varies from $10^{-4}$ to 10, spanning five orders of magnitude, while all other training configurations are fixed. This indicates that our method is numerically robust to the choice of $\lambda$, a practically desirable property. This robustness is enabled by the staged training procedure described later in "Effective Training Strategies".

The SI subsection "Pre-training and Inference-Informed Continual Training" has been revised as follows.

### Pre-training and Inference-Informed Continual Training

Several difficulties arise in the actual optimization of (14). We assume $\mathcal{F}_\theta$ is a neural network optimized with a stochastic first-order optimization method $\mathcal{M}$. The update rule is denoted by

$$\phi_\theta^{(k+1)} \leftarrow \mathcal{M}\left(\phi_\theta^{(k)}, \nabla_\theta \mathcal{L}_\theta^{(k)}\right).$$

Our focus remains on (14), as the other formulation can be viewed as a special case.

***Memory Challenges and Tail-Set Selection in $W_1$ Estimation.*** A practical challenge stems from approximating the $W_1$ term, even with its tractable one-dimensional quantile representation. Accurate estimation requires evaluating $g \circ \phi_\theta$ over the auxiliary set $\mathcal{D}_{\tilde{x}}$, which is typically much larger than the labeled training set. If this full auxiliary evaluation were included in every gradient-carrying update, the memory cost would be substantial.

To address this, we note that only $n_q$ specific inputs from $\mathcal{D}_{\tilde{x}}$ are ultimately needed to compute the empirical quantile loss, namely the inputs whose current observable values realize the

quantile levels in $\mathcal{Q}$. This motivates the Inference-Informed Continual Training (IICT) procedure. Periodically, we run inference on the full auxiliary set $\mathcal{D}_{\tilde{x}}$ and compute

$$\{g(\phi_\theta(\tilde{x}_j))\}_{j=1}^{n_{\tilde{x}}}.$$

We then identify the indices of the inputs corresponding to the prescribed quantile levels in $\mathcal{Q}$ and store them in a set $\mathcal{I}$. Between refreshes, the $W_1$ loss is evaluated only on the $\mathcal{I}$-indexed inputs. Thus, full inference on $\mathcal{D}_{\tilde{x}}$ is used only to identify the relevant tail samples, while the gradient-carrying $W_1$ update is restricted to the selected quantile samples. The refresh frequency is controlled by a hyperparameter $\omega$, with $\mathcal{I}$ updated every $\omega$ iterations. This procedure substantially reduces memory usage while preserving an accurate estimate of the tail contribution to the $W_1$ loss.

***Gradient Conflict and ERM Pre-training.*** The tail-set selection procedure also explains a potential source of optimization instability. The $W_1$ loss is evaluated on inputs selected according to the current model-induced observable distribution $(g \circ \phi_\theta)_{\#}\mu$. Early in training, before the supervised input-output relation has been learned, the ordering of the auxiliary observable values $g(\phi_\theta(\tilde{x}_j))$ can be unreliable. Consequently, inputs that should correspond to bulk behavior under an ERM-consistent predictor may be spuriously selected into $\mathcal{I}$ and treated as high-quantile samples. The $W_1$ term then pulls these samples toward upper quantiles of the reference distribution $\nu_0$, while the ERM loss pulls the model toward the labeled input-output relation. These two signals can be incompatible, leading to oscillatory or unstable optimization.

Our main stabilization mechanism is ERM pre-training. We first minimize the ERM loss alone and initialize $\phi_\theta$ with the resulting ERM estimator $u_\xi$ when solving (14). The initial index set $\mathcal{I}$ used by the $W_1$ penalty is also computed from this ERM-pretrained model. Thus, the first $W_1$ correction is applied to quantile samples selected under a predictor that already captures the observed input-output relation. This substantially reduces the chance that bulk samples are spuriously treated as tail samples because of unreliable early predictions. During $\eta$-learning, $\mathcal{I}$ is periodically refreshed as $\phi_\theta$ evolves, keeping the $W_1$ correction aligned with the current observable distribution.

***Choice of $\lambda$.*** The balancing parameter $\lambda$ controls the relative magnitude of the ERM and $W_1$ gradient components. Writing

$$\mathcal{L}_\theta = L_{\mathrm{ERM}}(\theta) + \lambda L_{W_1}(\theta),$$

we have

$$\nabla_\theta \mathcal{L}_\theta = \nabla_\theta L_{\mathrm{ERM}}(\theta) + \lambda \nabla_\theta L_{W_1}(\theta).$$

A practical scale for $\lambda$ is obtained by choosing it so that the scaled $W_1$ gradient has a comparable Euclidean norm to the ERM gradient during the early $\eta$-learning iterations:

$$\lambda \|\nabla_\theta L_{W_1}(\theta)\|_2 = \|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2 \implies \lambda = \frac{\|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2}{\|\nabla_\theta L_{W_1}(\theta)\|_2 + \varepsilon},$$

with a small $\varepsilon > 0$ for numerical stability. This prevents the $W_1$ correction from overwhelming the supervised anchor.

However, $\lambda$ only rescales the $W_1$ gradient; it does not change its direction. Therefore, while magnitude balancing could be potentially beneficial, it alone cannot resolve the directional incompatibility caused by unreliable tail-set selection. For this reason, ERM pre-training and periodic tail-set refresh are the primary mechanisms for stable optimization, while $\lambda$ serves as a secondary balancing parameter that controls the strength of the residual tail correction. This interpretation

is consistent with the robustness experiment in SI Fig. 4, where the learned observable distribution remains stable as $\lambda$ varies from $10^{-4}$ to 10 while all other training configurations are fixed.

**Methodology comment #vi.** *Theorem 6 relies on Assumption 5, which requires the model to make systematically biased errors without positive and negative errors canceling out. As this is hard to constrain during training, the authors should clarify how its failure impacts the point-wise reliability of the $\eta$-map.*

**Response.** We thank the reviewer for this comment. We agree that Assumption 5 being uniformly satisfied can be hard to impose and should not be interpreted as a property that is automatically enforced by the training objective. We have revised the discussion following Theorem 6 at line 443 to clarify its role and the consequence of its failure. That paragraph now reads as follows.

Theorem 6 reveals an important insight: effective recovery of extreme events is intrinsically linked to the distributional structure of the observable, justifying the $W_1$ term in $\eta$-learning as a distributional regularizer in limited-data regimes. However, the condition that Assumption 5 is uniformly satisfied as $W_1\left(y_{\xi\#}\mu, y_{\#}\mu\right) \to 0$ can be hard to impose during training without additional information. Therefore, if Assumption 5 cannot be assured, convergence in the observable distribution does not, by itself, imply point-wise reliability of the learned $\eta$-map. The learned map should then be interpreted as producing distributionally calibrated candidate extremes, rather than as guaranteeing recovery of their true location, time, or mechanism. Point-wise reliability requires additional identifying information, such as richer observables, spatial or temporal references, or structural and physical constraints during training.

**Methodology comment #vii.** *In the Optimality section, Assumptions 8a and 8b introduce the bounding parameters $\epsilon_1$ and $\epsilon_2$ to formalize $L^1$ deviations in the non-extreme set. A brief physical intuition of these parameters should be provided to improve readability.*

**Response.** We thank the reviewer for the suggestion. We have added this explanation in the Optimality section, immediately after Assumptions 8a and 8b, at line 501. We have also revised the surrounding text to clarify the rationale behind these two assumptions and to improve readability and flow. The revised passage reads as follows.

Intuitively, $\varepsilon_1$ controls the residual error of the ordinary supervised estimator in the non-extreme region, where labeled data are abundant. It therefore represents the baseline bulk approximation error. The parameter $\varepsilon_2$ controls the deviation of the $\eta$-estimator deviates from this supervised estimator in the same non-extreme region. Hence, Assumption 8a requires the baseline estimator to be reasonably accurate in the non-extreme region, while Assumption 8b further requires the discrepancy between the estimators $y_\xi$ and $y_\eta$ to be controlled there. These are natural conditions because both estimators are constrained by the same supervised-learning objective on the data-rich non-extreme region. With these assumptions in place, we are prepared to derive bounds expressed in terms of the exact approximation errors.

**Results comment #i.** *The toy models inadvertently expose the fragility of using a scalar observable $g(u)$. In the 2D-to-1D problem, the physical location of the extreme shifts across initializations (SI Figure 2); a metric, e.g. ensemble spatial variance, should quantify this. Furthermore, in the 2D-to-2D problem, the scalar penalty causes notable spatial distortion via interference between $u_1$*

*and $u_2$ (Figure 2). This interference effect should be discussed as a limitation of the 1D Wasserstein penalty.*

**Response.** We thank the reviewer for this helpful suggestion. We have revised the toy-example discussion to make this limitation explicit, while keeping the main-text discussion concise and moving the quantitative diagnostics to the SI.

For the 2D-to-1D problem, we now quantify the ensemble spatial variability of the inferred peak location. Specifically, using $K = 20$ independently initialized $\eta$-estimators with the same training configurations, we compute

$$\widehat{x}_k = \arg\max_{x \in \mathcal{X}} y_\eta^{(k)}(x), \qquad \bar{x} = \frac{1}{K} \sum_{k=1}^{K} \widehat{x}_k,$$

and

$$\sigma_{\text{loc}}^2 = \frac{1}{K} \sum_{k=1}^{K} \|\widehat{x}_k - \bar{x}\|_2^2.$$

For this problem, the ensemble mean inferred location is

$$\bar{x} = (-0.113, -0.461),$$

and the ensemble spatial variance is

$$\sigma_{\text{loc}}^2 = 15.935, \qquad \sigma_{\text{loc}} = 3.992.$$

These diagnostics show that independently initialized $\eta$-estimators can place the inferred high-output region at substantially different spatial locations. They are now reported in the SI subsection "The 2D-to-1D Toy Problem":

**Ensemble spatial variance.** To quantify the spatial variability of the inferred high-output region across neural-network initializations, we train $K = 20$ independent $\eta$-estimators while keeping the rest of the training configurations fixed. For realization $k$, define

$$\widehat{x}_k = \arg\max_{x \in \mathcal{X}} y_\eta^{(k)}(x).$$

We report the ensemble mean location

$$\bar{x} = \frac{1}{K} \sum_{k=1}^{K} \widehat{x}_k$$

and the ensemble spatial variance

$$\sigma_{\text{loc}}^2 = \frac{1}{K} \sum_{k=1}^{K} \|\widehat{x}_k - \bar{x}\|_2^2.$$

For the 2D-to-1D toy problem, the true grid maximizer is

$$x_\star = (2.020, -2.020),$$

whereas the ensemble mean inferred location is

$$\bar{x} = (-0.113, -0.461).$$

The ensemble spatial variance is

$$\sigma_{\text{loc}}^2 = 15.935, \qquad \sigma_{\text{loc}} = 3.992.$$

Thus, independently initialized $\eta$-estimators can place the inferred high-output region at substantially different spatial locations, quantifying the non-uniqueness presented in SI Fig. 3.

