Supplementary Information: *Extreme Event Aware (η-) Learning*

Kai Chang[1,2] and Themistoklis P. Sapsis[1*]

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, 02139, MA, USA.
[2]Center for Computational Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, 02139, MA, USA.

*Corresponding author(s). E-mail(s): sapsis@mit.edu;
Contributing authors: kaichang@mit.edu;

Supplementary Methods

Supplementary Figures

Supplementary Tables

Supplementary Methods

To improve readability, we restate all theorems and assumptions in the main text below and refer to their corresponding numbering here, rather than the numbering used in the main text, unless explicitly stated otherwise.

1 Proof of Theorem 4 of the Main Text

In this section, we provide all the intermediate results and proofs used to establish Theorem 4 of the main text. We first recall the definition of a data-consistent estimator.

Definition S1 (Data-Consistent Estimator) Let y_ξ be an estimator selected from some function class to fit the dataset $\mathcal{D}$. Let S be some smoothness parameter of the true mapping y. Recall that E denotes the extreme set that satisfies the inequality.

$$0 < \mathbb{P}\{X \in E\} = \mathbb{P}\{Y \geq t_*\} \leq \delta, \tag{1}$$

We say that y_ξ is a data-consistent estimator if it satisfies either (or both) of the following conditions (i) and (ii). (i) When $\mathbf{x}_i \notin E$ for any i, we have

$$\left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \mathrm{d}\mu \right| \leq \tilde{C}(n, E, \mathcal{F}, S) \cdot \left| \int_E \left(y - y_\xi\right) \mathrm{d}\mu \right|, \tag{2}$$

where

$$\tilde{C} := \inf_{C \geq 0} \left\{ \left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \mathrm{d}\mu \right| \leq C \cdot \left| \int_E \left(y - y_\xi\right) \mathrm{d}\mu \right| \right\} < 1. \tag{3}$$

(ii) When $\mathbf{x}_i \notin E$ for any i, we have

$$\int_{\mathcal{X}\setminus E} \left|y - y_\xi\right| \mathrm{d}\mu \leq \hat{C}(n, E, \mathcal{F}, S) \cdot \int_E \left|y - y_\xi\right| \mathrm{d}\mu, \tag{4}$$

where

$$\hat{C} := \inf_{C \geq 0} \left\{ \int_{\mathcal{X}\setminus E} \left|y - y_\xi\right| \mathrm{d}\mu \leq C \cdot \int_E \left|y - y_\xi\right| \mathrm{d}\mu \right\} < \infty. \tag{5}$$

Here, $\tilde{C}$ and $\hat{C}$ are some constants possibly depending on the function class $\mathcal{F}$, the smoothness of the true mapping (characterized by the parameter S), the number of data points n, and the extreme set E.

Our goal is to characterize when and how an estimator fails to accurately capture the output statistics. One natural approach is to establish a lower bound on the distance between the push-forward measures, thereby demonstrating that the two distributions must be substantially different. The W_1 distance is especially well-suited for this purpose due to its duality property, as formalized in the following theorem.

Theorem S2 (Kantorovich Duality of W_1 (Theorem 1.21 in [1])) Define $\mathrm{Lip}_1 = \{f : \mathcal{Y} \to \mathbb{R} | f \text{ is 1-Lipschitz}\}$. Then we have

$$W_1(\nu_1, \nu_2) = \sup_{f \in \mathrm{Lip}_1} \left\{ \int f \, \mathrm{d}\nu_1 - \int f \, \mathrm{d}\nu_2 \right\}. \tag{6}$$

The Kantorovich Duality offers a convenient way to derive a lower bound on the W_1 distance between two probability measures: taking any 1-Lipschitz function and substituting f in (6) with it, the resulting quantity will be a suitable lower bound. However, plugging $y_{\xi\#}\mu$ and $y_{\#}\mu$ into (6) will generally make them appear as measure differentials in the integral. This is not ideal: since our goal is to relate the comparison of the two push-forward measures to the two mappings, it is crucial that y_ξ and y appear in the integrand instead of the measure differential. To do this, we take advantage of the following lemma.

Lemma S3 Take $h : \mathcal{X} \to \mathcal{Y}$ as a $\mathcal{B}(\mathcal{X})$-measurable function with respect to $\mathcal{B}(\mathcal{Y})$. Let μ be a probability measure on $\mathcal{X}$. Suppose $f : \mathcal{Y} \to \mathbb{R}$ is an $L^1(h_{\#}\mu)$ function. Then,

$$\int_{\mathcal{Y}} f \, \mathrm{d}h_{\#}\mu = \int_{\mathcal{X}} f \circ h \, \mathrm{d}\mu.$$

We give a proof of this seemingly simple lemma here, as we have not found one in a standard measure theory textbook [2].

Proof Our proof follows a standard technique used in real analysis [2]: we start with the simple case where f is assumed to be a characteristic function, move on with a generic positive function, and in the end generalize to an arbitrary measurable function. We begin by analyzing the scenario where $f = \mathbb{1}_B$ for some measurable set $B \subset \mathcal{Y}$, where

$$\mathbb{1}_B(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in B \\ 0, & \text{otherwise} \end{cases}.$$

Define $h^{-1}(B) = \{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) \in B\}$. By definition, we have

$$\mathbb{1}_{h^{-1}(B)}(\mathbf{x}) = 1 \iff \mathbb{1}_B(h(\mathbf{x})) = 1.$$

Therefore,

$$\int_{\mathcal{Y}} f \, \mathrm{d}h_{\#}\mu = h_{\#}\mu(B) = \mu(h^{-1}(B)) = \int_{\mathcal{X}} \mathbb{1}_{h^{-1}(B)}(\mathbf{x}) \, \mathrm{d}\mu,$$

which can be further rewritten as:

$$\int_{\mathcal{X}} \mathbb{1}_{h^{-1}(B)}(\mathbf{x}) \, \mathrm{d}\mu = \int_{\mathcal{X}} \mathbb{1}_B(h(\mathbf{x})) \, \mathrm{d}\mu = \int_{\mathcal{X}} f \circ h \, \mathrm{d}\mu,$$

which is the desired result.

Building on this, we further consider

$$f = \sum_{i=1}^{n} b_i \cdot \mathbb{1}_{B_i}, \tag{7}$$

where $b_i \in \mathbb{R}$ and each $B_i \subset \mathcal{Y}$ is measurable. By linearity, we have

$$\int_{\mathcal{Y}} f \, \mathrm{d}h_{\#}\mu = \int_{\mathcal{Y}} \sum_{i=1}^{n} b_i \cdot \mathbb{1}_{B_i} \, \mathrm{d}h_{\#}\mu = \sum_{i=1}^{n} b_i \int_{\mathcal{Y}} \mathbb{1}_{B_i} \, \mathrm{d}h_{\#}\mu.$$

Applying our earlier argument for each $\mathbb{1}_{B_i}$, we get

$$\int_{\mathcal{Y}} f \, \mathrm{d}h_{\#}\mu = \sum_{i=1}^{n} b_i \int_{\mathcal{X}} \mathbb{1}_{B_i} \circ h \, \mathrm{d}\mu = \int_{\mathcal{X}} f \circ h \, \mathrm{d}\mu.$$

Now, consider any $f \in L^1(h_{\#}\mu)$ with $f \geq 0$. Such a function can be approached as the pointwise limit of a sequence of functions $\{f_m\}$, where $f = \lim_{m\to\infty} f_m$ and each f_j is of the form shown in (7) [2]. By the monotone convergence theorem, we deduce that

$$\begin{aligned}\int_{\mathcal{Y}} f \,\mathrm{d}h_{\#}\mu &= \lim_{m\to\infty} \int_{\mathcal{Y}} f_m \,\mathrm{d}h_{\#}\mu \\ &= \lim_{m\to\infty} \int_{\mathcal{X}} f_m \circ h \,\mathrm{d}\mu \\ &= \int_{\mathcal{X}} f \circ h \,\mathrm{d}\mu.\end{aligned}$$

Finally, take $f \in L^1(h_{\#}\mu)$ as any function. We express f as the difference of its positive and negative parts: $f = f^+ - f^-$, where $f^+, f^- \geq 0$ and $f^+, f^- \in L^1(h_{\#}\mu)$. Since we have already shown the result for non-negative functions, the linearity of integration allows us to conclude:

$$\int_{\mathcal{Y}} f \,\mathrm{d}h_{\#}\mu = \int_{\mathcal{X}} f \circ h \,\mathrm{d}\mu.$$

Thus, the proof is complete.

With Theorem S2 and Lemma S3, we are able to prove the following lemma, which gives a lower bound on the 1-Wasserstein distance between one-dimensional push-forward measures with the absolute difference between the forward mappings appearing as an integrand in the bound.

Lemma S4 Let y_1 and y_2 be two functions in $L^1(\mu)$, mapping from $\mathcal{X}$ to $\mathcal{Y}$. Suppose that they both are $\mathcal{B}(\mathcal{X})$-measurable with respect to $\mathcal{B}(\mathcal{Y})$ and that the input of both of them follows the distribution μ. Then we have

$$W_1(y_{1\#}\mu, y_{2\#}\mu) \geq \left| \int_{\mathcal{X}} (y_1 - y_2) \mathrm{d}\mu \right|.$$

Proof Take $f_0 : \mathcal{Y} \mapsto \mathbb{R}$ as the identity map: that is, $f_0(x) = x$ for all $x \in \mathcal{Y}$. Then it is trivial to check that $f_0 \in \mathrm{Lip}_1$. Thus, by Theorem S2, we have

$$\begin{aligned}W_1\left(y_{1\#}\mu, y_{2\#}\mu\right) &= \sup_{f\in\mathrm{Lip}_1} \left\{ \int_{\mathcal{Y}} f \,\mathrm{d}y_{1\#}\mu - \int_{\mathcal{Y}} f \,\mathrm{d}y_{2\#}\mu \right\} \\ &\geq \int_{\mathcal{Y}} f_0 \,\mathrm{d}y_{1\#}\mu - \int_{\mathcal{Y}} f_0 \,\mathrm{d}y_{2\#}\mu.\end{aligned}$$

Applying Lemma S3, it follows that

$$\begin{aligned}W_1\left(y_{1\#}\mu, y_{2\#}\mu\right) &\geq \int_{\mathcal{Y}} f_0 \,\mathrm{d}y_{1\#}\mu - \int_{\mathcal{Y}} f_0 \,\mathrm{d}y_{2\#}\mu \\ &= \int_{\mathcal{X}} f_0 \circ y_1 \,\mathrm{d}\mu - \int_{\mathcal{X}} f_0 \circ y_2 \,\mathrm{d}\mu \\ &= \int_{\mathcal{X}} (y_1 - y_2) \,\mathrm{d}\mu.\end{aligned}$$

By symmetry and that y_1 and y_2 are one-dimensional, we arrive at the desired conclusion.

