MIND THE RESIDUAL GAP: PROBABILISTIC DOWNSCALING UNDER REAL-WORLD BIAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Probabilistic downscaling models the conditional distribution of high-resolution fields given coarse inputs and is a core challenge in atmospheric science, climate modeling, and other multiscale physical systems. A widely used paradigm decomposes the problem into a deterministic mean predictor followed by a stochastic residual generator. While effective in idealized settings, this mean–residual approach frequently produces biased and underdispersive ensembles in real-world applications. We identify a fundamental source of this failure: *residual target misspecification*, where the residual distribution induced during training systematically differs from the correction distribution required at test time, a mismatch amplified by downscaling bias. We theoretically connect this misspecification to underdispersion through the spread–skill ratio and show that conditional residual distribution matching controls both residual-energy and conditional-mean mismatch. Motivated by this analysis, we introduce **ReMatch** (Residual Distribution Matching), which aligns training residual targets toward a held-out calibration regime via optimal transport. On a controlled synthetic benchmark with varying bias levels and a real-world HRRR–ERA5 wind-field downscaling task, ReMatch substantially reduces underdispersion, improves calibration, and outperforms strong mean–residual and super-resolution baselines.

1 INTRODUCTION

Probabilistic downscaling seeks to model the conditional distribution of high-resolution fields given coarse-resolution inputs, enabling downstream tasks that require fine-scale dynamics such as hazard assessment, local forecasting, trajectory planning, and uncertainty-aware decision making. While downscaling inherently has unresolved subgrid variability, real-world applications face challenges like systematic biases and mismatches between coarse- and fine-resolution data sources—collectively referred to as *downscaling bias*— which introduce additional, often domain- or source-dependent discrepancies (Maraun & Widmann, 2018). These biases distort the low-to-high resolution relationship far beyond simple interpolation, rendering the required high-resolution correction both stochastic and systematically more difficult to infer.

A two-stage mean–residual decomposition provides a useful framework for modeling both effects: a deterministic model predicts the conditional mean, after which a stochastic generative model learns the remaining residual distribution. Variants of this strategy have been adopted in various real-world tasks such as atmospheric downscaling (Mardani et al., 2025), weather and geophysical forecasting (Couairon et al., 2026; Finn et al., 2024), precipitation nowcasting (Yu et al., 2024; Nguyen et al., 2025), medical image super-resolution (Han & Huang, 2024), probabilistic time-series and energy forecasting (Lai et al., 2025; Chen & Gao, 2025), and speech synthesis (Chen et al., 2022). Despite their benefits, mean–residual generative models exhibit underdispersive predictions across applications and models, including atmospheric downscaling (Mardani et al., 2025), sea-ice forecasting (Finn et al., 2024), and probabilistic weather forecasting (Couairon et al., 2026).

We argue that this underdispersion stems from a fundamental issue that we term *residual target misspecification*. In the mean–residual framework, the residual is an induced correction target defined relative to the fitted mean predictor. Consequently, the residual distribution observed during training

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

can systematically differ from the corrections required at test time, causing the residual generator to underestimate the required predictive variability.

To address this, we propose **ReMatch** (Residual distribution Matching). Using a calibration set as a proxy for the test regime, ReMatch explicitly aligns the training residual distribution with the test-time residual distribution via optimal transport (Wan et al., 2023). This preserves the statistical benefits of the mean–residual decomposition while providing the residual generator with targets that better reflect the corrections required at inference.

Our main contributions are threefold: (i) we formalize residual target misspecification and establish its connection to ensemble underdispersion through the spread–skill ratio (SSR); (ii) we propose ReMatch, a practical residual distribution matching method that improves both predictive accuracy and ensemble calibration; and (iii) we provide a comprehensive empirical evaluation on a synthetic benchmark (Chung et al., 2023) with systematically varied LR–HR bias and a real-world HRRR–ERA5 wind downscaling task (Dowell et al., 2022; Hersbach et al., 2020). Across these settings, ReMatch consistently outperforms strong deterministic and probabilistic baselines, supporting residual target misspecification as a major contributor to underdispersion in biased downscaling.