In the main text, we have rephrased the original analysis to improve readability, and added the new diagnostic at line 570:

***Analysis of the results.*** Fig. 1 shows the true map, the MSE map, and one realization of the $\eta$-map, while SI Fig. 1 compares the induced distributions $y_{\#}\mu$, $y_{\eta\#}\mu$, and $y_{\xi\#}\mu$. Additional $\eta$-map realizations and their corresponding PDF comparisons are shown in SI Figs. 2 and 3. The MSE map accurately fits the data-rich region but fails to infer the unobserved target extreme region, illustrating the data-consistent setting of Definition 3 and the fundamental limitation of MSE-based learning under tail data scarcity. In contrast, the $\eta$-map does not exactly recover the true extreme structure, but it introduces high-output mass in the tail while preserving the prediction in the data-rich region. This behavior is expected by theory: because the assumptions of Theorem 6 are not enforced as constraints during training, exact pointwise recovery is not guaranteed, and the available guarantee is convergence in $L^1(\mu)$. Distributionally, however, SI Fig. 1 shows that $y_{\eta\#}\mu$ is substantially closer to the reference distribution, which is the truth in this example, than the MSE-induced law, especially in the tail. Thus, $\eta$-learning reduces the epistemic uncertainty caused by missing extreme samples by producing statistically plausible extremes, rather than by uniquely identifying their exact spatial configuration.

The additional realizations in SI Fig. 2 show that distributional convergence does not uniquely determine the spatial location of the inferred high-output region. We quantify this variability in the SI by training $K = 20$ independently initialized $\eta$-estimators under the same configuration. The resulting maximizer locations have ensemble spatial standard deviation $\sigma_{\text{loc}} = 3.992$, while the 2-norm of the true maximizer is 2.86. Thus, independently initialized $\eta$-estimators can place the inferred high-output region at substantially different spatial locations.

The ensemble analysis highlights a broader implication of $\eta$-learning: enforcing scalar observable tail statistics generates candidate extremes, but does not necessarily identify the underlying physical structure, such as the peak location shown in this example. This variability reflects the residual physical uncertainty that remains after enforcing the observable tail statistics, rather than a failure to recover a uniquely identifiable hidden event. Nevertheless, two aspects are stable across realizations: predictions remain accurate in the data-rich region, and, as shown in SI Fig. 3, the induced output PDFs remain consistent with the reference distribution. In the Discussion section, we further describe how this uncertainty can be systematically characterized and used in downstream applications.

For the 2D-to-2D problem, rather than attributing the spatial distortion to noisy information in $u_2$, we now describe it as a structural limitation of using a scalar observable. We have also rephrased the previous analysis to improve readability. The result analysis now reads as follows at line 640:

***Analysis of the results.*** Fig. 2 shows the true map, the MSE map, and one realization of the $\eta$-map, while SI Fig. 5 compares the induced observable distributions $y_{\#}\mu$, $y_{\eta\#}\mu$, and $y_{\xi\#}\mu$. Additional $\eta$-map realizations and their corresponding PDF comparisons are shown in SI Figs. 6

and 7. As in the previous example, the MSE estimator accurately captures the map only in the data-rich region, but fails to detect the unobserved target extreme region. In contrast, the $\eta$-map uses the reference observable distribution to generate high-output responses in the missing tail region. Although the reconstructed extremes are not pointwise exact, the induced observable PDF remains closely aligned with the reference distribution, as shown in SI Fig. 5.

This example also exposes a limitation of constraining a scalar observable for high-dimensional states. Since $g(u) = 2|u_1| + \frac{1}{2}|u_2|$ aggregates the two state components into a single scalar quantity, the $W_1$ penalty constrains the observable law, but not how the extreme response decomposes between the two components. Compared with the 2D-to-1D case, this introduces an additional ambiguity: distortion in one state variable can be partially compensated by changes in the other while preserving similar scalar observable statistics. This scalarization-induced interference explains the more pronounced spatial distortion visible in Fig. 2, especially in $u_1$.

The additional realizations in SI Fig. 6 further show variability in the inferred extreme-region structure across neural-network initializations. Nevertheless, the induced observable PDFs remain consistently aligned with the reference distribution, as shown in SI Fig. 7. Overall, these results demonstrate that $\eta$-learning can generate plausible tail events under missing extreme data, but scalar-observable matching alone may leave residual uncertainty in the component-wise structure of high-dimensional states.

**Results comment #ii.** *For the first precipitation downscaling challenge, the degradation of the weighted coverage statistic at lower thresholds (SI Figures 8 and 9) indicates that penalizing the spatial maximum explicitly costs bulk spatial coherence. The authors should report standard spatial similarity metrics, e.g. SSIM or full-grid RMSE, alongside the extreme metrics to quantify how much spatial structure is sacrificed.*

**Response.** We thank the reviewer for this suggestion. We have added full-field spatial-fidelity diagnostics for the architecture-matched MSE and $\eta$ downscalers. Specifically, we now report full-grid RMSE as well as mean and standard deviation of SSIM. These metrics are computed over the complete HR precipitation fields.

The added diagnostics are evaluated over all $N = 9{,}044$ retained paired fields and over subsets defined by the true HR observable distribution. The bulk subsets are defined by $g(u) \leq Q_{\text{true}}(q)$ for $q \in \{0.7, 0.8, 0.9\}$, and the tail subsets are defined by $g(u) \geq Q_{\text{true}}(q)$ for $q \in \{0.95, 0.975, 0.99\}$.

The results confirm that emphasizing the spatial maximum incurs a measurable cost in pointwise fidelity, compared with the MSE-map. Over the full evaluation set, the full-grid RMSE increases from 2.877611 mm for the MSE downscaler to 3.111615 mm for the $\eta$ downscaler, an 8.13% relative increase. In the bulk subsets, the relative RMSE increases are smaller, ranging from 6.20% to 6.92%. In the tail subsets, the RMSE cost is larger, ranging from 9.21% to 10.30%, reflecting the greater difficulty of matching pointwise intensities on more extreme fields.

The SSIM diagnostics indicate that this pointwise cost is not accompanied by a comparable degradation of spatial structure. On the full evaluation set, $\overline{\text{SSIM}}$ changes from 0.947003 to 0.944207, a 0.30% relative decrease. Across all reported subsets, the largest relative $\overline{\text{SSIM}}$ decrease is 0.48%, occurring in the $q \geq 0.95$ tail regime. $\sigma_{\text{SSIM}}$ is slightly larger for the $\eta$ downscaler in each stratum, but the increase is small. These results refine the interpretation of SI Figs. 8 and 9: the lower-threshold weighted-coverage degradation reflects a real change in the magnitude-weighted allocation of precipitation above moderate thresholds, but it does not indicate a broad collapse of

spatial morphology.

We have revised the result analysis in the main text at line 718:

We further evaluate two statistics that are not directly enforced during training: the conditional mean $m(u_0; t, n_g)$ and the weighted coverage $c(u_0; t, n_g)$, where $u_0$ is a sample field, $t$ is a threshold, and $n_g$ is the number of spatial grid cells. The conditional mean measures the average magnitude above a threshold, while the weighted coverage measures the fraction of total field magnitude contributed by entries exceeding that threshold. Formal definitions are given in the SI. For $m$, we compute the PDFs across a range of thresholds from 154 to 214, and similarly for $c$. As shown in SI Fig. 8, the $u_\eta$-mapped samples consistently outperform the $u_\xi$-mapped samples with respect to the conditional mean statistic, further confirming the effectiveness of $\eta$-learning in capturing extreme values. SI Fig. 9 shows a trade-off at lower thresholds: the $u_\eta$-mapped weighted-coverage distribution agrees less well with the truth than the $u_\xi$-mapped distribution, especially in the bulk. This is expected because the training objective constrains the spatial maximum, not the magnitude-weighted exceedance fraction. However, as the threshold increases, the MSE-PDF progressively collapses into a Dirac delta, reflecting its failure to represent extremes, whereas the $\eta$-PDF continues to provide a valid proxy of the true distribution.

To further quantify the associated full-field cost, we provide RMSE and structural similarity index measure (SSIM) diagnostics in SI Table 1. Over all 9044 paired fields, the full-grid RMSE increases from 2.878 for $u_\xi$ to 3.112 for $u_\eta$, an 8.13% relative increase. Average SSIM ($\overline{\text{SSIM}}$) changes only from 0.947 to 0.944, a 0.30% relative decrease. We also evaluate subsets stratified by $\nu_0$: define $\mathcal{I}_{\le q} = \{1 \le j \le 9044 : \max(u_j) \le F_{\nu_0}^{-1}(q)\}$ and $\mathcal{I}_{\ge q}$ similarly, where $F_{\nu_0}^{-1}(q)$ denotes the $q$-quantile of $\nu_0$. For bulk subsets $\mathcal{I}_{\le 0.7}$, $\mathcal{I}_{\le 0.8}$, and $\mathcal{I}_{\le 0.9}$, the $u_\eta$-RMSE increase ranges from 6.20% to 6.92%, while the $u_\eta$-$\overline{\text{SSIM}}$ decrease remains below 0.27%. For tail subsets $\mathcal{I}_{\ge 0.95}$, $\mathcal{I}_{\ge 0.975}$, and $\mathcal{I}_{\ge 0.99}$, the $u_\eta$-RMSE increase ranges from 9.21% to 10.30%, while the $u_\eta$-$\overline{\text{SSIM}}$ decrease remains below 0.48%. These results show that emphasizing the scalar observable introduces a modest pointwise-intensity cost, with a larger effect on high-maximum fields, but does not produce a significant degradation in the local spatial structure measured by SSIM. Additional details of the spatial-fidelity analysis are provided in the SI.

Taken together, these diagnostics clarify the trade-off. On one hand, $\eta$-learning captures the prescribed extreme-observable statistics and transfers this improvement to related high-threshold statistics that are not explicitly enforced. On the other hand, this improvement comes with a modest spatial fidelity cost, primarily through pointwise amplitudes rather than a broad degradation of spatial morphology.

In SI subection *The Real-World Precipitation Downscaling Challenge*, immediately after the definitions of the conditional mean and weighted coverage statistics at line 1210, we add:

**Full-field spatial-fidelity metrics.** The conditional mean and weighted coverage characterize thresholded aggregate statistics of each precipitation field. They do not directly measure pointwise reconstruction accuracy or local spatial structure. We therefore supplement them with full-grid RMSE and structural similarity index measure (SSIM).