Lemma S4 was first explored and proved in [3] with exactly the same proof we independently develop and present here. The problem setting and consideration therein were not related to extreme events. Moreover, we give a different proof that applies to the general p-Wasserstein distance and does not rely on the Kantorovich duality below in Section 5.

The remainder of the analysis focuses on deriving key statistical properties of a data-consistent estimator. Importantly, under the IID assumption, the probability of encountering data in the extreme set E can be explicitly calculated. With these elements established, we are now equipped to state and prove a lower bound on $W_1\left(y_{\#}\mu, y_{\xi\#}\mu\right)$ for a data-consistent estimator. It is formalized in the theorem below.

Theorem S5 (Theorem 4 of the main text.) Suppose both y and y_ξ are in $L^1(\mu)$ and are $\mathcal{B}(\mathcal{X})$-measurable with respect to $\mathcal{B}(\mathcal{Y})$. Then for a data-consistent estimator y_ξ in the sense of (2), when $n \leq \frac{\log p}{\log(1-\delta)}$, with probability at least p, we have

$$W_1\left(y_{\#}\mu, y_{\xi\#}\mu\right) \geq \left(1 - \tilde{C}(n, E, \mathcal{F}, S)\right) \left| \int_E \left(y - y_\xi\right) \mathrm{d}\mu \right|, \tag{8}$$

where $\tilde{C}$ is the same as that in (2).

Proof By Lemma S4, we have

$$W_1\left(y_{\#}\mu, y_{\xi\#}\mu\right) \geq \left| \int_{\mathcal{X}} (y - y_\xi) \mathrm{d}\mu \right|.$$

Writing $\mathcal{X} = (\mathcal{X} \setminus E) \cup E$, we may rewrite the inequality above as

$$W_1\left(y_{\#}\mu, y_{\xi\#}\mu\right) \geq \left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \mathrm{d}\mu + \int_E (y - y_\xi) \mathrm{d}\mu \right| \tag{9}$$

$$\geq \left| \int_E (y - y_\xi) \mathrm{d}\mu \right| - \left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \mathrm{d}\mu \right|, \tag{10}$$

where the last inequality follows from the reverse triangle inequality.

We further realize that when $n \leq \frac{\log p}{\log 1-\delta}$, with probability at least p, $\mathbf{x}_i \notin E$ for any i. This fact can be easily proved by the IID property and the inequality (1). Therefore, for a data-consistent function family $\mathcal{F}$, by Definition S1, when $n \leq \frac{\log p}{\log 1-\delta}$, with probability at least p, we have

$$\left| \int_{\mathcal{X}\setminus E} (y - y_\xi) \mathrm{d}\mu \right| \leq \tilde{C}(n, E, \mathcal{F}, S) \cdot \left| \int_E \left(y - y_\xi\right) \mathrm{d}\mu \right|.$$

This combined with (10) gives

$$W_1\left(y_{\#}\mu, y_{\xi\#}\mu\right) \geq \left(1 - \tilde{C}(n, E, \mathcal{F}, S)\right) \left| \int_E \left(y - y_\xi\right) \mathrm{d}\mu \right|,$$

which is the desired result.

2 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 μ 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 δ, 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 fortuitously 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.

3 Proof of Theorem 6 of the Main Text

We first restate a key assumption, the explanation and intuition of which have already been provided in the main text.

Assumption S6 (Assumption 5 of the main text.) $\int_E \left|y_\xi - y\right| d\mu \leq K \cdot \left|\int_E \left(y_\xi - y\right) d\mu\right|$ for some constant $K \geq 1$.

We are now ready to prove the theorem.

Theorem S7 (Theorem 6 of the main text.) In the setting of Theorem S5, we have that if Definition S1 and Assumption S6 are uniformly satisfied as $W_1(y_{\xi\#}\mu, y_\#\mu) \to 0$, then $y_\xi \xrightarrow{\mathbb{P}} y$ on $\mathcal{X}$.

Proof We first note that

$$\begin{aligned}
\mathbb{E}\left[\left|y_\xi(\mathbf{x}) - y(\mathbf{x})\right|\right] &= \int_{\mathcal{X}} |y_\xi - y| \mathrm{d}\mu \\
&= \int_E |y_\xi - y| \mathrm{d}\mu + \int_{\mathcal{X}\setminus E} |y_\xi - y| \mathrm{d}\mu \\
&\leq (1 + \hat{C}) \int_E |y_\xi - y| \mathrm{d}\mu \\
&\leq K(1 + \hat{C}) \left| \int_E (y_\xi - y) \mathrm{d}\mu \right| \\
&\leq \frac{K(1 + \hat{C})}{1 - \tilde{C}} W_1(y_{\xi\#}\mu, y_\#\mu),
\end{aligned}$$

where the first inequality follows from (4), the second inequality follows from Assumption S6, and the third follows from Theorem S5. $\tilde{C}$ and $\hat{C}$ are defined as in (S1) and (4). Now, by Markov's inequality, we have that for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(\left|y_\xi(\mathbf{x}) - y(\mathbf{x})\right| \geq t\right) &\leq \frac{\mathbb{E}\left[\left|y_\xi(\mathbf{x}) - y(\mathbf{x})\right|\right]}{t} \\
&\leq \frac{K(1 + \hat{C})}{(1 - \tilde{C})} \cdot \frac{W_1(y_{\xi\#}\mu, y_\#\mu)}{t} \\
&\to 0,
\end{aligned}$$

where the convergence follows by the assumption that $W_1(y_{\xi\#}\mu, y_\#\mu) \to 0$. Since this holds for any t, we conclude that $y_\xi \to y$ in probability pointwise. The desired result then follows from the assumption that y_ξ and y are continuous and $\mathcal{X}$ is a Polish space.

4 Proof of Theorems 7 and 9 of the Main Text

We begin by proving a key lemma that underpins our subsequent derivations.

Lemma S8 (Exact Upper Bound on W_1) In the setting of Lemma S4, we have

$$W_1\left(y_{1\#}\mu, y_{2\#}\mu\right) \leq \int_{\mathcal{X}} |y_1(\mathbf{x}) - y_2(\mathbf{x})| \, \mathrm{d}\mu$$

Proof Recall the definition of W_1:

$$W_1\left(y_{1\#}\mu, y_{2\#}\mu\right) = \inf_{\gamma \in \Pi(y_{1\#}\mu, y_{2\#}\mu)} \int |y_1 - y_2| \, \mathrm{d}\gamma\left(y_1, y_2\right).$$

We define γ_0 as

$$\gamma_0(A \times B) = \mu\left(\{\mathbf{x} \mid y_1(\mathbf{x}) \in A, y_2(\mathbf{x}) \in B\}\right),$$

for any $A, B \in \mathcal{B}(\mathcal{Y})$. We realize that $\gamma_0 \in \Pi\left(y_{1\#}\mu, y_{2\#}\mu\right)$ because

$$\gamma_0(A \times \mathcal{Y}) = \mu\left(\{\mathbf{x} \mid y_1(\mathbf{x}) \in A\}\right) = y_{1\#}\mu(A)$$

and similarly

$$\gamma_0(\mathcal{Y} \times B) = \mu\left(\{\mathbf{x} \mid y_2(\mathbf{x}) \in B\}\right) = y_{2\#}\mu(B).$$

Since γ_0 is the joint distribution between $y_{1\#}\mu$ and $y_{2\#}\mu$ by definition, we obtain

$$\begin{aligned} W_1\left(y_{1\#}\mu, y_{2\#}\mu\right) &\leq \int |y_1 - y_2| \, \mathrm{d}\gamma_0\left(y_1, y_2\right) \\ &= \int_{\mathcal{X}} |y_1(\mathbf{x}) - y_2(\mathbf{x})| \, \mathrm{d}\mu, \end{aligned}$$

which is the desired result.

The lemma above gives an exact upper bound on the W_1 in terms of the L^1 difference in the maps. This is reminiscent of the Lemma S4. While the proof of Lemma S4 uses the dual formulation, the proof of Lemma S8 takes advantage of the primal formulation. It is precisely this connection between the two formulations that allows us to derive matching lower and upper bounds. We now prove Theorem 7 of the main text.

Theorem S9 (Theorem 7 of the main text.) In the setting of Theorem S7, we have

$$\frac{1-\tilde{C}}{K} \int_E |y_\xi - y| \mathrm{d}\mu \leq W_1(y_{\xi\#}\mu, y_{\#}\mu) \leq (1+\hat{C}) \int_E |y_\xi - y| \mathrm{d}\mu.$$

Proof First note that the lower bound directly follows from Lemma S4, Assumption S6, and the proof of Theorem S7. On the other hand, the upper bound directly follows from Theorem S8 and (4). The proof is thus complete.

Before presenting the proof of Theorem 9 of the main text, we restate two associated assumptions below.

Assumption S10a $\int_{\mathcal{X}\setminus E} |y_\xi - y| \mathrm{d}\mu \leq \varepsilon_1$.

Assumption S10b $\int_{\mathcal{X}\setminus E} |y_\xi - y_\eta| \mathrm{d}\mu \leq \varepsilon_2$.

We now prove the corresponding theorem.

Theorem S11 (Theorem 9 of the main text.) Suppose $\int_E |y_\eta - y| \, d\mu \leq K \cdot \left| \int_E (y_\eta - y) \, d\mu \right|$ for some constant $K \geq 1$. In the setting of Theorem S9, under Assumptions S10a and S10b, as $K \to 1$, we have

$$\left| W_1(y_{\eta\#}\mu, y_\#\mu) - \int_E |y_\eta - y| \mathrm{d}\mu \right| \leq \varepsilon_1 + \varepsilon_2.$$

Proof We first note that

$$\begin{aligned} W_1(y_{\eta\#}\mu, y_\#\mu) &\leq \int_{\mathcal{X}\setminus E} |y_\eta - y| \mathrm{d}\mu + \int_E |y_\eta - y| \mathrm{d}\mu \\ &\leq \int_E |y_\eta - y| \mathrm{d}\mu + \int_{\mathcal{X}\setminus E} |y_\eta - y_\xi| \mathrm{d}\mu \\ &+ \int_{\mathcal{X}\setminus E} |y_\xi - y| \mathrm{d}\mu \\ &\leq \int_E |y_\eta - y| \mathrm{d}\mu + \varepsilon_1 + \varepsilon_2, \end{aligned}$$

where the first and last inequalities follow from Theorem S8 and Assumptions S10a and S10b. For the other direction, we note that

$$\begin{aligned} W_1(y_{\eta\#}\mu, y_\#\mu) &\geq \left| \int_{\mathcal{X}\setminus E} (y_\eta - y) \mathrm{d}\mu + \int_E (y_\eta - y) \mathrm{d}\mu \right| \\ &\geq \left| \int_E (y_\eta - y) \mathrm{d}\mu \right| - \left| \int_{\mathcal{X}\setminus E} (y_\eta - y) \mathrm{d}\mu \right| \\ &\geq \left| \int_E (y_\eta - y) \mathrm{d}\mu \right| - \left| \int_{\mathcal{X}\setminus E} (y_\eta - y_\xi) \mathrm{d}\mu \right| \\ &- \left| \int_{\mathcal{X}\setminus E} (y_\xi - y) \mathrm{d}\mu \right| \\ &\geq \left| \int_E (y - y_\eta) \mathrm{d}\mu \right| - \varepsilon_1 - \varepsilon_2 \\ &\geq \frac{1}{K} \int_E |y - y_\eta| \mathrm{d}\mu - \varepsilon_1 - \varepsilon_2, \end{aligned}$$

where the first inequality follows from Theorem S5, the second inequality follows from the reverse triangle inequality, and the rest follows from the Assumptions. The result is thus established by taking $K \to 1$.