2 Preliminaries

2.1 Probabilistic Downscaling

We consider probabilistic downscaling of high-resolution geophysical fields $y \in \mathbb{R}^{c_{\text{out}} \times H \times W}$ from low-resolution inputs $x \in \mathbb{R}^{c_{\text{in}} \times h \times w}$, where $H > h$ and $W > w$. Here, c_{in} and c_{out} denote the numbers of input and output channels, respectively, such as velocity, temperature or other physical variables, while (h, w) and (H, W) denote the low- and high-resolution spatial grid sizes. The task is to model the conditional distribution of y given x. We use X and Y to denote the corresponding random input and target fields, with x and y denoting their realizations.

In many physical downscaling problems, this mapping is fundamentally non-deterministic. A useful abstraction is to assume that the low-resolution observation is obtained from the high-resolution state through an effective degradation process $x = D(y, s)$, where D may include not only resolution reduction, but also regridding, projection, source mismatch, simulator bias, and other systematic discrepancies, and s denotes latent source- or pipeline-dependent factors. For example, an ERA5-to-HRRR wind downscaling task combines differences in spatial resolution, numerical model formulation, data assimilation, pressure-level definition, and grid projection within the same effective degradation process D. In idealized super-resolution benchmarks, D is often close to a simple downsampling operator, whereas in real-world settings it may reflect substantial bias and cross-source mismatch. In either case, D is many-to-one, so the inverse problem is ill-posed. Hence the natural learning target is not a unique deterministic inverse, but the conditional distribution. This probabilistic viewpoint is important not only because the inverse problem is many-to-one, but also because downstream applications often require uncertainty-aware predictions. In multiscale physical systems, unresolved fine-scale variability can affect forecasting, decision making, and motion planning of agents moving through the predicted field. Successful downscaling should therefore model both a central prediction and the spread of physically plausible high-resolution realizations.

2.2 Mean–Residual Decomposition

We study a class of pipelines in which an upstream mean predictor $\mu_\phi(x) \in \mathbb{R}^{c_{\text{out}} \times H \times W}$ first produces a deterministic approximation of the target, and a residual generator then models the remaining. For a realization (x, y), define the predictor-induced residual as $r_\phi(x, y) := y - \mu_\phi(x)$. A residual generator $q_\theta(r \mid x, \mu_\phi(x))$ then produces stochastic residual samples $\hat{r} \sim q_\theta(\cdot \mid x, \mu_\phi(x))$, leading to k number of ensemble predictions $\hat{y}^{(i)} = \mu_\phi(x) + \hat{r}^{(i)}, \quad i = 1, \ldots, k.$

This separation allows the deterministic predictor to capture the primary predictable structure while the residual generator focuses its capacity on unresolved fine-scale variability and uncertainty. Moreover, the residual distribution can be easier to model than the full target distribution, as the upstream mean predictor removes much of the dominant, large-scale variation in the target, resulting in lower variance in the residuals (Mardani et al., 2025). However, this advantage comes with an important caveat: the residual is defined relative to the upstream predictor. Once μ_ϕ is fixed, the residual stage

is therefore trained on a predictor-induced target rather than an intrinsic component of the data distribution. This distinction is central to our analysis: the residual target can change when the reliability of μ_ϕ changes across domains, even if the underlying downscaling task remains the same.

2.3 In-Sample and Out-of-Sample Residuals

Let $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ denote the fitting sample used to obtain μ_ϕ. Conditional on the realized fitting sample $\mathcal{S}$, let $(X_{\texttt{in}}, Y_{\texttt{in}})$ denote an observation drawn uniformly from $\mathcal{S}$, and let $(X_{\texttt{out}}, Y_{\texttt{out}}) \sim P$ denote an independent observation from the same population. Unless otherwise stated, we suppress conditioning on $\mathcal{S}$ for notational simplicity. For $d \in \{\texttt{in}, \texttt{out}\}$, define the corresponding residual random variable as