Let $u_j$ and $\widehat{u}_j$ denote the true and predicted HR fields for sample $j$, and let $n_g = 80 \cdot 160$ denote the number of HR grid cells. We compute

$$\text{RMSE} = \left[ \frac{1}{N n_g} \sum_{j=1}^{N} \|\widehat{u}_j - u_j\|_2^2 \right]^{1/2}.$$

Here, $N$ depends on the selected subset. We also compute the SSIM for each prediction-truth pair on the full HR field. Let $\mathcal{W}$ denote the set of $7 \times 7$ local windows. For a window $w \in \mathcal{W}$, define

$$\mathrm{SSIM}_w(\widehat{u}_j, u_j) = \frac{(2m_{\widehat{u}_{j,w}} m_{u_j,w} + C_1)(2\sigma_{\widehat{u}_j u_j,w} + C_2)}{(m^2_{\widehat{u}_{j,w}} + m^2_{u_j,w} + C_1)(\sigma^2_{\widehat{u}_{j,w}} + \sigma^2_{u_j,w} + C_2)},$$

where $m_{\widehat{u}_{j,w}}$ and $m_{u_j,w}$ are the local means, $\sigma^2_{\widehat{u}_{j,w}}$ and $\sigma^2_{u_j,w}$ are the local variances, and $\sigma_{\widehat{u}_j u_j,w}$ is the local covariance. The stabilizing constants are

$$C_1 = (0.01L)^2, \qquad C_2 = (0.03L)^2,$$

with the common intensity range

$$L = \max_{j,z} u_j(z) - \min_{j,z} u_j(z),$$

where the maximum and minimum are taken over the true HR set. The SSIM is then

$$\mathrm{SSIM}(\widehat{u}_j, u_j) = \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \mathrm{SSIM}_w(\widehat{u}_j, u_j).$$

Numerically, SSIM is a normalized local similarity score whose maximum value is 1, attained when the predicted and true fields agree within the local window up to numerical precision. Values close to 1 indicate that the two fields have similar local mean intensity, contrast, and spatial covariation, even if their pointwise amplitudes are not identical. Thus, SSIM complements RMSE: RMSE measures pointwise amplitude error in physical units, whereas SSIM measures whether the local spatial pattern is preserved after normalization by the local means and variances. Using a common intensity range $L$ for all samples makes the SSIM values comparable across subsets. Using a common intensity range $L$ for all samples makes the SSIM values comparable across subsets. We report the sample mean $\overline{\mathrm{SSIM}}$ and sample standard deviation $\sigma_{\mathrm{SSIM}}$ for each selected subset.

For the $\eta$-map, we also report the relative RMSE increase compared to the MSE-map,

$$r_{\eta/\mathrm{MSE}} = \frac{\mathrm{RMSE}_\eta - \mathrm{RMSE}_{\mathrm{MSE}}}{\mathrm{RMSE}_{\mathrm{MSE}}}.$$

The bulk subsets are defined as

$$\mathcal{I}_{\le q} := \{1 \le j \le 9044 : \max(u_j) \le F^{-1}_{\nu_0}(q)\},$$

and the tail subsets are defined as

$$\mathcal{I}_{\ge q} := \{1 \le j \le 9044 : \max(u_j) \ge F^{-1}_{\nu_0}(q)\}.$$

Over the full evaluation set with 9044 samples, the $\eta$ downscaler has an 8.13% larger full-grid RMSE than the MSE downscaler. The $\overline{\mathrm{SSIM}}$ decrease is 0.30%, and $\sigma_{\mathrm{SSIM}}$ increases by 5.36%. Thus, the main full-field cost is pointwise intensity error rather than a large loss of local spatial structure.

The bulk subsets show the same pattern with a smaller RMSE cost. For $\mathcal{I}_{\le 0.7}$, $\mathcal{I}_{\le 0.8}$, and $\mathcal{I}_{\le 0.9}$, the relative RMSE increases are 6.20%, 6.55%, and 6.92%, respectively. The corresponding $\overline{\mathrm{SSIM}}$ decreases are 0.19%, 0.23%, and 0.27%. Hence, in the bulk and moderate regimes, the $\eta$-map sacrifices a small amount of pointwise fidelity while retaining high structural similarity.

| Subset | Method | $N$ | RMSE | $\overline{\text{SSIM}}$ | $\sigma_{\text{SSIM}}$ | $r_{\eta/\text{MSE}}$ |
|---|---|---|---|---|---|---|
| Full | MSE | 9044 | 2.877611 | 0.947003 | 0.043672 | – |
| | $\eta$ | | 3.111615 | 0.944207 | 0.046011 | 8.13% |
| $\mathcal{I}_{\leq 0.7}$ | MSE | 6331 | 1.739108 | 0.965508 | 0.029943 | – |
| | $\eta$ | | 1.846915 | 0.963688 | 0.031964 | 6.20% |
| $\mathcal{I}_{\leq 0.8}$ | MSE | 7235 | 2.008767 | 0.960048 | 0.033681 | – |
| | $\eta$ | | 2.140332 | 0.957853 | 0.035877 | 6.55% |
| $\mathcal{I}_{\leq 0.9}$ | MSE | 8139 | 2.344438 | 0.954014 | 0.037959 | – |
| | $\eta$ | | 2.506664 | 0.951470 | 0.040278 | 6.92% |
| $\mathcal{I}_{\geq 0.95}$ | MSE | 453 | 6.469183 | 0.877379 | 0.040998 | – |
| | $\eta$ | | 7.130227 | 0.873164 | 0.042823 | 10.22% |
| $\mathcal{I}_{\geq 0.975}$ | MSE | 227 | 7.191779 | 0.875916 | 0.040929 | – |
| | $\eta$ | | 7.932229 | 0.872659 | 0.042427 | 10.30% |
| $\mathcal{I}_{\geq 0.99}$ | MSE | 91 | 7.904075 | 0.870943 | 0.038264 | – |
| | $\eta$ | | 8.631702 | 0.869387 | 0.039693 | 9.21% |

Table 1: HR spatial-fidelity metrics for the precipitation downscalers. The subsets are selected by the true HR observable distribution: $\mathcal{I}_{\leq q}$ consists of sample indices whose corresponding true HR spatial maximum falls below the $q$-quantile, and $\mathcal{I}_{\geq q}$ is similarly defined. Here $r_{\eta/\text{MSE}} = (\text{RMSE}_\eta - \text{RMSE}_{\text{MSE}})/\text{RMSE}_{\text{MSE}}$ denotes the relative RMSE increase of the $\eta$-downscaler over the MSE-downscaler.

In the tail subsets, both downscalers have larger RMSE because the selected fields are more extreme. The $\eta$-downscaler has relative RMSE increases of 10.22%, 10.30%, and 9.21% for $\mathcal{I}_{\geq 0.95}$, $\mathcal{I}_{\geq 0.975}$, and $\mathcal{I}_{\geq 0.99}$ respectively. However, $\overline{\text{SSIM}}$ remains close to the MSE baseline, with relative decreases of 0.48%, 0.37%, and 0.18%. These results show that the spatial-maximum penalty induces a measurable but limited full-field cost. Together with SI Fig. 9, they indicate that the lower-threshold weighted-coverage degradation reflects a change in the magnitude-weighted allocation of precipitation above moderate thresholds, not a broad collapse of spatial morphology.

**Results comment #iii.** *The case study using a hypothesized GEVD prior (SI Figure 12) highlights that the model will faithfully hallucinate extremes to match a potentially flawed distribution. A quantitative sensitivity analysis, i.e. ablation study, showing how drastically physical fidelity degrades when $\nu_0$ is misspecified by a known margin would be valuable.*

**Response.** We thank the reviewer for this suggestion. We have added a controlled misspecification analysis for the hypothesized-GEVD experiment. The goal is to quantify how errors in the supplied scalar tail information affect the fidelity of the resulting HR precipitation fields.

We perturb the supplied reference values directly at the selected quantile levels $\mathcal{Q} = \{q_i\}_{i=1}^{n_q}$. Let $Q_{\text{true}}$ and $Q_{\text{GEVD}}$ denote the empirical HR and hypothesized GEVD quantile functions of the spatial-maximum observable. We define

$$Q_\alpha(q_i) = Q_{\text{true}}(q_i) + \alpha\,[Q_{\text{GEVD}}(q_i) - Q_{\text{true}}(q_i)]\,, \qquad q_i \in \mathcal{Q}.$$

Here, $\alpha = 0$ gives the empirical HR tail target and $\alpha = 1$ gives the hypothesized GEVD target.

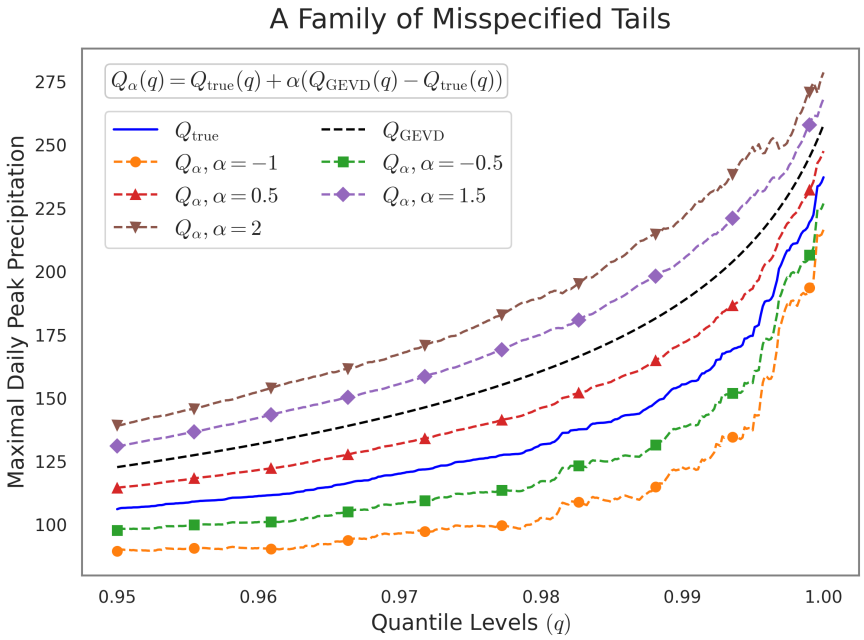



Figure 1: Sensitivity analysis for misspecified reference tail. Tail-target profiles obtained by shifting the true HR quantiles using the parameter $\alpha$.

Negative $\alpha$ shifts the target in the lighter-tail direction, while $\alpha > 1$ extrapolates beyond the original GEVD target. Defining

$$\Delta_\alpha = \frac{1}{n_q} \sum_{i=1}^{n_q} |Q_\alpha(q_i) - Q_{\text{true}}(q_i)|,$$

we have

$$\Delta_\alpha = |\alpha| \Delta_1.$$

Thus, $|\alpha|$ measures the size of the perturbation relative to the discrepancy between the hypothesized GEVD and empirical HR tail targets. We evaluated five misspecified-tail models with

$$\alpha \in \{-1, -0.5, 0.5, 1.5, 2\},$$

while holding the remaining training configurations fixed. The family of tail-shifted $\nu_0$ is provided in Fig. 1, and the corresponding $\eta$-estimated HR observable distributions are presented in Fig. 2.

The results show clear sensitivity of the pointwise spatial fidelity to heavier-tail misspecification. The full-grid RMSE increases from 2.926038 at $\alpha = -1$ to 3.527282 at $\alpha = 2$, a 20.55% increase. The intermediate values are ordered by tail strength: $\alpha = -0.5$ remains close to $\alpha = -1$, $\alpha = 0.5$ gives a moderate increase, and $\alpha = 1.5$ is close to the largest-error end of the sweep. At the same perturbation magnitude, the heavier-tail direction is more damaging than the lighter-tail direction: $\alpha = 0.5$ gives RMSE 3.193944, compared with 2.967890 for $\alpha = -0.5$. Within the positive heavy-tail sweep, increasing the perturbation from $0.5\Delta_1$ to $2\Delta_1$ increases RMSE from 3.193944 to 3.527282, a 10.43% increase.