5 Theoretical Extensions

Extension to Non-IID Data.

The IID assumption in Theorem S5 is used only to evaluate the probability that the training set does not intersect the extreme set E. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a possibly

dependent sequence with common marginal law μ, and define

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

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

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

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

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

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 $(\mathbf{x}_i)$ is a stationary Markov chain with invariant law μ, $q_n(E) = \mathbb{P}_\mu(\tau_E > n)$ 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 S5 is unchanged. Consequently, its lower bound holds with probability at least $q_n(E)$.

Generalize to Random Estimator and Dataset

We first relax the assumption that y_ξ and $\mathcal{D}$ are deterministic by allowing both to be random. We abuse the notation a little to let ξ denote the randomness of y_ξ. The consistency result below is agnostic to whether y_ξ is parametric or non-parametric, and closely parallels Theorem S7.

Theorem S12 If Assumption S6 and the setting of Theorem S5 are uniformly satisfied as

$$\mathbb{E}_{\xi,\mathcal{D}}\left[W_1(y_{\xi\#}\mu, y_{\#}\mu)\right] \to 0,$$

where the expectation is taken with respect to the distribution of the estimator as well as the training dataset, then $y_\xi \xrightarrow{\mathbb{P}} y$ on $\mathcal{X}$.

The proof follows directly from the proof of Theorem S7 by expanding the expectation into a conditional expectation.

This extension highlights a subtle but essential point in the context of neural network (NN)-based estimators. In practice, a learned NN-based model y_ξ inherits randomness from two sources: (i) the sampling variability of the training set $\mathcal{D}$ and (ii)

the stochasticity of the optimization procedure (random initial weights, data shuffling, dropout, etc.). Consequently, even with a fixed dataset, repeated training runs can yield a collection of estimators with indistinguishable empirical risks yet divergent predictions. Consistency therefore requires each realization y_ξ in this family to be (i) data-consistent and (ii) trained so that the Wasserstein loss is driven sufficiently close to 0 on average. When these conditions hold, the expectation in Theorem S12 converges to zero and the whole distribution of estimators collapses around the true map y; equivalently, y_ξ converges to y in probability on $\mathcal{X}$.

Extension to Composite Observables $g \circ u$

While the theory thus far has been developed for the complete output map y, all results hold verbatim when y is replaced by the composite observable $g \circ u$. If g captures the relevant physical notion of extremeness, then the same fundamental difficulty in the data-scarce regime persists, and the 1-Wasserstein distance remains an optimal metric for measuring tail discrepancy.

A natural question is whether properties of the latent state map u can be inferred from those of $g \circ u$. In general, this is not possible: since u typically resides in a much higher-dimensional space, recovering it from observations in a lower-dimensional (particularly 1-dimensional) space is an ill-posed inverse problem unless additional structure is imposed. A plausible structural assumption is that g is bi-Lipschitz; that is, there exists $L > 0$ such that for all $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}$,

$$\frac{1}{L}\|\mathbf{u}_1 - \mathbf{u}_2\| \leq |g(\mathbf{u}_1) - g(\mathbf{u}_2)| \leq L\|\mathbf{u}_1 - \mathbf{u}_2\|.$$

Under this condition, it is not hard to see that optimality results for y obtained in Theorems S9 and S11 directly transfer to u, which would be ideal. However, bi-Lipschitz maps must be bijective, and such bijectivity in general fails for functions $g : \mathbb{R}^m \to \mathbb{R}$ when $m > 1$.

Generalize to Higher Moments

The W_1 distance only requires the existence of the first moment. In Theorems S9 and S11, we have established that W_1 is optimal for capturing tail deviations in the first moment. A natural question arises: if higher moments are available, can similar optimality results be derived for W_p? This section explores that question.

To begin, we state the definition of the p-Wasserstein distance.

Definition S13 (p-Wasserstein Distance [1]) Let $\nu_1, \nu_2 \in \mathcal{P}_p(\mathcal{Y})$ be two probability measures, where $\mathcal{P}_p(\mathcal{Y})$ denotes the space of probability measures on $\mathcal{Y}$ with a finite p-th moment. Let $\Pi(\nu_1, \nu_2)$ be the set of all probability couplings between ν_1 and ν_2; that is,

$$\Pi(\nu_1, \nu_2) := \{\gamma \in \mathcal{P}(\mathcal{Y} \times \mathcal{Y}) : \forall \text{ measurable } A \subset \mathcal{Y}, \gamma(A \times \mathcal{Y}) = \nu_1(A), \gamma(\mathcal{Y} \times A) = \nu_2(A)\}.$$

The p-Wasserstein distance between ν_1 and ν_2 is defined as

$$W_p^p(\nu_1, \nu_2) = \inf_{\gamma \in \Pi(\nu_1, \nu_2)} \int \|y_1 - y_2\|^p \, \mathrm{d}\gamma(y_1, y_2), \tag{11}$$

where $\|\cdot\|$ denotes the Euclidean norm.

We generalize a key lemma used to prove the lower bound of W_1 earlier.

Lemma S14 (Generalize Lemma S4 to W_p.) Let y_1 and y_2 be two functions in $L^p(\mu)$ where $p \geq 1$, mapping from $\mathcal{X}$ to $\mathcal{Y}$. Suppose that they both are $\mathcal{B}(\mathcal{X})$-measurable with respect to $\mathcal{B}(\mathcal{Y})$ and that the input of both of them follows the distribution μ. Then we have

$$W_p(y_{1\#}\mu, y_{2\#}\mu) \geq \left| \int_{\mathcal{X}} (y_1 - y_2) \mathrm{d}\mu \right|.$$

Proof Take any coupling $\gamma \in \Pi(y_{1\#}\mu, y_{2\#}\mu)$. Derive that

$$\begin{aligned}
|\mathbb{E}[y_1 - y_2]|^p &= \left| \int y_1 \mathrm{d}(y_{1\#}\mu) - \int y_2 \mathrm{d}(y_{2\#}\mu) \right|^p \\
&= \left| \int y_1 \mathrm{d}\gamma(y_1, y_2) - \int y_2 \mathrm{d}\gamma(y_1, y_2) \right|^p \\
&= \left| \int (y_1 - y_2) \, \mathrm{d}\gamma(y_1, y_2) \right|^p \\
&\leq \int |y_1 - y_2|^p \, \mathrm{d}\gamma(y_1, y_2),
\end{aligned}$$

where the last inequality follows from Jensen's inequality. Now, since the relationship above holds for any coupling, it particularly holds for the optimal coupling γ^*. Therefore,

$$|\mathbb{E}[y_1 - y_2]|^p \leq \int |y_1 - y_2|^p \, \mathrm{d}\gamma^*(y_1, y_2) = W_p^p(y_{1\#}\mu, y_{2\#}\mu),$$

which implies the desired result.

Note that the proof here is different from the proof of Lemma S4; the Kantorovich duality is not invoked. Although the Lemma is generalizable to W_p, the lower bound is still in terms of the first moment rather than the p-th moment. To obtain lower bounds in terms of the p-th moment, we approach to another result in the literature.

Theorem S15 (Proposition 1.5 in [1]) Let $y_{1\#}\mu$ and $y_{2\#}\mu$ be two sub-exponential random variables. Then, for any $p > 1$ for some $k \geq 1$, it holds

$$W_p\left(y_{1\#}\mu, y_{2\#}\mu\right) \geq \frac{\left| \int \left(|y_1(\mathbf{x})|^p - |y_2(\mathbf{x})|^p \right) d\mu \right|}{\left(2p^2/(p-1)\right)^p}$$

We make several remarks on this bound. First, note that this lower bound is not sharp; in particular, when $p > 1$, the constant in the denominator, $(2p^2/(p-1))^p$, grows rapidly. The reason is in the proof of the theorem, since the attention is constrained on sub-exponential laws, the tail effect of the two distributions is controlled using sub-exponential tail bound, thereby obtaining a constant that under-estimates extreme behavior. The numerator, on the other hand, is ideal in the sense that it is in terms of $||y_1|^p - |y_2|^p|$ rather than $||y_1|^p - |y_2|^p|^{\frac{1}{p}}$. However, a growth rate of p^p in the denominator is in general insurmountable, leading to a non-sharp bound for general distributions.

We could indeed improve the constant in the denominator using a simple triangle inequality for general distributions while sacrificing the $||y_1|^p - |y_2|^p|$ in the denominator. The bound we obtain, despite only requiring elementary techniques, is sharp for general distributions. We summarize this result here.

Lemma S16 In the setting of Lemma S14, we have

$$W_p\left(y_{1\#}\mu, y_{2\#}\mu\right) \geq \left| \left(\int_{\mathcal{X}} |y_1(\mathbf{x})| \mathrm{d}\mu \right)^{\frac{1}{p}} - \left(\int_{\mathcal{X}} |y_2(\mathbf{x})| \mathrm{d}\mu \right)^{\frac{1}{p}} \right|$$

Proof The proof follows from triangle inequality by considering the coupling between $y_{1\#}\mu$ and δ_0 and that between $y_{2\#}\mu$ and δ_0, where δ_0 is the Dirac delta at the origin.

Finally, we generalize Lemma S8 to W_p to obtain a p-th moment upper bound.

Lemma S17 (Generalize Lemma S8 to W_p.) In the setting of Lemma S14, we have

$$W_p^p\left(y_{1\#}\mu, y_{2\#}\mu\right) \leq \int_{\mathcal{X}} |y_1(\mathbf{x}) - y_2(\mathbf{x})|^p \, \mathrm{d}\mu$$

Proof The techniques used to prove Lemma S8 apply verbatim, and the result directly follows.

Note that Lemma S14 (and Lemma S8) is sharp by considering y_1 and y_2 as constant maps.

We now observe that upper bound in Lemma S17 does not match with either the lower bound obtained in Theorem S15 or Lemma S16, the best moment p-th moment bounds we could have obtained using the proof techniques we have adopted. Therefore, optimality results like the ones we obtained for W_1 in Theorems S9 and S11 in the main text are not possible.