$$R_d := Y_d - \mu_\phi(X_d). \tag{1}$$

Even when the underlying observations are sampled from the same population, the fitted predictor is statistically coupled to the observations used in fitting but not to fresh observations. Consequently, $R_{\texttt{in}}$ and $R_{\texttt{out}}$ need not have the same distribution. We denote their conditional mean and variance by $m_d(x) := \mathbb{E}[R_d \mid X_d = x]$ and $V_d(x) := \operatorname{tr}\operatorname{Cov}(R_d \mid X_d = x)$ and define the residual energy as

$$M_d := \mathbb{E}\|R_d\|_2^2 = \mathbb{E}\left[\|m_d(X_d)\|_2^2 + V_d(X_d)\right]. \tag{2}$$

2.4 Spread-Skill Ratio

The spread–skill ratio (SSR) measures whether the variability represented by a predictive ensemble is equal to its prediction error. For a probabilistic model that generates predictions $\hat{Y} \sim p(\hat{Y} \mid X)$, we consider the squared SSR,

$$\mathrm{SSR}^2 = \frac{\mathbb{E}\left[\|\hat{Y} - \mathbb{E}[\hat{Y} \mid X]\|_2^2\right]}{\mathbb{E}\left[\|Y - \mathbb{E}[\hat{Y} \mid X]\|_2^2\right]}. \tag{3}$$

The numerator measures the ensemble variance, while the denominator measures the squared error of the ensemble mean. A calibrated predictive distribution has $\mathrm{SSR} = 1$; values below 1 indicate underdispersion, whereas values above 1 indicate overdispersion.

3 Residual Target Misspecification

Mean–residual generative models have repeatedly exhibited underdispersive behavior. Mardani et al. (2025) reports ensemble spreads that are systematically too small relative to ensemble-mean errors in atmospheric downscaling, while similar behavior has been observed in residual diffusion for sea-ice forecasting (Finn et al., 2024). More recently, Couairon et al. (2026) links underdispersion in probabilistic weather forecasting to overfitting of the deterministic predictor. In this section, we formalize this phenomenon as *residual target misspecification* and show how the mean–residual framework itself can induce underdispersion.

Fitting-induced residual-energy optimism. In the mean–residual structure introduced in Section 2.2, $\hat{Y} = \mu_\phi(X) + \hat{R}, \quad \hat{R} \sim q_\theta(\cdot \mid X, \mu_\phi(X))$, the corresponding residual target is induced by the fitted mean predictor as in Eq. 1.

Even without distribution shift, empirical risk minimization (ERM) favors lower residual energy on the samples used for fitting. Suppose μ_ϕ is obtained by minimizing empirical squared error over i.i.d. samples from P. Averaging over the randomness of the fitting sample, ERM yields the expected residual-energy inequality

$$\mathbb{E}_{\mathcal{S}}[M_{\texttt{in}}] \le \mathbb{E}_{\mathcal{S}}[M_{\texttt{out}}], \tag{4}$$

where the $\texttt{out}$ residual is evaluated on an independent observation from the same distribution. Intuitively, the fitted predictor is explicitly selected to minimize error on the observed sample, whereas an independent sample receives no such optimization advantage. Appendix A.2 formalizes this argument using an independent ghost sample and shows that Equation 4 follows directly from ERM optimality. This fitting-induced optimism is a standard consequence of ERM Vapnik (1999); Efron (1986).

How residual optimism produces under-dispersion. As in equation 2, we define mean and variance of predicted residual as $\hat{m}(x) := \mathbb{E}_{q_\theta}[\hat{R} \mid X = x]$ and $\hat{V}(x) := \operatorname{tr}\operatorname{Cov}_{q_\theta}(\hat{R} \mid X = x)$. Denote the corresponding predicted residual energy as $\hat{M} := \mathbb{E}\left[\|\hat{m}(X)\|_2^2 + \hat{V}(X)\right]$. Substituting the mean–residual structure into Equation 3 gives

$$\mathrm{SSR}^2 = \frac{\mathbb{E}[\hat{V}(X_{\mathtt{out}})]}{\mathbb{E}\left[V_{\mathtt{out}}(X_{\mathtt{out}}) + \|m_{\mathtt{out}}(X_{\mathtt{out}}) - \hat{m}(X_{\mathtt{out}})\|_2^2\right]}. \tag{5}$$