$\overline{\text{SSIM}}$ remains high across the sweep, between 0.940689 and 0.947280. It is not strictly monotone: the $\alpha = 2$ model has the largest RMSE but a slightly higher $\overline{\text{SSIM}}$ than the $\alpha = 1.5$ model. $\sigma_{\text{SSIM}}$

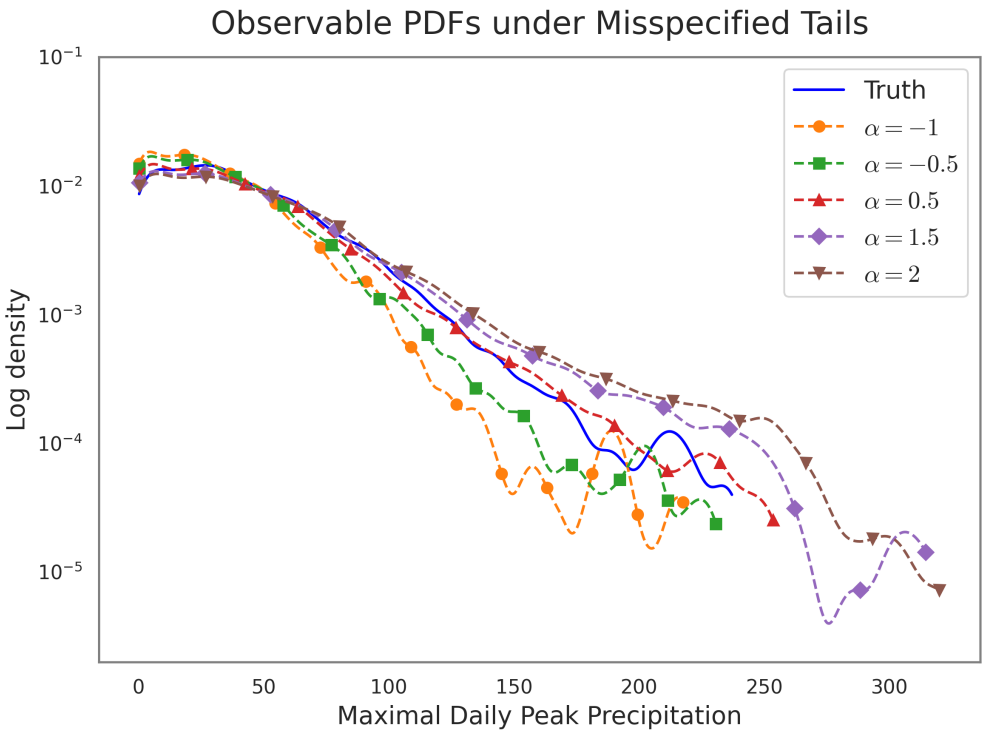



Figure 2: Sensitivity analysis for misspecified reference tail. Resulting observable distributions from $\eta$-downscalers trained with the corresponding shifted $\nu_0$ shown in Fig. 1.

increases mildly as the tail is made heavier, from 0.043792 at $\alpha = -1$ to 0.047701 at $\alpha = 2$. These results indicate that the misspecification primarily affects pointwise fidelity, while the aggregate spatial structure measured by SSIM remains comparatively stable.

We have revised the result analysis in the main text at line 917:

***Analysis of the results.*** The visualizations of the resulting downscaled fields and the corresponding observable PDFs are presented in SI Fig. 12 and Fig. 6. The observable law of the $\eta$-downscaled fields is shifted toward the hypothesized GEVD and therefore has a heavier tail than the true HR law. This experiment shows that the framework can generate fields conditional on tail information extending beyond the observed record, thereby broadening our understanding of potential extreme scenarios by leveraging only hypothetical tail information—a critical advantage in risk-sensitive applications.

At the same time, the experiment does not validate the hypothesized $\nu_0$ or the amplified precipitation values produced under that hypothesis. This distinction is important: since the distributional regularizer is designed to enforce the supplied tail statistics, if the tail is misspecified, the learned map can generate correspondingly misspecified extreme amplitudes. We quantify this effect through the controlled sensitivity analysis reported in SI Table 2, with details provided in the same SI subsection. The supplied tail quantiles are perturbed by a scalar parameter $\alpha$, for which the misspecification magnitude is proportional to $|\alpha|$. As the target is moved in the heavier-tail direction from $\alpha = -1$ to $\alpha = 2$, the full-grid $u_\eta$-RMSE increases from 2.926 to 3.527, a 20.55% increase. $u_\eta$-$\overline{\text{SSIM}}$ remains between 0.941 and 0.947, indicating that the main degradation is in

amplitude-sensitive pointwise fidelity rather than in the spatial morphology measured by SSIM.

Accordingly, SI Fig. 12 should be interpreted as showing scenarios conditional on the hypothesized GEVD. The fields illustrate how the learned downscaling map responds when an assumed heavier tail is imposed, rather than providing evidence that the amplified extremes will occur or that their magnitudes are physically validated. The experiments therefore demonstrate both the scenario-generation capability of the framework and its dependence on the accuracy of the supplied reference information. Such conditional scenarios can help inform risk assessment and decision-making when the reference tail is treated as an explicit modeling hypothesis and accompanied by proper sensitivity analysis.

In SI subsection "$\eta$-Downscaling with Hypothesized Heavier-Tailed Distributions", we add:

**Sensitivity to misspecified tails.** We examine how errors in the supplied tail information propagate to the downscaled precipitation fields. The $W_1$ estimator uses $\nu_0$ through its quantiles at the selected probability levels

$$\mathcal{Q} = \{q_i\}_{i=1}^{n_q}, \qquad n_q = 350, \qquad q_i \in [0.95, 1].$$

We thus perturb $\nu_0$ directly at $\mathcal{Q}$. Let $Q_{\text{true}}(q)$ denote the true HR observable quantile, and let $Q_{\text{GEVD}}(q)$ denote the corresponding quantile of the hypothesized truncated GEVD. We define

$$Q_\alpha(q_i) = Q_{\text{true}}(q_i) + \alpha \left[ Q_{\text{GEVD}}(q_i) - Q_{\text{true}}(q_i) \right], \qquad q_i \in \mathcal{Q}.$$

Here, $\alpha = 0$ gives the empirical HR target, while $\alpha = 1$ gives the hypothesized GEVD target. Negative $\alpha$ moves the supplied tail information in the lighter-tail direction, whereas $\alpha > 1$ extrapolates beyond the GEVD target.

The average magnitude of the imposed tail perturbation is

$$\Delta_\alpha = \frac{1}{n_q} \sum_{i=1}^{n_q} \left| Q_\alpha(q_i) - Q_{\text{true}}(q_i) \right|.$$

Since

$$Q_\alpha(q_i) - Q_{\text{true}}(q_i) = \alpha \left[ Q_{\text{GEVD}}(q_i) - Q_{\text{true}}(q_i) \right],$$

we have

$$\Delta_\alpha = |\alpha| \Delta_1.$$

Thus, $|\alpha|$ specifies the misspecification magnitude relative to the discrepancy between the hypothesized GEVD tail and the true HR tail. For each value of

$$\alpha \in \{-1, -0.5, 0.5, 1.5, 2\},$$

we use the same training configuration except that $\nu_0$ varies across the perturbed family. We evaluate the complete predicted HR fields using RMSE together with $\overline{\text{SSIM}}$ and $\sigma_{\text{SSIM}}$. All metrics are computed over the $N = 9{,}044$ HR fields.

The full-grid RMSE increases as the supplied target is shifted in the heavier-tail direction. The lowest RMSE occurs at $\alpha = -1$, with RMSE $= 2.926038$, while the largest occurs at $\alpha = 2$, with RMSE $= 3.527282$. This is a 20.55% increase relative to the $\alpha = -1$ case. The intermediate values follow the same ordering by tail strength: $\alpha = -0.5$ remains close to $\alpha = -1$, $\alpha = 0.5$ is moderately larger, and $\alpha = 1.5$ is close to the largest-error end of the sweep. At equal misspecification magnitude, the heavier-tail direction is more damaging: $\alpha = 0.5$ gives RMSE 3.193944, whereas

| $\alpha$ | RMSE | $\overline{\text{SSIM}}$ | $\sigma_{\text{SSIM}}$ |
|---|---|---|---|
| $-1.0$ | 2.926038 | 0.947280 | 0.043792 |
| $-0.5$ | 2.967890 | 0.946510 | 0.044193 |
| 0.5 | 3.193944 | 0.944542 | 0.045382 |
| 1.5 | 3.423614 | 0.940689 | 0.048021 |
| 2.0 | 3.527282 | 0.941217 | 0.047701 |

Table 2: Sensitivity of the full-field fidelity of the $\eta$-mapped fields to misspecification of the supplied reference tail. All metrics are computed over the full $N = 9044$ paired HR fields.

$\alpha = -0.5$ gives RMSE 2.967890. Within the positive heavy-tail sweep, increasing the target discrepancy from $0.5\Delta_1$ to $2\Delta_1$ increases RMSE by 10.43%.

The SSIM behavior is more stable. $\overline{\text{SSIM}}$ remains between 0.940689 and 0.947280 for all misspecified-tail models. It is not strictly monotone in $\alpha$: the $\alpha = 2$ model has the largest RMSE but a slightly higher $\overline{\text{SSIM}}$ than the $\alpha = 1.5$ model. $\sigma_{\text{SSIM}}$ increases mildly from 0.043792 at $\alpha = -1$ to 0.047701 at $\alpha = 2$. Thus, heavier-tail misspecification mainly degrades amplitude-sensitive pointwise fidelity, while the aggregate spatial structure measured by SSIM remains relatively close to the truth.

This experiment confirms that the pointwise fidelity of the $\eta$-generated fields is sensitive to the supplied reference tail. A heavier-than-correct reference profile can induce mismatched pointwise precipitation amplitudes even when the fields retain high SSIM. The GEVD-based outputs should therefore be interpreted conditionally on the hypothesized tail law, rather than as physically validated predictions.

**Results comment #iv.** *The algorithm requires dynamic approximation of the 1-Wasserstein distance, yet a discussion on computational complexity is missing. The authors should include a brief quantitative analysis to set realistic expectations, particularly concerning the overhead of adding this to already expensive generative pipelines like Flow Matching.*

**Response.** We thank the reviewer for raising this concern. We have added a detailed computational complexity analysis to the SI and referenced it in the main text at line 1140:

A detailed computational complexity analysis of the full procedure is provided in SI subsection "Computational Complexity".

We emphasize that, in the generative-modeling experiment, $\eta$-learning does not modify the training or sampling procedure of the Flow Matching (FM) model. The low-resolution (LR) FM model is trained and used to generate LR samples exactly as when no $\eta$-Learning is involved. The $\eta$-map is trained separately and applied only after FM sampling. This requires one additional forward pass per generated LR sample. In particular, FM sampling involves no $W_1$ computation or backward propagation through the $\eta$-map. Therefore, aside from the separate one-time training of the $\eta$-map, the only generation-time overhead relative to the baseline FM pipeline is the forward propagation of the generated LR samples through the trained $\eta$-map.

The complexity analysis subsection added to the SI at line 942 is attached below.