6 Experimental Setups and Details

We provide additional details for the numerical results presented in the Applications section of the main text, including the problem setups and training and testing procedures.

6.1 The 2D-to-1D Toy Problem

Training and testing details.

The explicit formula of the true map y is given by

$$\begin{aligned} y(x_1, x_2) = &\frac{1.5}{2\pi\sqrt{0.5}} \exp\left(-\left((x_1-2)^2 + (x_2-2)^2\right)\right) \\ &+ \frac{1.5}{2\pi\sqrt{0.7}} \exp\left(-\frac{1}{1.4}\left((x_1+1)^2 + (x_2+1)^2\right)\right) \\ &+ \frac{1.0}{2\pi\sqrt{0.3}} \exp\left(-\frac{1}{0.6}\left((x_1-2)^2 + (x_2+2)^2\right)\right) \\ &+ \frac{0.5}{2\pi\sqrt{0.9}} \exp\left(-\frac{1}{1.8}\left((x_1-0)^2 + (x_2-1)^2\right)\right) \\ &+ \frac{1.25}{2\pi\sqrt{0.6}} \exp\left(-\frac{1}{1.2}\left((x_1-0.5)^2 + (x_2+0.5)^2\right)\right) \end{aligned}$$

To obtain ν_0, we sample 1,000,000 $\mathbf{x}_i$ from the input distribution, evaluate $y(\mathbf{x}_i)$ based on the analytic formula, and compute the PDF. We form $\mathcal{Q}$, the set of quantile levels, with the following probability values:

$$\mathcal{Q} = \begin{Bmatrix} \text{linsp}(0, 10^{-4}, 10),\ \text{linsp}(10^{-4}, 10^{-3}, 10), \\ \text{linsp}(10^{-3}, 10^{-2}, 10), \text{linsp}(10^{-2}, 10^{-1}, 9), \\ \text{linsp}(10^{-1}, 1-10^{-1}, 20), \\ \text{linsp}(1-10^{-1}, 1-10^{-2}, 21), \text{linsp}(1-10^{-2}, 1-10^{-3}, 21), \\ \text{linsp}(1-10^{-3}, 1-10^{-4}, 21), \text{linsp}(1-10^{-4}, 1-10^{-5}, 21), \\ \text{linsp}(1-10^{-5}, 1-10^{-6}, 21), \text{linsp}(1-10^{-6}, 1-10^{-7}, 21) \end{Bmatrix},$$

where the function $\text{linsp}(a, b, n)$ is defined as:

$$\text{linsp}(a, b, n) = \left\{a,\ a + \frac{b-a}{n-1},\ a + 2\frac{b-a}{n-1},\ \ldots,\ b\right\};$$

i.e. it generates n evenly spaced points between a and b inclusively. Here, we also include several quantiles before the 0.9 probability level to reduce the mismatch in the bulk as well as the left tail. The primary focus remains the right tail. There is no particular strategy for forming $\mathcal{Q}$. We use standard multi-layer perceptron (MLP) with 3 hidden layers of 256 neurons as the network architecture. The activation function is

chosen as a smoothed ReLU function: $\frac{x}{1+e^{-4x}}$. Algorithm 1 of the main text is executed with the chosen $\mathcal{D}$, $\mathcal{Q}$, and $\lambda = 1.0$. We use Adam as the optimizer with default hyperparameter values in PyTorch [6], and we train the network for 4000 gradient descent iterations. No ad hoc hyperparameter tuning or architecture selection is performed. Additionally, a separate neural network is trained under identical configurations to obtain an MSE estimator for baseline comparison.

Robustness of λ.

Regarding the hyperparameter λ, we found in our experiments that our method is robust to its value. To illustrate this, we include an additional experiment here in Supplementary Figure 4, which shows the push-forward densities of the η-estimators in the 2D-to-1D toy example under a wide range of λ values. The results demonstrate that the learned densities consistently align well with the ground truth, when λ varies across several orders of magnitude from 0.0001 to 10. Importantly, all other hyperparameters (e.g., learning rate, number of epochs) were held fixed in this experiment, further confirming robustness with respect to λ.

In the main-text subsection A 2D-to-1D Toy Problem: Description of the Results, we replace the final paragraph concerning another key feature of the η-map with the following:

Ensemble spatial variance.

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

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

We report the ensemble mean location

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

and the ensemble spatial variance

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

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

$$\mathbf{x}_\star = (2.020, -2.020),$$

whereas the ensemble mean inferred location is

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

The ensemble spatial variance is

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

Thus, independently initialized η-estimators can place the inferred high-output region at substantially different spatial locations, quantifying the non-uniqueness presented in Supplementary Figure 2.

6.2 The 2D-to-2D Toy Problem

In this example, we model the intermediate state map u as follows:

$$u(x_1, x_2) = \begin{bmatrix} u_1(x_1, x_2) \\ u_2(x_1, x_2) \end{bmatrix}$$

where

$$u_1(x_1, x_2) = y(x_1, x_2),$$

and $y(x_1, x_2)$ is the one defined in the 2D-to-1D toy problem. Here, $u_2(x_1, x_2)$ is a Fourier mode and has a formula of

$$u_2(x_1, x_2) = -0.1 \sin\left(\frac{\pi}{3} x_1\right) \sin\left(\frac{\pi}{4} x_2\right)$$

u_2 can be interpreted as some noisy information that complicates the state space. It is the extreme that u_1 exhibits we seek to discover. We consider the observable function g as

$$g(\mathbf{u}) = 2|u_1| + \frac{|u_2|}{2},$$

where the unbalanced weight is to reflect that the extreme observable aligns more with u_1 than u_2. To compute the PDF of $y(x_1, x_2) = g(u(x_1, x_2))$, we generate 1,000,000 samples from the input space and evaluate $y(x_1, x_2)$. This resulting PDF serves as ν_0, and the quantile set $\mathcal{Q}$ is constructed as

$$\mathcal{Q} = \begin{Bmatrix} \text{linsp}(0, 1 - 10^{-1}, 41), \\ \text{linsp}(1 - 10^{-1}, 1 - 10^{-2}, 21), \text{linsp}(1 - 10^{-2}, 1 - 10^{-3}, 21), \\ \text{linsp}(1 - 10^{-3}, 1 - 10^{-4}, 21), \text{linsp}(1 - 10^{-4}, 1 - 10^{-5}, 21), \\ \text{linsp}(1 - 10^{-5}, 1 - 10^{-6}, 21), \text{linsp}(1 - 10^{-6}, 1 - 10^{-7}, 21) \end{Bmatrix}.$$

6.3 The Real-World Precipitation Downscaling Challenge

Data processing.

We use the ERA5-Land dataset [7] and download in total 25-year hourly total precipitation (the variable `tp`) data from 1999 to 2023 on the main continent of the United States. We then select the particular mid-east region with a latitude ranging from 30.8°N to 38.7°N and a longitude ranging from 97.6°W to 81.7°W. To emphasize the presence of extreme events, all hourly snapshots are aggregated by taking the

maximum precipitation value on each spatial grid over 24-hour intervals. This process reduces the total number of snapshots by a factor of 24, resulting in 9044 daily maximal precipitation fields. The data resolution is 0.1° in both spatial dimensions, yielding a dataset of size $9044 \times 80 \times 160$. This high-resolution dataset is denoted as $\mathcal{D}_{\text{hi}}$. To generate the corresponding coarse-scale dataset, we subsample each snapshot by retaining one field value for every 10 grid points in each dimension. This subsampling results in a dataset of resolution $9044 \times 8 \times 16$, referred to as $\mathcal{D}_{\text{lo}}$. A crucial assumption here is the availability of well-aligned LR-HR pairs, which are not always available for turbulent dynamics due to their chaotic nature. In that case, techniques such as nudging ([8]) could be employed to find a trustworthy paired dataset. This is not a focus of our work.

Training and testing details.

We assume the ratio of the number of HR snapshots available for training to the number of LR snapshots available is 1:50. Specifically, we take only 0.5 years (from 1999 to 1999.5) of data out of $\mathcal{D}_{\text{hi}}$ and their corresponding downsampled LR snapshots out of $\mathcal{D}_{\text{lo}}$ as training data, representing a scenario of extreme data scarcity. In total, 180 pairs of snapshots are used in training as the ERM loss. The ν_0 is constructed using the PDF of $\{\max(\mathbf{u}_i)\}_{i=1}^{9044}$, and the full training algorithm follows Algorithm 1 of the main text. The tail-cutoff parameter τ in Algorithm 1 of the main text is chosen to be 0.989 which corresponds to a quantile value of 150, and we form $\mathcal{Q}$ with 102 probability values between τ and 1 for training. It is important to emphasize that ν_0 provides information only about values (i.e., quantiles) and not the specific times or spatial locations at which these quantiles occur. Consequently, a naively implemented training algorithm is unlikely to produce physically plausible dynamics and can induce prohibitively expensive memory cost, as discussed in the Practicality Section, due to the high dimensionality of the discretization grids and preferably accurate quantile estimations. The training techniques developed in Algorithm 1 of the main text address these challenges, enabling the generation of physically meaningful dynamics and statistically consistent observables while allowing for a low memory cost. We use a standard 2D convolutional neural network with upsampling layers in between convolutional layers as the model architecture. The balancing parameter λ is set to be 1. For optimization, we employ the Adam optimizer [9] with a learning rate of 0.0003. The window size ω in Algorithm 1 of the main text is selected to be 30, and we train the network for 150 epochs. Additionally, another neural network is trained under identical configurations to compute an MSE estimator for baseline comparison. After training, we obtain the u_η-map and the u_ξ-map, and thus the y_η-map and the y_ξ-map. We collect the downscaled snapshots $\{u_\eta(\mathbf{x}_i)\}_{i=1}^{9044}$ and $\{u_\xi(\mathbf{x}_i)\}_{i=1}^{9044}$ as well as the observable data $\{y_\eta(\mathbf{x}_i)\}_{i=1}^{9044}$ and $\{y_\xi(\mathbf{x}_i)\}_{i=1}^{9044}$, visualize the snapshots, compute the PDFs of the observable, and compare against the ground truth.

Further, we consider a conditional mean statistic to examine the magnitude of extremes by focusing on the tail as well as a weighted coverage statistic to assess the accuracy of the η-map in reconstructing the spatial fraction of field values exceeding a chosen threshold across the entire domain and time interval. For a sample field $\mathbf{u}_0$, we denote the conditional mean as $m(\mathbf{u}_0; t, n_g)$ and the weighted coverage as $c(\mathbf{u}_0; t, n_g)$,

where t is a threshold parameter and n_g is the number of discretizing grids of the input field $\hat{u}$. They are defined as

$$m(\mathbf{u}_0; t, n_g) := \frac{\sum_i^{n_g} u_0^{(i)} \cdot \mathbb{I}\{u_0^{(i)} \geq t\}}{\sum_i^{n_g} \mathbb{I}\{u_0^{(i)} \geq t\}}, \tag{12}$$

and

$$c(\mathbf{u}_0; t, n_g) := \frac{\sum_i^{n_g} u_0^{(i)} \cdot \mathbb{I}\{u_0^{(i)} \geq t\}}{\sum_i^{n_g} u_0^{(i)}}, \tag{13}$$

where $\mathbb{I}$ denotes the indicator function. Intuitively, c quantifies the total magnitude of $\mathbf{u}_0$ contributed by grid cells whose values exceed t, thus capturing the spatial extent of extreme events in a magnitude-weighted sense.