Define predictive residual-energy gap as $\Delta\hat{M} := M_{\mathtt{out}} - \hat{M}$. Then, as shown in Appendix A.2.3,

$$\mathrm{SSR} \le 1 \quad \Longleftrightarrow \quad \mathbb{E}\left[\hat{m}(X_{\mathtt{out}})^\top (m_{\mathtt{out}}(X_{\mathtt{out}}) - \hat{m}(X_{\mathtt{out}}))\right] \le \frac{\Delta\hat{M}}{2}. \tag{6}$$

Thus, underdispersion depends jointly on the residual energy supplied by the generator and the mismatch between its predicted conditional residual mean and the out-of-sample residual mean.

To isolate residual target misspecification from residual-generator approximation error, consider the optimistic setting in which the residual generator exactly matches its in-sample training target,

$$\hat{m}(x) = m_{\mathtt{in}}(x), \qquad \hat{V}(x) = V_{\mathtt{in}}(x). \tag{7}$$

Here, $m_{\mathtt{in}}$ and $V_{\mathtt{in}}$ characterize the residual target on the fitting distribution, whereas $\hat{m}$ and $\hat{V}$ are moments of the learned residual generator, which is subsequently applied to unseen inputs $X_{\mathtt{out}}$.

For a fixed fitting sample with $\Delta\hat{M} \ge 0$, Eq. 6 includes two special cases. If $m_{\mathtt{in}}(x) = 0$ for all x, meaning that the fitted mean predictor $\mu_\phi(x)$ recovers the in-sample conditional mean, then Eq. 7 implies $\hat{m}(x) = 0$ for all x. Evaluating this function at the random test input $X_{\mathtt{out}}$ therefore gives $\hat{m}(X_{\mathtt{out}}) = 0$, while the predictive variance remains $\hat{V}(X_{\mathtt{out}})$.

$$\mathrm{SSR} = \sqrt{\frac{\hat{M}}{M_{\mathtt{out}}}} \le 1. \tag{8}$$

Alternatively, if $\hat{m}(x) = m_{\mathtt{out}}(x)$, meaning that there is no mismatch in the predicted residual and the target residual mean, then

$$\mathrm{SSR} = \sqrt{\frac{\mathbb{E}[\hat{V}(X_{\mathtt{out}})]}{\mathbb{E}[V_{\mathtt{out}}(X_{\mathtt{out}})]}} \le 1. \tag{9}$$

More generally, in this optimistic setting, underdispersion holds whenever the cross second moment between the in-sample and out-of-sample residual means satisfies

$$\mathbb{E}[\hat{m}(X_{\mathtt{out}})^\top m_{\mathtt{out}}(X_{\mathtt{out}})] \le \mathbb{E}[\|\hat{m}(X_{\mathtt{out}})\|_2^2] + \frac{M_{\mathtt{out}} - \hat{M}}{2}.$$

The right hand side sums the second moment of the predict residual mean (always positive) with half the residual energy gap (positive in expectation). Overall, the analysis above identifies fitting-induced residual-energy optimism as one mechanism through which predictor-induced residual targets can lead to underdispersion. More generally, though, we see that the difference between the predicted residual distribution relative to the target residual distribution is the source of this problem. This motivates our focus on aligning residual distributions more broadly.