### Computational Complexity

Here, a detailed complexity analysis of the full $\eta$-Learning algorithm, i.e. Algorithm 1, is provided. We distinguish the inference-only computation used to identify the relevant tail samples from the gradient-carrying computation used to update the neural network. Let $n_{\tilde{x}} = |\mathcal{D}_{\tilde{x}}|$ be the number of inputs used to estimate the push-forward distribution, $n_q = |\mathcal{Q}|$ the number of quantiles for estimating $W_1$, and $C_\mathrm{f}$ and $C_\mathrm{b}$ the per-sample forward- and backward-propagation costs through $g \circ \phi_\theta$, respectively. Repeated input indices are allowed when different quantile levels correspond to the same empirical sample; hence, the gradient-carrying $W_1$ loss always contains $n_q$ samples.

We consider the conservative case $\omega = 1$, in which case the quantile indices are updated at every optimization iteration. First, $g \circ \phi_\theta$ is evaluated on all $n_{\tilde{x}}$ inputs in inference mode, and the resulting scalar values are sorted to identify the prescribed quantiles. This step has cost

$$O\left(n_{\tilde{x}} C_\mathrm{f} + n_{\tilde{x}} \log n_{\tilde{x}}\right).$$

Because this full-sample evaluation is forward-only, it does not require retaining a computational graph. The $W_1$ loss is then computed and back-propagated using only on the $n_q$ quantile-indexed samples, with cost

$$O\left(n_q (C_\mathrm{f} + C_\mathrm{b})\right).$$

Therefore, the additional per-iteration cost of the $W_1$ regularization is

$$O\left(n_{\tilde{x}} C_\mathrm{f} + n_{\tilde{x}} \log n_{\tilde{x}} + n_q (C_\mathrm{f} + C_\mathrm{b})\right).$$

The main practical distinction is that forward propagation is performed on the full input pool, whereas the more expensive backward propagation is restricted to the $n_q$ dynamically selected tail samples. Note that $n_{\tilde{x}}$ and $n_q$ can be mini-batched if the memory cap is of concern.

In the precipitation experiment, $n_{\tilde{x}} = 9044$, and the matched tail begins at approximately the 0.975 quantile. The quantile grid therefore contains approximately

$$n_q \simeq (1 - 0.975) n_{\tilde{x}} \simeq 226$$

terms. Thus, although all 9044 inputs are used in the inference-only forward pass, only approximately 2.5% as many samples are used in the gradient-carrying $W_1$ update.

For the Flow Matching experiment, the LR Flow Matching model is trained on the LR data and used to generate the 25-year-equivalent LR samples exactly as it would be without $\eta$-learning. The trained $\eta$-map is then applied once to each generated LR sample to obtain a tail-corrected HR sample. For fair comparison, $n_{\tilde{x}}$ LR samples are generated, and the additional $\eta$-evaluation cost is therefore

$$O\left(n_{\tilde{x}} C_\mathrm{f}\right).$$

This is a single inference-only post-processing pass per sample. It does not alter the training cost of the Flow Matching model, does not add computations to its numerical sampling steps, and requires no $W_1$ evaluation or backward propagation. The other additional expense is the separate, one-time training of the $\eta$-map, whose per-iteration complexity is given above.

**Results comment #v.** *While the Discussion highlights the framework's versatility, it lacks critical evaluation. The authors must add a "Limitations and Future Work" subsection summarizing the major caveats identified above. Acknowledging these will strengthen the manuscript and guide future research toward multi-dimensional observables.*

**Response.** We thank the reviewer for the suggestion. We have added a dedicated subsection titled "Limitations and Future Work" at line 1033 to acknowledge the major caveats and discuss important future directions. The subsection reads as follows.

### Limitations and Future Work

While $\eta$-learning offers a promising approach for modeling extreme events under data scarcity, several important limitations remain. Here, we identify key limitations and corresponding directions for future work.

A central limitation is that matching the distribution of a prescribed observable does not identify the dynamical mechanism that generates extreme events. The $W_1$ regularizer constrains observable statistics, but does not by itself determine the temporal evolution, spatial organization, or broader physical coherence of the generated extremes. To the extent that these properties are recovered, they are inherited from the supervised input-output data rather than supplied by the $W_1$ term. Consequently, when distinct mechanisms or state configurations induce the same observable law, they are not distinguishable under the present objective. An important direction is therefore to extend the current framework with temporal and spatial constraints, conservation laws, or other physical priors.

This non-identifiability also arises at the level of individual events: the prescribed observable law does not generally determine the precise location, occurrence time, or full spatial structure of an unobserved extreme. Variability in these features across independently trained $\eta$-maps should therefore be interpreted as residual uncertainty. Importantly, this uncertainty can be exploited systematically. Ensembles of $\eta$-maps can generate diverse yet statistically consistent candidate extremes, which can then be organized according to physically meaningful attributes such as event location, occurrence time, or spatial pattern. Representative scenarios can guide subsequent rounds of adaptive data acquisition, such as targeted high-fidelity simulations or experiments, allowing the corresponding hypotheses to be tested, refined, or rejected while progressively reducing uncertainty in data-scarce regions.

Another limitation is the reliance on an accurate reference law $\nu_0$. Because the tail-emphasized $W_1$ estimation is designed to match the supplied tail, misspecification of $\nu_0$, or a shift in the target tail relative to $\nu_0$, can be transferred directly to the generated extremes. A promising extension is a distributionally robust formulation that replaces a single reference law with an ambiguity set of plausible laws, with particular emphasis on uncertainty and shifts in the tail. Worst-case or uncertainty-weighted tail objectives could reduce overcommitment to any single hypothesized distribution while preserving awareness of extreme behavior.

Finally, the present theory and algorithms are developed for scalar observables, for which $W_1$ enables efficient computation and underpins the tail optimality results. For a vector observable $G = (g_1, \ldots, g_r)$, one may either introduce a physically meaningful severity function $s : \mathbb{R}^r \to \mathbb{R}$ and apply the present framework to $g = s \circ G$, or use a sum of componentwise $W_1$ penalties when matching the marginal laws is sufficient. The latter scales linearly with $r$, but does not constrain the correlation structure among the components. When the joint law of $G$ is essential, genuinely multivariate discrepancies based on, for example, sliced Wasserstein distance or entropically regularized optimal transport are required [1, 2]. Extending the present tail-focused theory and scalable algorithms to these settings is another important direction.

**Results comment #vi.** *The empirical reproducibility of the work is overall high. The authors provide thorough details regarding the ERA5-Land dataset processing, hyperparameter choices, and*

*algorithmic training loops.*

**Response.** We appreciate this assessment. We have further improved reproducibility by expanding the code documentation, pinning dependencies, adding hardware/runtime notes, and providing a minimal ERA5-Land test pipeline as described below.

**Code availability.** *First, it should be noted that I reviewed the provided code repository without installing or running the scripts. The repository maps to the experiments in the manuscript. The authors provide separate notebooks for the different case studies, the 2D toy problems, ERA5-Land downscaling, and the generative model integration. While the code for the mathematical implementation is there, it would be very helpful if the authors added inline comments linking the core definitions directly to the numbered equations in the text. This would make it much easier for readers to follow the translation from theory to code.*

*Regarding usability, the README needs to be updated with proper environment setup instructions, ideally through a* `requirements.txt` *file. Please ensure that the exact versions of the core dependencies are pinned so the codebase does not break due to future library updates. Finally, a brief note on the expected hardware requirements, such as required GPU VRAM and approximate processing times, should be added. This is useful for anyone looking to reproduce these results.*

*Finally, for the real-world precipitation challenge, the code's usability relies on the availability of the ERA5-Land dataset. The repository could include a minimal, pre-processed subset of the data, or an automated script that downloads a small sample, so users can easily test the pipeline without needing to download and format massive climate datasets.*

**Response.** We thank the reviewer for examining the code repository. We have revised the implementation and documentation to address all of the requested reproducibility components. The core training and evaluation routines now contain inline comments linking the implementation to the corresponding equations and algorithms in the manuscript. The README also includes an equation-to-code map identifying the implementations of the supervised and $W_1$-regularized objectives, the tail-quantile approximation, the quantile-index update, and the inference-informed continual training algorithm.

We have further added a pinned software environment and complete setup instructions, including the tested Python, PyTorch, CUDA, and supporting-library versions. The README reports the tested hardware configuration, data dimensions, principal hyperparameters, and approximate runtime and peak GPU-memory requirements for the toy problems, ERA5-Land downscaling, Flow Matching training and sampling, and GEVD-prior downscaling. Finally, we have provided a minimal ERA5-Land reproducibility workflow, together with the associated preprocessing and smoke-test instructions, so that users can verify data loading, loss evaluation, and model execution without reproducing the complete full-scale dataset pipeline.

## Response to Reviewer #2

**General assessment.** *In this paper, an Extreme Event Aware (e2a or eta) or $\eta$-learning algorithm is developed. It does not assume the existence of extreme events in the available data. It reduces uncertainty even in "uncharted" extreme-event regions by enforcing the extreme-event statistics of an observable indicative of extremeness during training, which can be supported by qualitative arguments or estimated from unlabeled data.*

*Overall, the method is novel, and the paper is well-written. The topic is also a good fit for the scope of Nature Communications. That said, a few points are not very clear, and I believe they need to be explained and/or validated more systematically. I suggest a revision of this paper.*

**Response.** We thank the reviewer for the positive assessment and for identifying points where the method should be explained more systematically.

**Comment #1.** *It looks like $g$ is a very low-dimensional output function. In the paper, it says $g$ belongs to $\mathbb{R}$, which is one dimension. However, in practice, the observable may not always be a scalar. It can be an extreme pattern or a combination of several features. What is the scalability of the method?*

**Response.** We agree that certain applications may require more than a scalar observable. The present work formulates $g$ as scalar because the one-dimensional quantile representation of $W_1$ underlies both the efficient implementation and the tail-error optimality results. The framework can nevertheless be extended to multivariate observables in a computationally scalable manner. There are three distinct settings to consider.

If a vector of features $G = (g_1, \dots, g_r)$ admits a physically meaningful scalar severity score $s : \mathbb{R}^r \to \mathbb{R}$, one can take $g = s \circ G$ and apply the present scalar framework directly. This includes pattern-based events whenever a suitable scalar score is available.

When scalarization is not desired and calibration of the componentwise marginal laws is sufficient, a direct extension is

$$\sum_{j=1}^{r} W_1((g_j \circ \phi_\theta)_{\#}\mu, \nu_{0,j}) \,,$$

where $\nu_{0,j}$ denotes the reference law for the $j$-th observable. A detailed complexity analysis for the single-obsevable case is provided in newly added SI subsection Computational Complexity. We briefly restate the analysis here for assistance. Let $n_{\tilde{x}}$ be the number of auxiliary inputs used to estimate the push-forward distributions, $n_q$ the number of quantile levels per observable, and $C_\mathrm{f}$ and $C_\mathrm{b}$ the costs of one forward and backward model propagation, respectively. At each quantile-index refresh, the model is evaluated once on all $n_{\tilde{x}}$ auxiliary inputs, and these outputs are shared across the $r$ observables, giving a cost of $O(n_{\tilde{x}} C_\mathrm{f})$. Locating the quantile samples requires sorting $n_{\tilde{x}}$ values for each observable, with total cost $O(r n_{\tilde{x}} \log n_{\tilde{x}})$. The subsequent gradient computation uses at most $r n_q$ selected quantile samples and therefore costs $O(r n_q (C_\mathrm{f} + C_\mathrm{b}))$. Thus, under the worst-case assumption that the quantile indices are refreshed at every iteration, the componentwise Wasserstein regularization has total cost

$$O(n_{\tilde{x}} C_\mathrm{f} + r n_{\tilde{x}} \log n_{\tilde{x}} + r n_q (C_\mathrm{f} + C_\mathrm{b})) \,.$$

The extension therefore scales linearly with the number of observables $r$, while the full inference

pass over the auxiliary inputs is shared across all observables.