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 $\mathbf{u}_j$ and $\widehat{\mathbf{u}}_j$ 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

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

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 7×7 local windows. For a window $w \in \mathcal{W}$, define

$$\mathrm{SSIM}_w(\widehat{\mathbf{u}}_j, \mathbf{u}_j) = \frac{(2m_{\widehat{\mathbf{u}}_j,w} m_{\mathbf{u}_j,w} + C_1)(2\sigma_{\widehat{\mathbf{u}}_j \mathbf{u}_j,w} + C_2)}{(m_{\widehat{\mathbf{u}}_j,w}^2 + m_{\mathbf{u}_j,w}^2 + C_1)(\sigma_{\widehat{\mathbf{u}}_j,w}^2 + \sigma_{\mathbf{u}_j,w}^2 + C_2)},$$

where $m_{\widehat{\mathbf{u}}_j,w}$ and $m_{\mathbf{u}_j,w}$ are the local means, $\sigma_{\widehat{\mathbf{u}}_j,w}^2$ and $\sigma_{\mathbf{u}_j,w}^2$ are the local variances, and $\sigma_{\widehat{\mathbf{u}}_j \mathbf{u}_j,w}$ 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

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

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. We report the sample mean $\overline{\text{SSIM}}$ and sample standard deviation σ_{SSIM} for each selected subset.

For the η-map, we also report the relative RMSE increase compared to the MSE-map,

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

The bulk subsets are defined as

$$\mathcal{I}_{\leq q} := \{1 \leq j \leq 9044 : \max(\mathbf{u}_j) \leq F^{-1}_{\nu_0}(q)\},$$

and the tail subsets are defined as

$$\mathcal{I}_{\geq q} := \{1 \leq j \leq 9044 : \max(\mathbf{u}_j) \geq F^{-1}_{\nu_0}(q)\}.$$

Over the full evaluation set with 9044 samples, the η downscaler has an 8.13% larger full-grid RMSE than the MSE downscaler. The $\overline{\text{SSIM}}$ decrease is 0.30%, and σ_{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 $\mathcal{I}_{\leq 0.7}$, $\mathcal{I}_{\leq 0.8}$, and $\mathcal{I}_{\leq 0.9}$, the relative RMSE increases are 6.20%, 6.55%, and 6.92%, respectively. The corresponding $\overline{\text{SSIM}}$ decreases are 0.19%, 0.23%, and 0.27%. Hence, in the bulk and moderate regimes, the η-map sacrifices a small amount of pointwise fidelity while retaining high structural similarity.

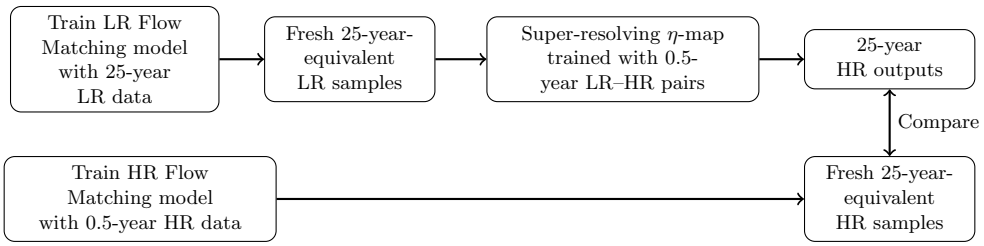
In the tail subsets, both downscalers have larger RMSE because the selected fields are more extreme. The η-downscaler has relative RMSE increases of 10.22%, 10.30%, and 9.21% for $\mathcal{I}_{\geq 0.95}$, $\mathcal{I}_{\geq 0.975}$, and $\mathcal{I}_{\geq 0.99}$ respectively. However, $\overline{\text{SSIM}}$ 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 Supplementary Figure 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.

6.4 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 η-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.



Training and testing details.

In this experiment, we again use only 0.5 years of HR data from $\mathcal{D}_{\text{hi}}$ to train the η-map, while retaining full access to $\mathcal{D}_{\text{lo}}$ during training. The training set-up for u_η stays the same as the vanilla downscaling case. A Flow Matching (FM) model [10], denoted p_{hi}, is trained on the 180 HR data points, and a separate FM model p_{lo} is trained on $\mathcal{D}_{\text{lo}}$. To ensure a fair comparison, both models adopt the widely validated linear conditional flow with the McCann interpolant [10], and the velocity fields are implemented using the well-established U-Net architecture [11]. After extensive hyperparameter tuning, the velocity field for p_{lo} is trained for 80 epochs with a batch size of 128, while that for p_{hi} is trained for 50 epochs using the same batch size. Both models are optimized using the Adam optimizer with a fixed learning rate of 0.0003, without a learning rate scheduler. During sampling, 9044 samples are generated from each model by integrating the respective flow with the Dormand–Prince 5(4) scheme—a 5th order Runge–Kutta method—for 200 timesteps. The samples from p_{lo} are then processed through the same u_η module used in the previous test case to obtain super-resolved versions. Finally, observable statistics are computed for both the HR samples from p_{hi} and the super-resolved samples from p_{lo}. Visualizations of the resulting η-generated HR fields and comparisons of the observable PDFs are provided in Supplementary Figure 10 and Fig. 5.

6.5 η-Downscaling with Hypothesized Heavier-Tailed Distributions

Determining the reference distribution.

To determine a suitable hypothesized distribution, we fit a heavy-tailed distribution to the observables of $\mathcal{D}_{\text{hi}}$ as defined above. Consequently, the fitted distribution exhibits a heavier tail than the true observables, reflecting the hypothesized extreme behavior.

For this purpose, the Generalized Extreme Value Distributions (GEVDs) are specifically chosen, as they are fundamental in modeling a wide range of extreme phenomena [12].

The PDF of the GEVD is defined as:

$$p(y;\kappa,\zeta,\sigma) = \frac{1}{\sigma}\left(1+\kappa\frac{y-\zeta}{\sigma}\right)^{-(1+\frac{1}{\kappa})}\exp\left(-\left(1+\kappa\frac{y-\zeta}{\sigma}\right)^{-1/\kappa}\right),$$

where κ is the shape parameter, ζ is the location parameter, and $\sigma > 0$ is the scale parameter. The GEVD encompasses three types of extreme value distributions based on the value of κ: for $\kappa > 0$, it corresponds to the Fréchet distribution that is heavy-tailed; for $\kappa = 0$, it reduces to the Gumbel distribution that has an exponentially decaying tail; and for $\kappa < 0$, it corresponds to the Weibull distribution with a bounded upper tail. The special case of the Gumbel distribution is expressed as:

$$p(y;\zeta,\sigma) = \frac{1}{\sigma}\exp\left(-\frac{y-\zeta}{\sigma} - \exp\left(-\frac{y-\zeta}{\sigma}\right)\right).$$

One particularly useful property of the GEVDs is the analytic formula for their quantiles. Specifically, suppose $Y \sim$GEVD. Then for a given value q is between 0 and 1, we have

$$\mathbb{P}\left(Y \le Q(q;\zeta,\sigma,\kappa)\right) = q,$$

where

$$Q(q;\zeta,\sigma,\kappa) = \begin{cases} \zeta - \sigma\ln(-\ln q), & \text{if } \kappa = 0, \\ \zeta + \frac{\sigma}{\kappa}\left((-\ln q)^{-\kappa} - 1\right), & \text{if } \kappa \neq 0. \end{cases}$$

We first use the Maximum Likelihood Estimation (MLE) method. This method identifies the parameters κ, ζ, and σ that maximize the likelihood of observing the dataset under the GEVD model. The MLE process iteratively optimizes the parameters by evaluating the negative log-likelihood function of the GEVD with respect to the data. More theoretical and algorithmic details on the MLE can be found in [12]. After the optimally fitted GEVD is determined by MLE, we manually tweak the scale parameter σ to make the tail heavier so that the hypothesized distribution extends beyond the range of the original dataset. Since our (and any real-world) dataset is of bounded support, we introduce a cutoff threshold at the tail of the fitted GEVD. All values beyond this cutoff are ignored. Suppose the tail is truncated at point $y_{1-\gamma}$ with a tail probability of γ. To ensure the truncated density remains a valid PDF, it needs to be divided by $1-\gamma$. We denote Y_γ as the random variable that follows this truncated distribution. After the normalization, it is straightforward to see that the quantiles still exhibit closed-form expressions: i.e.,

$$\mathbb{P}\left(Y_\gamma \le Q_\gamma(q;\zeta,\sigma,\kappa)\right) = q,$$

where

$$Q_\gamma(q;\zeta,\sigma,\kappa) = Q\left(q(1-\gamma);\zeta,\sigma,\kappa\right).$$

In our case, the MLE-determined parameters are $\kappa = -0.179$, $\zeta = 25.077$, and $\sigma = 21.928$, with σ being manually added by 4 to 25.928 for reasons above. We chose $\gamma = 0.00470$ and the corresponding truncation point is 258.0.

Training and testing details.

The configuration of the training data, neural network architectures, and optimization procedure stays the same as the vanilla downscaling case. The only difference in this case is the selection of the reference PDF and, as a result, the tail cut-off parameter τ used to approximate the regularization term. Specifically, τ is chosen as 0.95, which corresponds to a quantile value of approximately 122 in the original ground truth observable dataset. $\mathcal{Q}$ is then formed with 350 probability values lying between τ and 1. The testing and evaluation procedures stay the same as the vanilla case.

Sensitivity to misspecified tails.