Empirical evidence and implications for ReMatch. To empirically examine the source of this predictive mismatch, we measure the residual target energies $M_{\mathtt{in}}$ and $M_{\mathtt{out}}$ induced by the fitted mean predictor. On the ERA5–HRRR dataset, the samplewise target residual magnitudes yield $M_{\mathtt{in}} = 0.55$ and $M_{\mathtt{out}} = 4.29$, corresponding to a substantial realized target-energy gap of $\Delta M = 3.74$. However, residual-energy mismatch captures only one aspect of residual target misspecification: the full residual distributions also exhibit a large normalized Wasserstein discrepancy of 0.94. This motivates ReMatch to align the residual targets at the distribution level rather than correcting residual energy alone. Transport increases the training residual energy from 0.55 to 2.19, closer to the `out` sample residual target energy, 4.29, while reducing the normalized Wasserstein discrepancy from 0.94 to 0.34. Full distributional statistics are provided in Appendix A.3.

4 REMATCH: RESIDUAL DISTRIBUTION MATCHING

We introduce Residual Distribution Matching (ReMatch), a probabilistic downscaling framework that combines deterministic mean prediction, residual target distribution alignment, and conditional diffusion-based residual generation.

4.1 RESIDUAL DISTRIBUTION MATCHING FOR RESIDUAL TARGET MISSPECIFICATION

Section 3 shows that residual-energy optimism provides one mechanism for underdispersion, but correcting residual energy alone is not sufficient for calibration. In particular, equation 6 shows that calibration depends jointly on the predicted residual energy and the mismatch between the predicted and target conditional residual means.

This motivates a stronger target: matching the conditional residual distribution itself. For a fixed input x, define $Q_x := q_\theta(\hat{R} \mid X = x)$ and $P_x := P(R_{\mathtt{out}} \mid X = x)$. Define the integrated conditional Wasserstein distance as $\mathcal{W}^2_{\mathrm{cond}} := \mathbb{E}_{X_{\mathtt{out}}}\left[W_2^2(Q_{X_{\mathtt{out}}}, P_{X_{\mathtt{out}}})\right]$. We next show that the $\mathcal{W}^2_{\mathrm{cond}}$ controls both components governing the SSR condition in equation 6: the conditional residual-mean mismatch and the residual-energy mismatch.

(i) Conditional residual-mean mismatch. For a fixed x, let $\gamma_x \in \Pi(Q_x, P_x)$ denote a coupling, i.e., a joint distribution whose marginals are Q_x and P_x. Applying Jensen's and Cauchy–Schwarz inequalities to $\mathbb{E}_{\gamma_x}[\hat{R} - R_{\mathtt{out}}]$ and minimizing over all couplings gives

$$\mathbb{E}_{X_{\mathtt{out}}} \|\hat{m}(X_{\mathtt{out}}) - m_{\mathtt{out}}(X_{\mathtt{out}})\|_2^2 \leq \mathcal{W}^2_{\mathrm{cond}}.$$

Thus, $\mathcal{W}_{\mathrm{cond}} \to 0$ implies convergence of the predicted conditional residual mean to the out-of-sample conditional residual mean in L^2. The full derivation is provided in Appendix A.4.

(ii) Residual-energy mismatch. Since the square root of residual energy corresponds to the L^2 norm of the residual, the reverse triangle inequality and minimization over the conditional couplings give

$$\left|\sqrt{\hat{M}} - \sqrt{M_{\mathtt{out}}}\right| \leq \mathcal{W}_{\mathrm{cond}}.$$

Thus, $\mathcal{W}_{\mathrm{cond}} \to 0$ implies $\Delta\hat{M} \to 0$. The full derivation is provided in Appendix A.4.

Together, these results show that conditional residual distribution matching jointly controls the two sources of mismatch appearing in the general SSR condition: the residual-energy gap and the conditional residual-mean mismatch. In particular, exact conditional distribution matching, $Q_x = P_x, \quad \forall x$ implies $\hat{m}(x) = m_{\mathtt{out}}(x)$, $\hat{V}(x) = V_{\mathtt{out}}(x)$ and $\Delta\hat{M} = 0$. Consequently, by Eq. equation 5, $\mathrm{SSR} = 1$. ReMatch approximates this ideal target using held-out calibration residuals as an observable proxy for the out-of-sample correction distribution.