Third, if the joint dependence or correlation structure of $G$ is also important, then marginal matching is insufficient and a genuinely multivariate discrepancy is needed. Tools such as sliced Wasserstein distance and entropically regularized optimal transport can be integrated into the framework in place of the scalar $W_1$ term. Both have been widely adopted in machine learning as computationally tractable approaches to multivariate distribution comparison: sliced Wasserstein reduces the problem to one-dimensional projections, while entropic regularization enables efficient and parallelizable Sinkhorn iterations. We expect tail-enhanced versions of these discrepancies to be important when the goal is to correct tail bias in a multivariate extreme-event setting.

We have further emphasized the scalar-observable setup and pointed out the extensions to multivariate observables in the Problem Statement at line 168:

Throughout this work, we focus on scalar-valued observables; remarks on extensions to multivariate settings are provided under "Limitations and Future Work".

We have also added the following discussion to the newly added "Limitations and Future Work" under the section "Discussion" at line 1067:

Finally, the present theory and algorithms are developed for scalar observables, for which $W_1$ enables efficient computation and underpins the tail optimality results. For a vector observable $G = (g_1, \ldots, g_r)$, one may either introduce a physically meaningful severity function $s : \mathbb{R}^r \to \mathbb{R}$ and apply the present framework to $g = s \circ G$, or use a sum of componentwise $W_1$ penalties when matching the marginal laws is sufficient. The latter scales linearly with $r$, but does not constrain the correlation structure among the components. When the joint law of $G$ is essential, genuinely multivariate discrepancies based on, for example, sliced Wasserstein distance or entropically regularized optimal transport are required [1, 2]. Extending the present tail-focused theory and scalable algorithms to these settings is another important direction.

**Comment #2.** *The use of the Wasserstein Distance needs to be explained more systematically. While it is true that it serves as a useful statistical regularization, the advantages and disadvantages should be discussed. What will be the conditions that make this regularization particularly suitable for extreme events? What about K-L divergence and other statistical metrics? Even qualitative comparisons or discussions can be very helpful in supporting the use of the method proposed here.*

**Response.** We thank the reviewer for this helpful suggestion. We agree that the comparison with other divergences/metrics could be made more systematic. We have revised the Discussion subsection "Practical Advantages of $W_1$" to clarify the comparison specifically with KL divergence and the $L^1$-log metric, which are the most relevant alternatives for the present setting.

In particular, for the usual forward KL form

$$D_{\mathrm{KL}}(p\|q) = \int p(y) \log \frac{p(y)}{q(y)}\, dy,$$

tail discrepancies are weighted by the small probability density $p(y)$. Therefore, even large discrepancies in low-probability regions may contribute weakly to the objective. The $L^1$-log metric, defined as

$$\int_{\mathcal{Y}} |\log p(y) - \log q(y)|\, dy.$$

mitigates this issue by comparing log densities directly and therefore gives stronger emphasis to tail discrepancies. However, this advantage comes with a practical limitation: the metric requires

reliable density estimates and becomes numerically unstable when either density is zero or extremely small. In particular, when the supports of the learned observable distribution and the reference distribution do not fully overlap, the log-density discrepancy diverges. This issue is especially common in the data-scarce regime considered here, where the learned model may initially assign little or no probability mass to extreme values. In such cases, KL-based objectives and the $L^1$-log metric require additional numerical devices to remain well defined.

In contrast, $W_1$ avoids this support-mismatch obstruction: for probability measures with finite first moments, it gives a finite and stable discrepancy even when the learned and reference observable laws have poorly overlapping tails. A $W_1$-based regularizer thus provides a stable training signal in this regime. The practical usefulness of $W_1$ also comes from the fact that, in one dimension, $W_1$ admits the quantile representation

$$W_1(\nu_1, \nu_2) = \int_0^1 \left| F_{\nu_1}^{-1}(q) - F_{\nu_2}^{-1}(q) \right| \, dq, \tag{1}$$

which makes the discrepancy straightforward to estimate from samples. More importantly, this representation naturally allows the numerical approximation to be concentrated on the tail; see Methods for details. As a result, $W_1$ allows for a numerically robust mechanism for making the regularization tail-aware.

The revised subsection now reads as follows.

### Practical Advantages of $W_1$

Aside from its theoretical optimality for controlling tail errors, $W_1$ offers several practical advantages over alternatives such as KL divergence and the $L^1$-log metric [3]. The key issue with the usual forward KL divergence,

$$D_{\mathrm{KL}}(p\|q) = \int p(y) \log \frac{p(y)}{q(y)} \, dy,$$

is that discrepancies in the tail are weighted by $p(y)$, which is small precisely in rare-event regions. Consequently, even substantial disagreement between $p$ and $q$ in the tail can contribute weakly to the total KL objective. In the present setting, this is undesirable because the main goal to recover statistically meaningful behavior in low-probability extreme regions.

The $L^1$-log metric mitigates this issue by comparing log densities directly. It is therefore more explicitly tail-aware and could in principle serve as a regularizer for $\eta$-learning. However, this advantage comes with a practical limitation: the metric requires reliable density estimates and becomes numerically unstable when either density is zero or extremely small. In particular, when the supports of the learned observable distribution and the reference distribution do not fully overlap, the log-density discrepancy diverges. This issue is especially common in the data-scarce regime considered here, where the learned model may initially assign little or no probability mass to extreme values. In such cases, KL-based objectives and the $L^1$-log metric require additional numerical devices to remain well defined.

In contrast, $W_1$ avoids this support-mismatch obstruction: for probability measures with finite first moments, it gives a finite and stable discrepancy even when the learned and reference observable laws have poorly overlapping tails. A $W_1$-based regularizer thus provides a stable training signal in this regime. The practical usefulness of $W_1$ also comes from the fact that, in one dimension, $W_1$ admits the quantile representation

$$W_1(\nu_1, \nu_2) = \int_0^1 \left| F_{\nu_1}^{-1}(q) - F_{\nu_2}^{-1}(q) \right| \, dq, \tag{2}$$

which makes the discrepancy straightforward to estimate from samples. More importantly, this representation naturally allows the numerical approximation to be concentrated on the tail; see Methods for details. As a result, $W_1$ allows for a numerically robust mechanism for making the regularization tail-aware.

**Comment #3.** *Following the above question, the Wasserstein Distance is a statistical regularization, which is also the emphasis of this paper. However, how does the method, with such a statistical regularization, help learn the dynamics of extreme events? In principle, with only statistical information, the mechanisms of the extreme events generated can be very different from the truth. This is certainly not the case with the numerical results. Therefore, something behind the scenes must play some roles here. Adding more reasoning and even potential caveats will be very necessary.*

**Response.** We thank the reviewer for raising this important point. We agree that the $W_1$ penalty is a statistical regularization and, by itself, does not identify or learn the dynamical mechanism that generates extreme events.

In the precipitation example, the dynamical and structural information enters through the supervised component of the learning problem. The dataset contains low-resolution (LR) precipitation snapshots over the full period $t \in [0, T]$, with $T = 25$ years, together with paired high-resolution (HR) snapshots over the first 0.5 years. Each LR snapshot provides a coarse representation of the corresponding physical state, while the LR–HR pairs provide labeled spatial correspondence for learning the downscaling map. The $W_1$ term does not add new mechanistic information; rather, it constrains this supervised map so that its induced observable statistics match the prescribed tail law.

Therefore, the method should be interpreted as a distributionally constrained supervised-learning framework, not as a mechanism-discovering method based on statistical information alone. If distinct mechanisms induce the same scalar observable law, the present formulation cannot distinguish them without additional identifying information, such as temporal or spatial structures and physics-informed constraints. In fact, one promising future direction is to develop learning frameworks that jointly infer the governing mechanisms while enforcing consistency with the prescribed extreme-event statistics.

We have incorporated an explicity caveat in the result analysis for the vanilla precipitation downscaling example. Line 706 now reads:

It is also crucial to note that the dynamical structure in the generated extreme events does not come from the Wasserstein regularization alone. Rather, it is inherited from the supervised problem: each LR snapshot provides a coarse representation of the corresponding physical state, while the paired LR–HR snapshots over the first 0.5 years provide labeled spatial correspondence.

The limitation that $\eta$-Learning does not identify the dynamical mechanism of extreme events and the correpsonding future direction are discussed in detail in the newly added subsection "Limitations and Future Work" at line 1038:

A central limitation is that matching the distribution of a prescribed observable does not identify the dynamical mechanism that generates extreme events. The $W_1$ regularizer constrains observable statistics, but does not by itself determine the temporal evolution, spatial organization, or broader physical coherence of the generated extremes. To the extent that these properties are recovered, they are inherited from the supervised input-output data rather than supplied by the $W_1$ term.

Consequently, when distinct mechanisms or state configurations induce the same observable law, they are not distinguishable under the present objective. An important direction is therefore to extend the current framework with temporal and spatial constraints, conservation laws, or other physical priors.

**Comment #4.** *Is there a way to automatically choose the hyperparameters to make the algorithm more objective? Otherwise, some additional prior knowledge is needed, which can sometimes complicate the problem.*

**Response.** We thank the reviewer for raising this important point. In the present framework, there are two types of hyperparameters. Some parameters, such as the quantile set $\mathcal{Q}$ or the cutoff $\tau$, specify which part of the observable distribution is regarded as the tail region of interest. These necessarily encode the target rare-event probability level and are therefore tied to the scientific or engineering objective; in other words, they are more application-specific. The balancing parameter $\lambda$, however, plays a different role: it controls the relative strength of the supervised ERM loss and the $W_1$ tail correction. We agree with the reviewer that it will be very desirable for $\lambda$ not to require delicate problem-specific tuning.

To address this concern, we have added a more explicit discussion of the robustness of our method with respect to $\lambda$. As reported in SI Fig. 4, the learned observable distribution remains stable as $\lambda$ varies from $10^{-4}$ to 10, while all other training configurations are fixed. This indicates that the method is numerically robust to the choice of $\lambda$ over a wide range. Thus, at least in the tested settings, $\lambda$ does not need to be finely tuned in order for the method to recover the target tail behavior.

This robustness is closely tied to the effective training strategies used in the algorithm. In particular, we do not introduce the $W_1$ correction from an untrained model. Instead, we first train the model using only the supervised ERM loss, and then initialize both the $\eta$-learning stage and the initial tail set used to approximate the $W_1$ penalty from the pre-trained supervised estimator. After this step, the model-induced ordering used to select tail samples is anchored to a predictor that already captures the observed input-output relation. Consequently, the $W_1$ term primarily acts as a residual tail correction around an ERM-consistent solution, rather than determining the entire learned map from scratch. The periodic tail-set refresh further keeps the $W_1$ correction aligned with the evolving observable distribution. This is why the method is much less sensitive to the precise value of $\lambda$ than a direct, unstaged optimization of the combined ERM+$W_1$ objective would be.