We examine how errors in the supplied tail information propagate to the downscaled precipitation fields. The W_1 estimator uses ν_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 ν_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 α 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

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

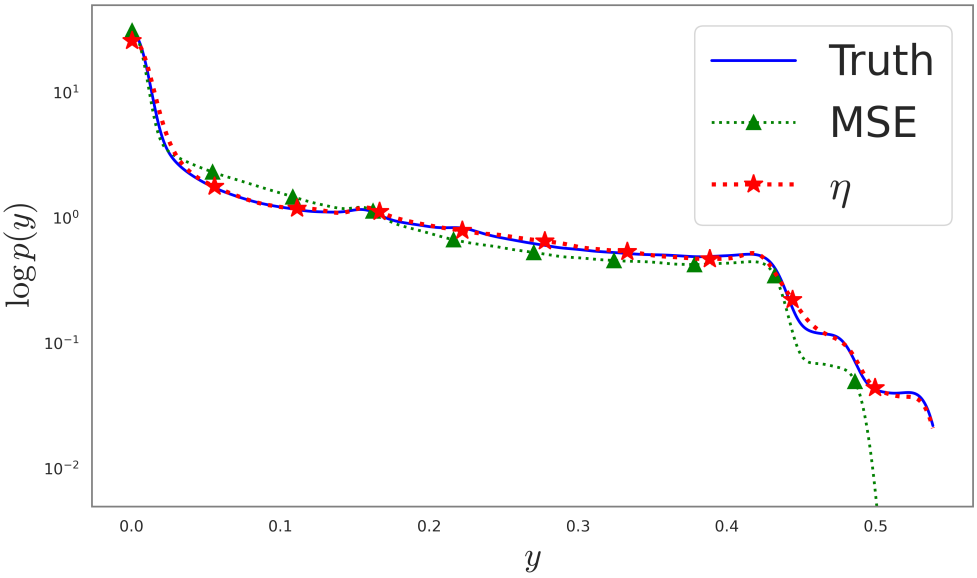
we use the same training configuration except that ν_0 varies across the perturbed family. We evaluate the complete predicted HR fields using RMSE together with $\overline{\text{SSIM}}$ and σ_{SSIM}. All metrics are computed over the $N = 9{,}044$ HR fields.

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 $\alpha = -0.5$ 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%.

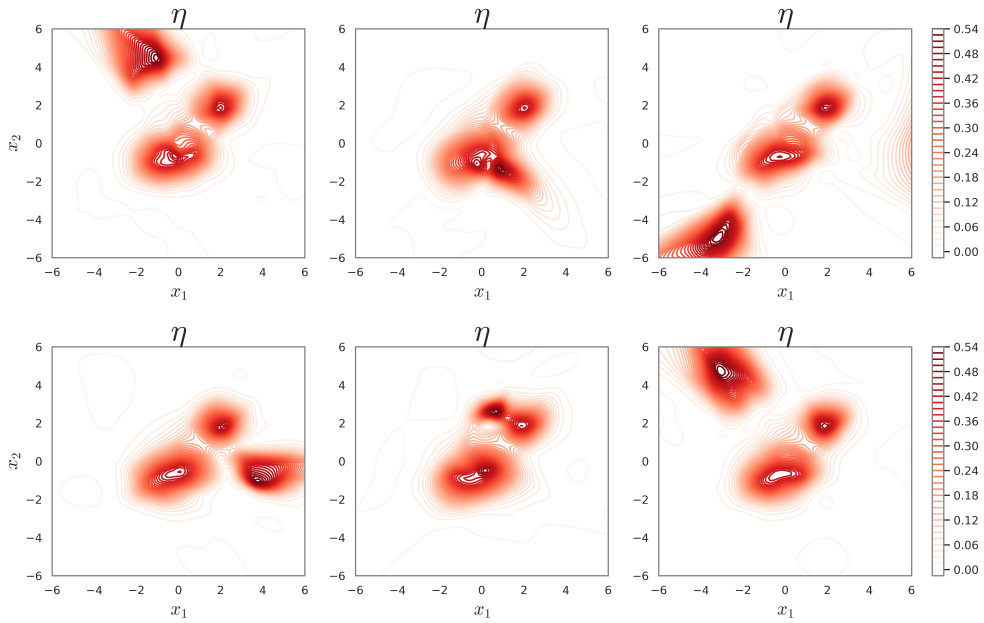
The SSIM behavior is more stable. $\overline{\text{SSIM}}$ remains between 0.940689 and 0.947280 for all misspecified-tail models. It is not strictly monotone in α: the $\alpha = 2$ model has the largest RMSE but a slightly higher $\overline{\text{SSIM}}$ than the $\alpha = 1.5$ model. σ_{SSIM} increases 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 η-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.

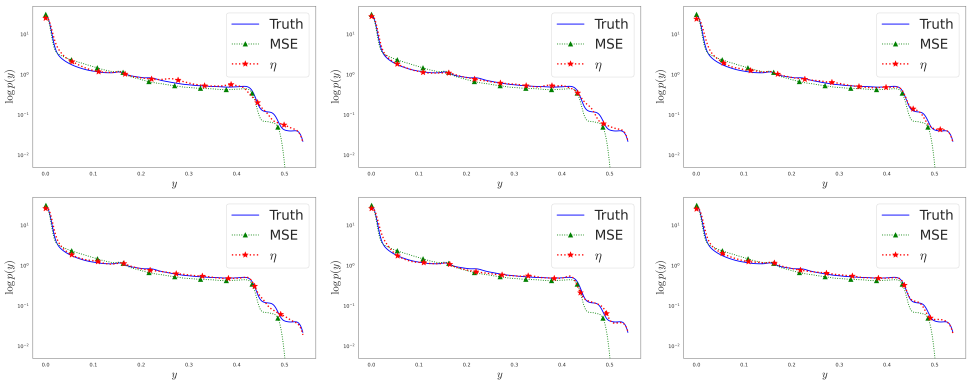
Supplementary Figures



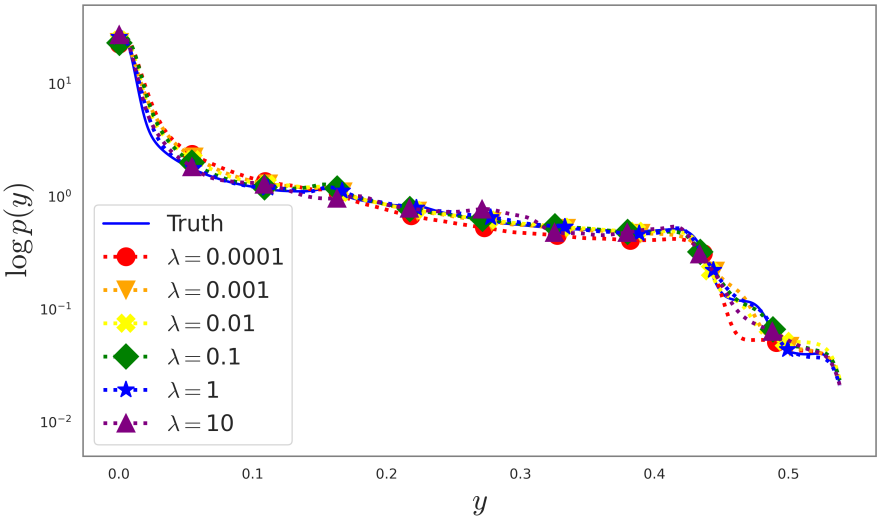
Supplementary Figure 1: The output PDF comparison of different estimators against the ground truth for the 2D-to-1D toy example. The solid blue curve: the ground truth. The dotted green curve with triangle markers: the MSE estimator. The dotted red curve with star markers: the η-estimator.



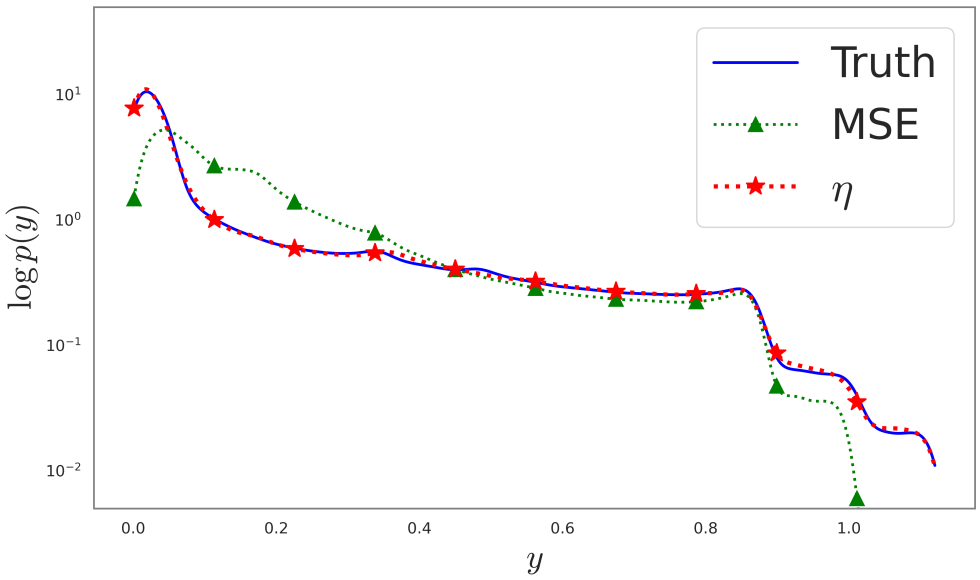
Supplementary Figure 2: The contour plot of 6 more independent realizations of the η-estimator trained with the same dataset and reference distribution as the realization shown in the rightmost plot of Fig. 1 in the main text.



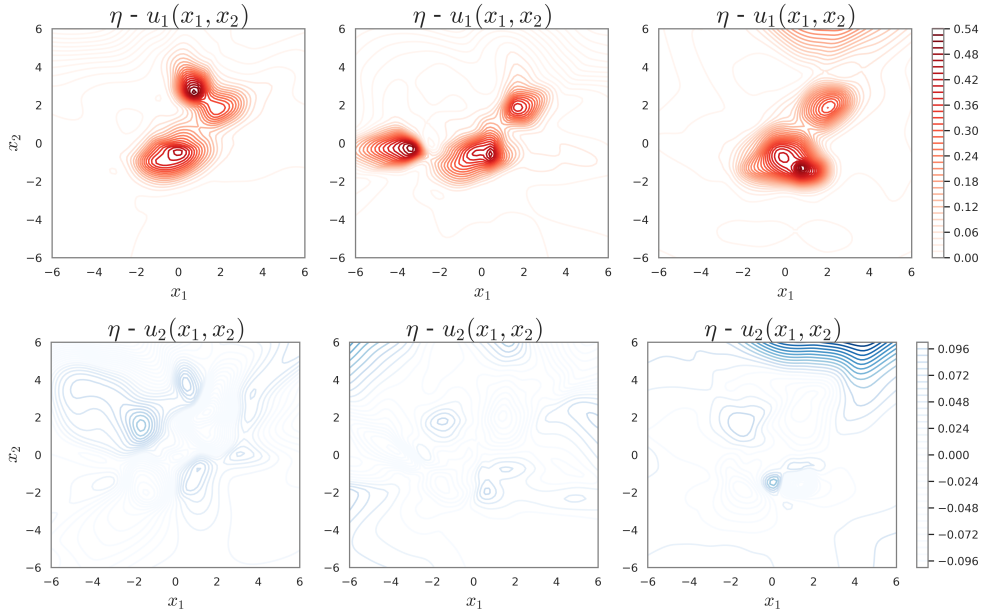
Supplementary Figure 3: The output PDF comparisons for the extra realizations of the η-estimators shown in Supplementary Figure 2. Each subplot in the present figure corresponds to the realization shown in the same subplot position in Supplementary Figure 2; for instance, the subplot in the second row and third column of the present figure is the PDF comparison plot for the η-realization in the second row and third column of Supplementary Figure 2. The solid blue curve: the ground truth. The dotted green curve with triangle markers: the MSE estimator. The dotted red curve with star markers: the η-estimator.