Nevertheless, to make the selection of $\lambda$ more objective in practice, we have added a gradient-scale balancing rule. Since

$$\nabla_\theta L(\theta) = \nabla_\theta L_{\mathrm{ERM}}(\theta) + \lambda \nabla_\theta L_{W_1}(\theta),$$

a potentially beneficial choice is to select $\lambda$ so that the scaled $W_1$ gradient has comparable Euclidean norm to the ERM gradient, namely

$$\lambda \|\nabla_\theta L_{W_1}(\theta)\|_2 = \|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2 \implies \lambda = \frac{\|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2}{\|\nabla_\theta L_{W_1}(\theta)\|_2 + \varepsilon},$$

with a small $\varepsilon > 0$. This provides an objective scale for $\lambda$, while the robustness study shows that the method does not require delicate tuning once the staged training procedure is used.

We have revised the main text to make these points explicit. In particular, we now explain in

subsection "Effective Training Strategies" that $\lambda$ is a secondary balancing parameter, while ERM pre-training and tail-set refresh are the primary mechanisms responsible for stable optimization:

### Effective Training Strategies

Accurately estimating tail quantiles at every iteration can be memory- and compute-intensive, because it naively requires evaluating the model on a large auxiliary sample set $\tilde{x} \sim \mu$ at each update. To mitigate this, we use an inference-informed continual training strategy: periodically, we run inference on the auxiliary set to identify the inputs whose current observable values realize the target quantiles in $\mathcal{Q}$, and then restrict subsequent $W_1$ estimations to these identified inputs until the next refresh. This preserves a reliable estimate of the tail loss while substantially reducing per-iteration cost. A detailed computational complexity analysis of the full procedure is provided in SI Subsection "Computational Complexity".

This tail-set selection procedure also creates a potential optimization issue. The $W_1$ term is evaluated on inputs selected according to the current model-induced observable distribution $(g \circ \phi_\theta)_{\#}\mu$. Early in training, before the supervised input-output relation has been learned, this model-induced ordering can be unreliable. As a result, samples that should correspond to bulk behavior under an ERM-consistent predictor can be spuriously selected as tail samples and pulled toward upper quantiles of $\nu_0$, while the ERM loss pulls the model toward the labeled pairs. The two loss components therefore provide conflicting gradient signals, leading to oscillatory or unstable optimization.

We mitigate this issue using a staged training procedure. We first train the model using only the supervised ERM loss, and then use the ERM estimator to initialize both the subsequent $\eta$-learning stage as well as the initial tail set used to approximate the $W_1$ penalty. After this pre-training step, the ordering used to select tail samples is anchored to a predictor that already respects the observed input-output relation. Consequently, the $W_1$ penalty primarily acts as a tail correction to an ERM-consistent map, while the ERM term continues to stabilize the learned map around the supervised solution. During the subsequent $\eta$-learning iterations, the tail set is periodically refreshed as the learned map evolves, keeping the samples used to estimate $W_1$ separated from the ERM-induced bulk. This procedure systematically mitigates the conflict caused by unreliable tail-sample selection and has empirically yielded consistently stable and robust training.

The balancing parameter $\lambda$, on the other hand, controls the strength of this residual tail correction. In practice, choosing $\lambda$ to make the $W_1$ gradient comparable in magnitude to the ERM gradient can be beneficial. However, $\lambda$ only rescales the $W_1$ contribution; it does not by itself resolve the directional incompatibility that can arise from unreliable early-stage tail-set selection. Thus, the main stabilization mechanism is ERM pre-training together with tail-set refresh, while $\lambda$ serves as a secondary balancing parameter. This interpretation aligns with the robustness study in SI Fig. 4, where $y_{\eta\#}\mu$ remains stable over a broad range of $\lambda$. A more detailed discussion of the gradient conflict and the practical choice of $\lambda$ is provided in SI Subsection "Pre-training and Inference-Informed Continual Training".

We have also revised the SI subsection "Pre-training and Inference-Informed Continual Training" to give a more detailed account of the gradient conflict, the role of tail-set selection, and the practical choice of $\lambda$:

### Pre-training and Inference-Informed Continual Training

Several difficulties arise in the actual optimization of (14). We assume $\mathcal{F}_\theta$ is a neural network optimized with a stochastic first-order optimization method $\mathcal{M}$. The update rule is denoted by

$$\phi_\theta^{(k+1)} \leftarrow \mathcal{M}\left(\phi_\theta^{(k)}, \nabla_\theta \mathcal{L}_\theta^{(k)}\right).$$

Our focus remains on (14), as the other formulation can be viewed as a special case.

***Memory Challenges and Tail-Set Selection in $W_1$ Estimation.*** A practical challenge stems from approximating the $W_1$ term, even with its tractable one-dimensional quantile representation. Accurate estimation requires evaluating $g \circ \phi_\theta$ over the auxiliary set $\mathcal{D}_{\tilde{x}}$, which is typically much larger than the labeled training set. If this full auxiliary evaluation were included in every gradient-carrying update, the memory cost would be substantial.

To address this, we note that only $n_q$ specific inputs from $\mathcal{D}_{\tilde{x}}$ are ultimately needed to compute the empirical quantile loss, namely the inputs whose current observable values realize the quantile levels in $\mathcal{Q}$. This motivates the Inference-Informed Continual Training (IICT) procedure. Periodically, we run inference on the full auxiliary set $\mathcal{D}_{\tilde{x}}$ and compute

$$\{g(\phi_\theta(\tilde{x}_j))\}_{j=1}^{n_{\tilde{x}}}.$$

We then identify the indices of the inputs corresponding to the prescribed quantile levels in $\mathcal{Q}$ and store them in a set $\mathcal{I}$. Between refreshes, the $W_1$ loss is evaluated only on the $\mathcal{I}$-indexed inputs. Thus, full inference on $\mathcal{D}_{\tilde{x}}$ is used only to identify the relevant tail samples, while the gradient-carrying $W_1$ update is restricted to the selected quantile samples. The refresh frequency is controlled by a hyperparameter $\omega$, with $\mathcal{I}$ updated every $\omega$ iterations. This procedure substantially reduces memory usage while preserving an accurate estimate of the tail contribution to the $W_1$ loss.

***Gradient Conflict and ERM Pre-training.*** The tail-set selection procedure also explains a potential source of optimization instability. The $W_1$ loss is evaluated on inputs selected according to the current model-induced observable distribution $(g \circ \phi_\theta)_{\#}\mu$. Early in training, before the supervised input-output relation has been learned, the ordering of the auxiliary observable values $g(\phi_\theta(\tilde{x}_j))$ can be unreliable. Consequently, inputs that should correspond to bulk behavior under an ERM-consistent predictor may be spuriously selected into $\mathcal{I}$ and treated as high-quantile samples. The $W_1$ term then pulls these samples toward upper quantiles of the reference distribution $\nu_0$, while the ERM loss pulls the model toward the labeled input-output relation. These two signals can be incompatible, leading to oscillatory or unstable optimization.

Our main stabilization mechanism is ERM pre-training. We first minimize the ERM loss alone and initialize $\phi_\theta$ with the resulting ERM estimator $u_\xi$ when solving (14). The initial index set $\mathcal{I}$ used by the $W_1$ penalty is also computed from this ERM-pretrained model. Thus, the first $W_1$ correction is applied to quantile samples selected under a predictor that already captures the observed input-output relation. This substantially reduces the chance that bulk samples are spuriously treated as tail samples because of unreliable early predictions. During $\eta$-learning, $\mathcal{I}$ is periodically refreshed as $\phi_\theta$ evolves, keeping the $W_1$ correction aligned with the current observable distribution.

***Choice of $\lambda$.*** The balancing parameter $\lambda$ controls the relative magnitude of the ERM and $W_1$ gradient components. Writing

$$\mathcal{L}_\theta = L_{\mathrm{ERM}}(\theta) + \lambda L_{W_1}(\theta),$$

we have

$$\nabla_\theta \mathcal{L}_\theta = \nabla_\theta L_{\mathrm{ERM}}(\theta) + \lambda \nabla_\theta L_{W_1}(\theta).$$

A practical scale for $\lambda$ is obtained by choosing it so that the scaled $W_1$ gradient has a comparable Euclidean norm to the ERM gradient during the early $\eta$-learning iterations:

$$\lambda \|\nabla_\theta L_{W_1}(\theta)\|_2 = \|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2 \implies \lambda = \frac{\|\nabla_\theta L_{\mathrm{ERM}}(\theta)\|_2}{\|\nabla_\theta L_{W_1}(\theta)\|_2 + \varepsilon},$$

with a small $\varepsilon > 0$ for numerical stability. This prevents the $W_1$ correction from overwhelming the supervised anchor.

However, $\lambda$ only rescales the $W_1$ gradient; it does not change its direction. Therefore, while magnitude balancing could be potentially beneficial, it alone cannot resolve the directional incompatibility caused by unreliable tail-set selection. For this reason, ERM pre-training and periodic tail-set refresh are the primary mechanisms for stable optimization, while $\lambda$ serves as a secondary balancing parameter that controls the strength of the residual tail correction. This interpretation is consistent with the robustness experiment in SI Fig. 4, where the learned observable distribution remains stable as $\lambda$ varies from $10^{-4}$ to 10 while all other training configurations are fixed.

**Comment #5 .** *One thing I feel very confused about is the toy model examples. I do see you get extreme events. However, the locations and amplitudes differ from the truth, e.g. in Fig 1. While this can still give you the right spatiotemporal statistics, I wonder what the target is here. I guess, in addition to sampling extreme events across the domain, in SI Fig 2, what is more important is to get the point-wise statistics, not the spatiotemporal average. It is also practically useful to predict the locations and times of extreme events. This relates to my previous question about the dynamical mechanisms that generate extreme events.*

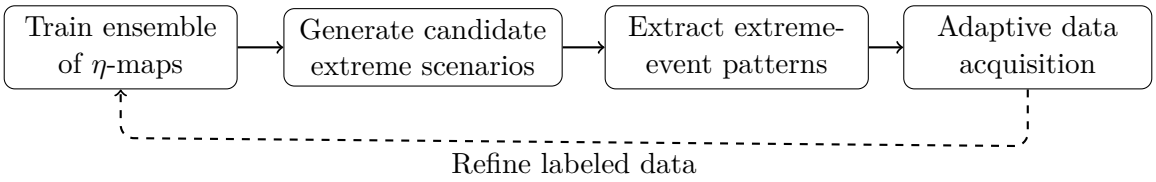
**Response.** We thank the reviewer for this question and agree that this point should be further clarified. The target of the toy examples, and of $\eta$-Learning in general, is not exact recovery of the *unobserved* extreme event location, amplitude, or time from training data that contain no such events. Such recovery cannot be expected from conventional supervised methods without extreme-event data, as shown in Theorem 4, and is non-identifiable from the scalar observable distribution alone.[1] Rather, the target is to learn maps that remain accurate in the data-rich region while matching the prescribed tail law, thereby generating *plausible candidate extremes* that are statistically consistent with the observable tail.

This distinction is important for interpreting Fig. 1 and SI Fig. 2. In these examples, the different peak locations obtained across independent $\eta$-maps should not be interpreted as failures to recover a hidden truth. They instead represent the residual spatial uncertainty that remains after enforcing the scalar tail statistics. If point-wise statistics are the intended target, then the corresponding point-wise or spatially conditioned observables should be included in the statistical constraint. With only a scalar extreme observable, $\eta$-Learning constrains the tail behavior of that observable, but it does not uniquely determine the full spatial arrangement of the corresponding extreme field.

One promising downstream use of the generated scenarios is precisely to expose this residual uncertainty in a systematic and algorithmic way. In applications where additional simulations or experiments can be performed, one can train an ensemble of $\eta$-maps, generate candidate extreme

[1] An exception is when Assumption 5 holds, which then would imply Theorem 6. However, enforcing Assumption 5 will likely require more information than the observable distribution. See line 443 for more discussion on this point.

scenarios, extract common patterns according to extreme-event time, location, or other physically meaningful features, and then use the resulting representatives to guide adaptive data acquisition. The pipeline is summarized in the following illustration.




The role of $\eta$-Learning in this workflow is therefore not to replace mechanistic validation, but to *propose* statistically consistent extreme-event hypotheses in regions where the original data provide no direct evidence. Subsequent high-fidelity simulations, experiments, or additional observations can then test, refine, or reject these candidate modes.

We have revised the result analysis for the 2D-to-1D toy example in the main text at line 595 to read:

The ensemble analysis highlights a broader implication of $\eta$-learning: enforcing scalar observable tail statistics generates candidate extremes, but does not necessarily identify the underlying physical structure, such as the peak location shown in this example. This variability reflects the residual physical uncertainty that remains after enforcing the observable tail statistics, rather than a failure to recover a uniquely identifiable hidden event. [...] In the Discussion section, we further describe how this uncertainty can be systematically characterized and used in downstream applications.

Moreover, we have added a detailed account of the downstream algorithmic pipeline outlined above in the newly added subsection "Limitations and Future Work" at line 1048:

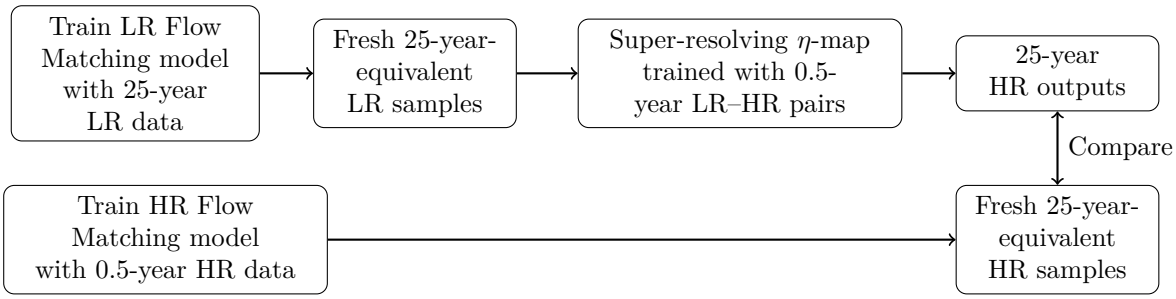
This non-identifiability also arises at the level of individual events: the prescribed observable law does not generally determine the precise location, occurrence time, or full spatial structure of an unobserved extreme. Variability in these features across independently trained $\eta$-maps should therefore be interpreted as residual uncertainty. Importantly, this uncertainty can be exploited systematically. Ensembles of $\eta$-maps can generate diverse yet statistically consistent candidate extremes, which can then be organized according to physically meaningful attributes such as event location, occurrence time, or spatial pattern. Representative scenarios can guide subsequent rounds of adaptive data acquisition, such as targeted high-fidelity simulations or experiments, allowing the corresponding hypotheses to be tested, refined, or rejected while progressively reducing uncertainty in data-scarce regions.

**Comment #6 .** *I also noticed that, for example, in Figure 6 and maybe Figure 5 as well, the estimated tail is above the true tail. Is this because of the over-regularization? Is there a general criterion to mitigate this, or does it need a few trials to get the optimal solution?*

**Response.** We thank the reviewer for raising this concern. We clarify two distinct cases.

For Fig. 5, the curve labeled "HR $\eta$-Generated Samples (0.5-Yr Training)" is produced by a two-stage generative pipeline: we train an LR Flow Matching model using 25 years of LR data, draw a fresh 25-year set of LR samples from the trained model, pass those LR samples through the super-resolving $\eta$-map trained with 0.5 years of LR–HR pairs (i.e., the map that produces Fig. 4), and compute the observable distribution of the resulting HR outputs. The workflow is summarized

schematically as



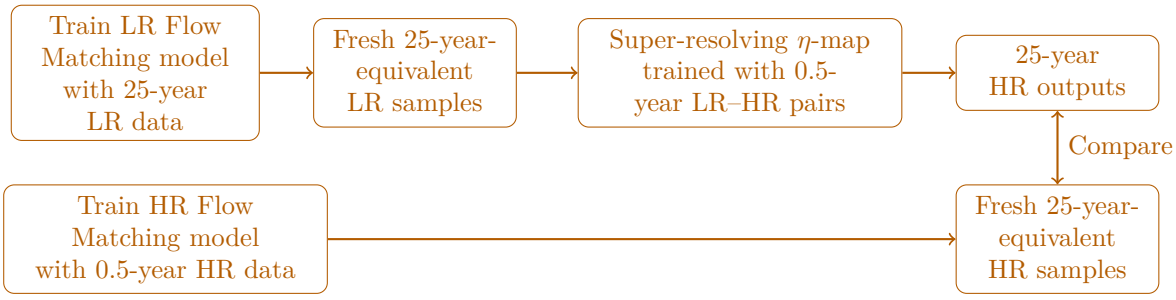

Thus, the star curve in Fig. 5 is not obtained by applying $u_\eta$ to the fixed empirical 25-year LR dataset. It is the observable distribution obtained from a finite set of samples generated by the LR Flow Matching model and subsequently post-processed by $u_\eta$. Therefore, the apparent tail elevation in Fig. 5 should be interpreted as uncertainty propagated through the low-resolution generative pipeline instead of over-regularization of $\eta$-learning.

We have added the following clarification in the main text at line 827:

The tail elevation relative to the ground truth reflects uncertainty from drawing a fresh finite 25-year LR sample from the LR FM model, rather than direct evidence of over-regularization of the $\eta$-map.

To faciliate the exposition, we have also added the schematic description to the SI in section "Experimental Setups and Details" under subsection "Correcting Statistical Biases of Deep Generative Models":

**Experimental Setup.** The experimental setup is summarized schematically below. We first train a low-resolution (LR) Flow Matching (FM) model using the full 25-year LR dataset. We then draw a fresh 25-year-equivalent collection of LR samples from this trained LR FM model and pass these samples through the super-resolving $\eta$-map trained with only 0.5 years of paired LR–HR data. As a baseline, we also train a high-resolution (HR) FM model using only the available 0.5-year HR dataset and draw a fresh 25-year-equivalent collection of HR samples from it. The observable distributions of these generated samples are then compared against the empirical 25-year HR observable distribution.




This workflow is also referenced in the main text at line 816:

A schematic description of this workflow is provided in the SI.

For Fig. 6, the interpretation is different. In the hypothesized GEVD experiment, the prescribed reference law $\nu_0$ is intentionally chosen to be heavier-tailed than the empirical 25-year HR observable

distribution. Therefore, the resulting heavier-tailed $y_{\eta\#}\mu$ is expected: the $\eta$-map is designed to be faithful to the prescribed reference law. This experiment serves as a prior-conditioned stress test rather than as an estimate of the true climatological observable distribution. Practically, this setting can represent, for example, an hypothetical adverse climate-tail scenario in climate change, where the goal is to generate statistically consistent extreme precipitation events for downstream analysis under a specified tail hypothesis.

We have revised the clarification in the main text at line 913:

We emphasize that here, $\nu_0$ is a deliberately heavier-tailed reference law used to explore what HR precipitation fields are statistically consistent with the paired LR-HR training data under a specified adverse-tail hypothesis.

**Comment #7 .** *Did you consider the case with observational noise? Can you filter the noise at the same time, so you will not treat the noisy observation as the truth when applying your method?*

**Response.** We thank the reviewer for raising this point. The present formulation is a supervised-learning framework rather than a filtering or data-assimilation framework. In particular, the setup does not assume an explicit latent dynamical model, observation operator, or measurement-noise model. Therefore, Kalman-type filters, particle filters, smoothing methods, and related data-assimilation procedures are not generic components of the proposed algorithm. If such additional structure is available in a particular application, these methods, or bias-correction procedures more generally, can be used as a *data-preprocessing* step to construct denoised training targets before applying $\eta$-Learning. The precipitation example is consistent with this viewpoint: we use reanalysis precipitation fields as the data source, so bias correction belongs to the construction of the training data rather than to $\eta$-Learning itself.

We have clarified the role of observational noise in Problem Statement at line 195:

We study the data-driven setting where we have access to a dataset $\mathcal{D} = \{(x_i, u_i, y_i)\}_{i=1}^n$, where $u_i = u(x_i)$ and $y_i = g(u_i)$. When the raw measurements used to construct $u_i$ may be noisy, $u_i$ should be understood as the processed training target obtained after any application-specific data-assimilation or bias-correction step. Such preprocessing requires additional assumptions, such as a measurement-noise model or an observation operator, and is therefore not a generic component of our setup.

A separate issue, more specific to $\eta$-learning, is the noise or misspecification in the reference law $\nu_0$. Since the $W_1$ term is designed to enforce agreement with the supplied $\nu_0$, distributional uncertainty in $\nu_0$, especially in the tail, can be transferred to the generated extremes. We have therefore added a sensitivity analysis for shifted reference laws and discussed distributionally robust extensions in which a single reference law is replaced by an ambiguity set of plausible tail laws in 'Limitations and Future Work". It reads as follows.

Another limitation is the reliance on an accurate reference law $\nu_0$. Because the tail-emphasized $W_1$ estimation is designed to match the supplied tail, misspecification of $\nu_0$, or a shift in the target tail relative to $\nu_0$, can be transferred directly to the generated extremes. A promising extension is a distributionally robust formulation that replaces a single reference law with an ambiguity set of plausible laws, with particular emphasis on uncertainty and shifts in the tail. Worst-case or uncertainty-weighted tail objectives could reduce overcommitment to any single hypothesized distribution while preserving awareness of extreme behavior.

## References


[1] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. doi:10.1007/s10851-014-0506-3.

[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

[3] Mustafa A. Mohamad and Themistoklis P. Sapsis. Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 115(44):11138–11143, 2018. doi:10.1073/pnas.1813263115.

[4] D. A. Levin and Y. Peres. *Markov Chains and Mixing Times*. 2nd ed. American Mathematical Society, Providence, RI, 2017. With contributions by E. L. Wilmer. doi:10.1090/mbk/107.

[5] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wissenschaften, Vol. 260. Springer, Berlin, Heidelberg, 3rd revised and enlarged edition, 2012. doi:10.1007/978-3-642-25847-3.