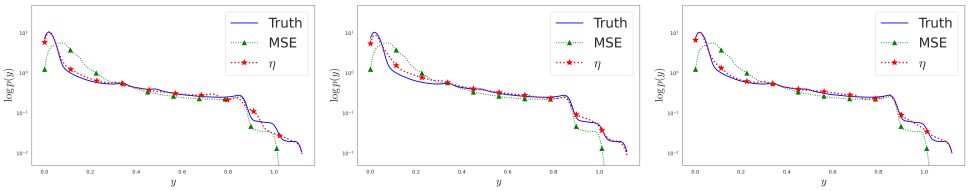
Supplementary Figure 4: The push-forward density of η-estimators under various values of λ during training in the 2D-to-1D toy example.



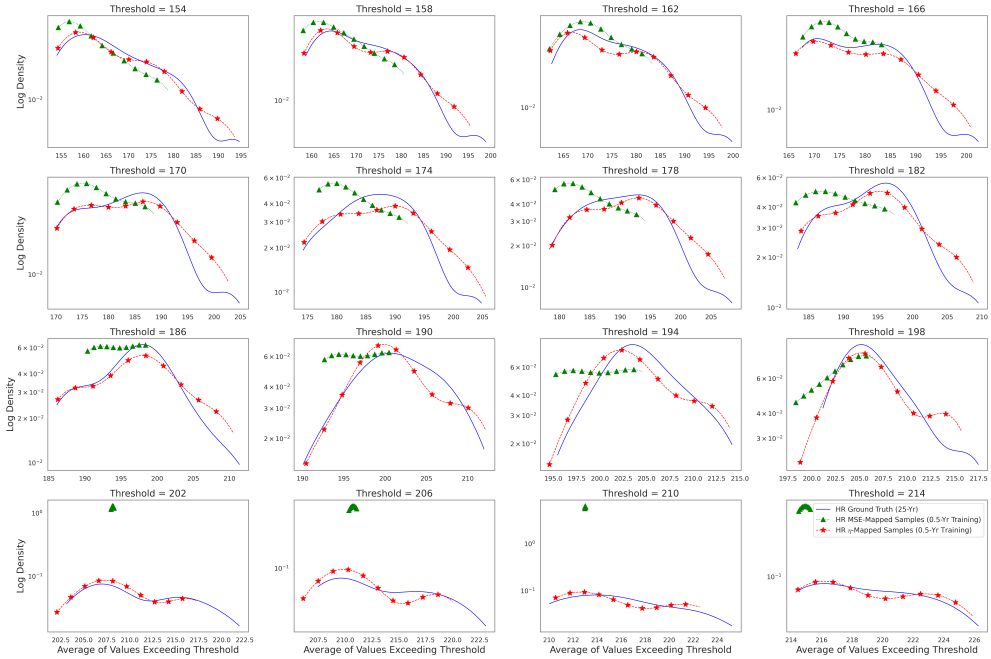
Supplementary Figure 5: The observable PDF comparison of different estimators against the ground truth for the 2D-to-2D toy example. The solid blue curve: the ground truth. The dotted green curve with triangle markers: the MSE estimator. The dotted red curve with star markers: the η-estimator.



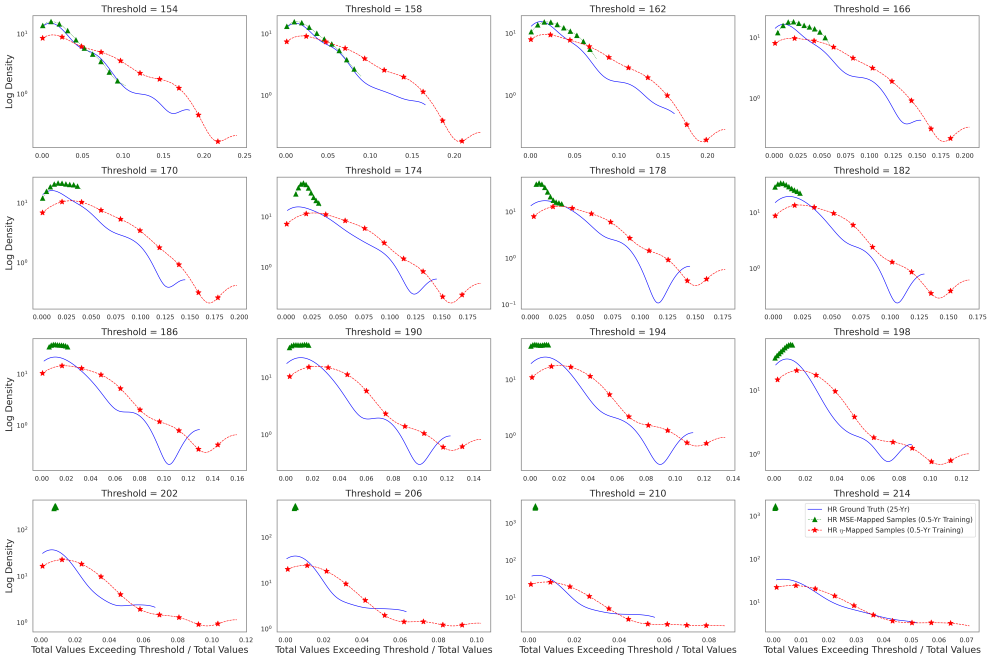
Supplementary Figure 6: The contour plot of 3 more independent realizations of the η-estimator trained with the same dataset and reference distribution as the realization shown in the rightmost plot of Fig. 2 in the main text. Each column corresponds to a different realization. Top row: the realization in the first coordinate. Bottom row: the realization in the second coordinate.



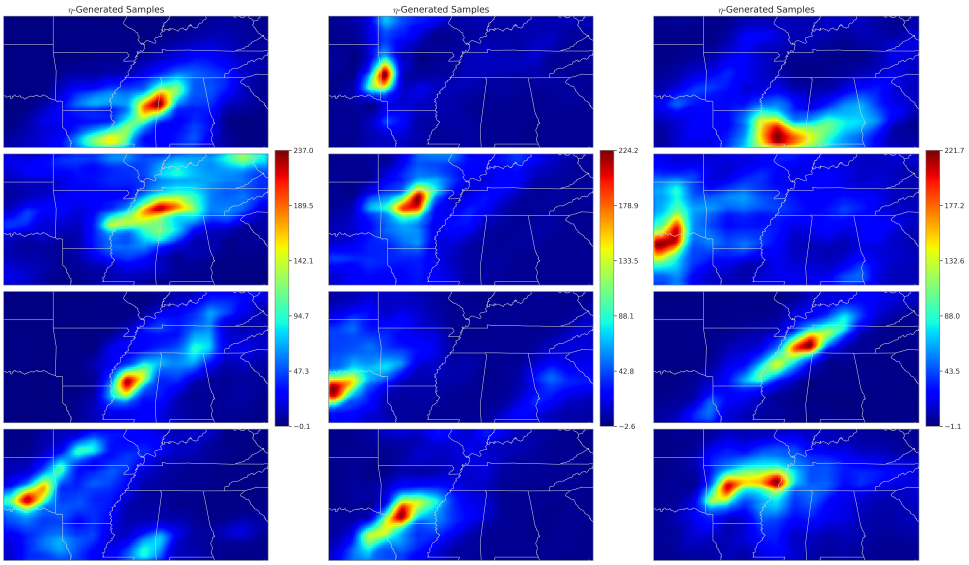
Supplementary Figure 7: The observable PDF comparisons for the 3 more realizations of the η-estimators shown in Supplementary Figure 6. Each subplot in the present figure corresponds to the realization shown in the same subplot position in Supplementary Figure 6; for instance, the subplot in the second column of the present figure is the observable PDF comparison plot for the η-realization in the second column of Supplementary Figure 6. The solid blue curve: the ground truth. The dotted green curve with triangle markers: the MSE estimator. The dotted red curve with star markers: the η-estimator.



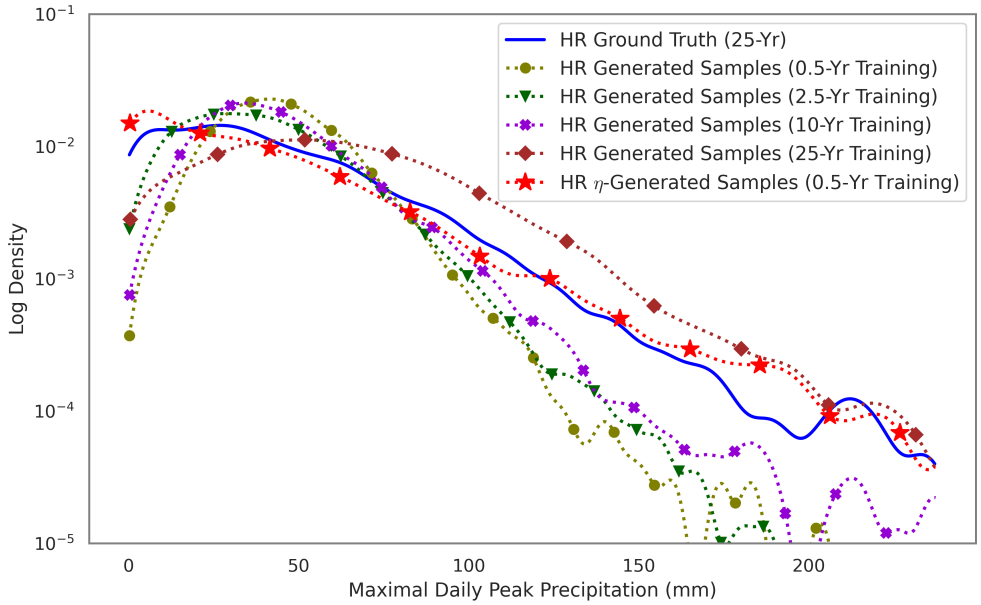
Supplementary Figure 8: The conditional mean PDF comparisons for a range of threshold values. The solid blue curve: the ground truth HR data. The dotted green curve with triangle markers: the HR samples super-resoluted from the 25-year ground truth LR samples through the MSE map. The dotted red curve with star markers: the HR samples super-resoluted from the 25-year ground truth LR samples through the η-map.



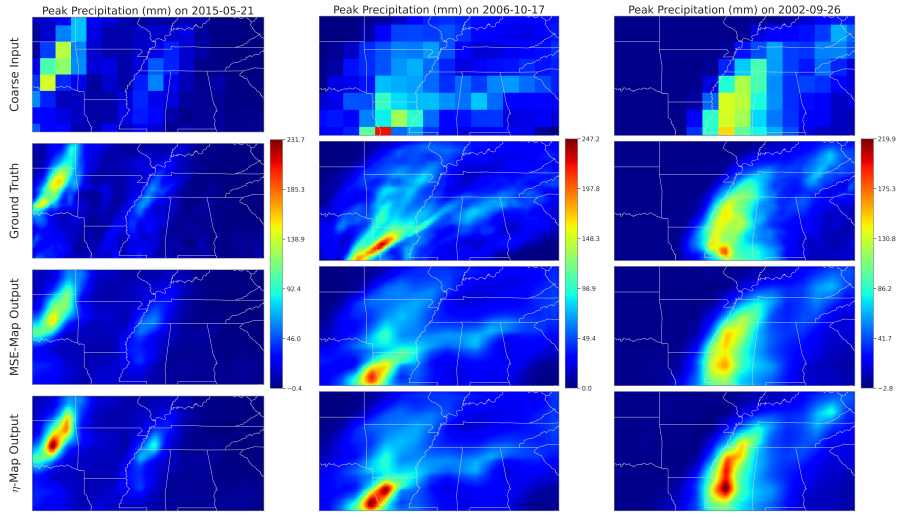
Supplementary Figure 9: The weighted coverage PDF comparisons for a range of threshold values. The solid blue curve: the ground truth HR data. The dotted green curve with triangle markers: the HR samples super-resoluted from the 25-year ground truth LR samples through the MSE map. The dotted red curve with star markers: the HR samples super-resoluted from the 25-year ground truth LR samples through the η-map.



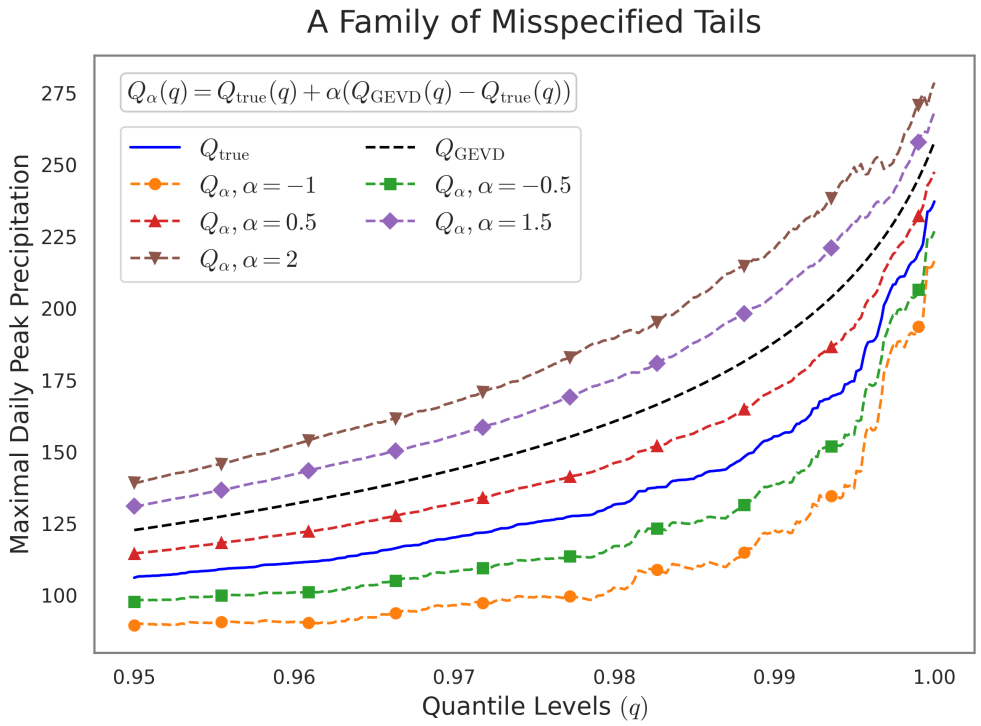
Supplementary Figure 10: Visualization of sample HR precipitation fields downscaled from the samples of the low-resolution (LR) generative model trained with 25-year LR data. There is no particular relationship among snapshots unlike other figures. All the samples are first drawn from the LR generative model and then downscaled by the η-map.



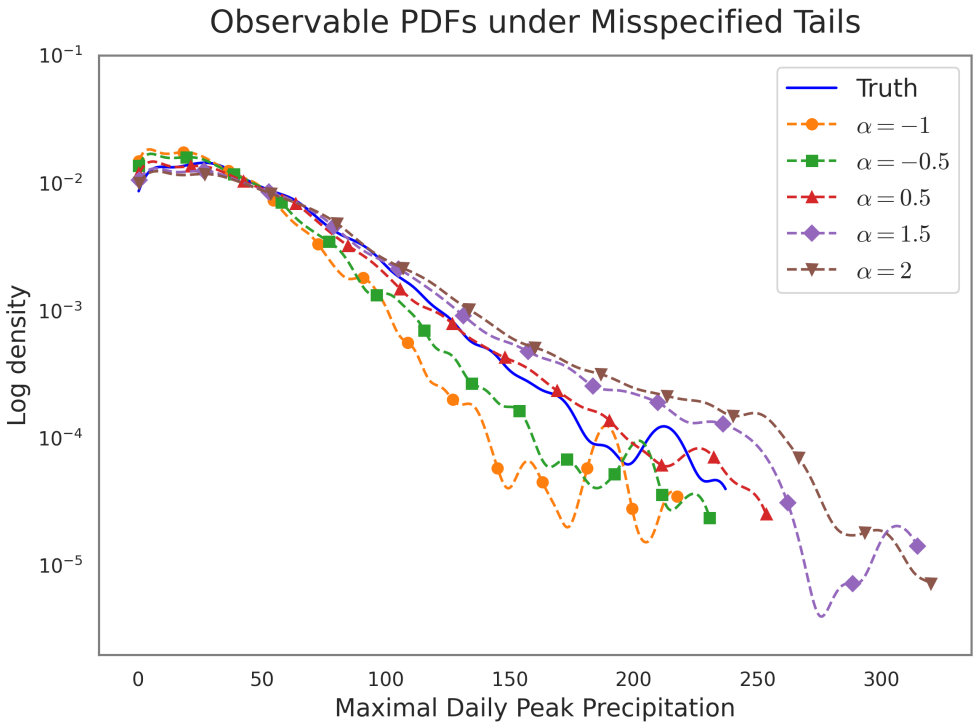
Supplementary Figure 11: The observable PDF comparison of the η-generated HR maximal daily peak precipitation fields against the observable PDFs of HR fields generated by Flow Matching (FM) trained with different amount of data. The solid blue curve: the ground truth HR data. The dotted olive curve with circle markers: the samples drawn from the HR FM model trained with 0.5-year data. The dotted dark green curve with flipped triangle markers: the samples drawn from the HR FM model trained with 2.5-year data. The dotted violet curve with cross markers: the samples drawn from the HR FM model trained with 10-year data. The dotted brown curve with rotated square markers: the samples drawn from the HR FM model trained with 25-year data. The dotted red curve with star markers: the HR samples generated from the η-map by mapping the samples drawn from the LR FM model through the η-map.



Supplementary Figure 12: Visualization of sample precipitation snapshots. Each column corresponds to a particular day in the test set. Each row corresponds to a particular processing method. First row: downsampled precipitation fields. Second row: ground truth HR fields. Third row: HR fields downscaled from the first row by the MSE map. Last row: HR fields downscaled from the first row by the η-map where the η-map is trained with a hypothesized GEVD that exhibits a much heavier tail than the truth.



Supplementary Figure 13: Sensitivity analysis for misspecified reference tail. Tail-target profiles obtained by shifting the true HR quantiles using the parameter α.



Supplementary Figure 14: Sensitivity analysis for misspecified reference tail. Resulting observable distributions from η-downscalers trained with the corresponding shifted ν_0 shown in Supplementary Figure 13.

Supplementary Tables

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

Supplementary 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. N denotes the size of the corresponding sample index set. RMSE is the full-field root-mean-square error, while $\overline{\text{SSIM}}$ and σ_{SSIM} are the sample mean and standard deviation, respectively, of the full-field SSIM scores. Here $r_{\eta/\text{MSE}} = (\text{RMSE}_\eta - \text{RMSE}_{\text{MSE}})/\text{RMSE}_{\text{MSE}}$ denotes the relative RMSE increase of the η-downscaler over the MSE-downscaler.

α	RMSE	$\overline{\mathrm{SSIM}}$	σ_{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

Supplementary Table 2: Sensitivity of the full-field fidelity of the η-mapped fields to misspecification of the supplied reference tail. All metrics are computed over the full $N = 9044$ paired HR fields. α controls the deviation from the ground-truth HR observable law. RMSE is the full-field root-mean-square error, while $\overline{\mathrm{SSIM}}$ and σ_{SSIM} are the sample mean and standard deviation, respectively, of the full-field SSIM scores.

Supplementary References

[1] Chewi, S., Niles-Weed, J., Rigollet, P.: Statistical optimal transport. arXiv preprint arXiv:2407.18163 (2024)

[2] Folland, G.B.: Real Analysis: Modern Techniques and Their Applications, 2nd edn. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. John Wiley & Sons, New York (1999)

[3] Sagiv, A.: The Wasserstein distances between pushed-forward measures with applications to uncertainty quantification. Communications in Mathematical Sciences **18**(3), 707–724 (2020) https://doi.org/10.4310/CMS.2020.v18.n3.a6 . Publisher: International Press of Boston. Accessed 2025-03-18

[4] Levin, D.A., Peres, Y.: Markov Chains and Mixing Times, 2nd edn., p. 447. American Mathematical Society, Providence, RI (2017). https://doi.org/10.1090/mbk/107 . With contributions by Elizabeth L. Wilmer

[5] Freidlin, M.I., Wentzell, A.D.: Random Perturbations of Dynamical Systems, 3rd edn. Grundlehren der mathematischen Wissenschaften, vol. 260. Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-25847-3

[6] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[7] Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., *et al.*: Era5-land: A state-of-the-art global reanalysis dataset for land applications. Earth system science data **13**(9), 4349–4383 (2021)

[8] Barthel Sorensen, B., Charalampopoulos, A., Zhang, S., Harrop, B., Leung, L., Sapsis, T.P.: A non-intrusive machine learning framework for debiasing long-time coarse resolution climate simulations and quantifying rare events statistics. Journal of Advances in Modeling Earth Systems **16**(3), 2023–004122 (2024)

[9] Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[10] Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)

[11] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241 (2015). Springer

[12] Coles, S.: An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics. Springer, London (2001). https://doi.org/10.1007/978-1-4471-3675-0