4.2 RESIDUAL DISTRIBUTION MATCHING WITH OPTIMAL TRANSPORT

Since the out-of-sample residual distribution is unavailable during training, ReMatch uses residuals from a held-out calibration split as its observable proxy. With the regression model trained as in Section 2.2, ReMatch then transports the training residual targets toward this calibration regime. The calibration split is excluded from mean-predictor fitting and therefore serves as an observable proxy for the corrections required outside the fitting sample. We use the domain notation $d \in \{\mathtt{tr}, \mathtt{cal}, \mathtt{te}\}$ to denote the training, calibration, and test splits, respectively.

At a high level, ReMatch replaces each training residual with an optimal-transport-weighted combination of selected calibration residuals. Because the desired alignment is conditional on the low-resolution input, the transport cost combines residual similarity with low-resolution conditioning consistency. This discourages transport between samples with incompatible large-scale inputs and provides a practical approximation to the conditional residual alignment motivated above.

Direct optimal transport in the full residual space is computationally expensive and can be statistically unstable for high-dimensional spatial fields (Choi et al., 2023; Balaji et al., 2020). We therefore compute the transport cost in compact PCA representations (Agustsson et al., 2017; Yuan et al., 2024; Liu et al., 2018). We fit separate PCA bases to the concatenated training and calibration

residuals and to their corresponding low-resolution inputs. Let $z_i^{\mathtt{tr}}$ and $z_j^{\mathtt{cal}}$ denote residual PCA representations, and let $\ell_i^{\mathtt{tr}}$ and $\ell_j^{\mathtt{cal}}$ denote LR-conditioning PCA representations. We define

$$c_{ij} = \alpha \|z_i^{\mathtt{tr}} - z_j^{\mathtt{cal}}\|_2^2 + \lambda_{\mathrm{cond}} \|\ell_i^{\mathtt{tr}} - \ell_j^{\mathtt{cal}}\|_2^2.$$

Here, the first term encourages residual distribution alignment, while the second restricts transport to samples with compatible low-resolution conditions.

Using c_{ij}, we solve the optimal transport problem and obtain a transport plan π between the training and calibration residual samples. For each training sample i, let $\mathcal{T}_i^{(k)}$ denote the top-k calibration samples selected according to the transport weights. We renormalize the corresponding weights as $\tilde{\pi}_{ij} = \frac{\pi_{ij}}{\sum_{j' \in \mathcal{T}_i^{(k)}} \pi_{ij'}}, j \in \mathcal{T}_i^{(k)}$, and construct the transported residual representation and residuals:

$$\tilde{z}_i = \sum_{j \in \mathcal{T}_i^{(k)}} \tilde{\pi}_{ij} z_j^{\mathtt{cal}}. \qquad \tilde{r}_i = U_K \tilde{z}_i + \bar{r}, \qquad \tilde{\mu}_i = y_i - \tilde{r}_i.$$

The resulting pair $(\tilde{\mu}_i, \tilde{r}_i)$ preserves the target identity while replacing the original fitting residual target with one shifted toward the held-out correction regime. The choice of k, PCA dimensionality, transport constraints, and additional implementation details are provided in Appendix A.6.

Finally, we train a conditional diffusion model $q_\theta(r \mid x, \mu_\phi(x))$ to generate stochastic residual corrections with the dataset union of the transported training samples and original calibration samples $\mathcal{D}_{\mathrm{res}} = \{(x_i, \tilde{\mu}_i, \tilde{r}_i)\}_{i=1}^{n_{\mathtt{tr}}} \cup \{(x_j, \mu_{\phi^\star}(x_j), r_j^{\mathtt{cal}})\}_{j=1}^{n_{\mathtt{cal}}}$. Algorithm A.5 provides an overview of the three-stage ReMatch training pipeline.

Implementation Note. ReMatch is modular: the mean predictor and stochastic residual generator can be instantiated with different architectures; implementation details are in Appendix A.10.

5 Experiments and Results

We evaluate ReMatch on a controlled synthetic benchmark and a real-world atmospheric downscaling task. The synthetic experiment validates the mechanism of bias-induced residual target misspecification by systematically varying the LR–HR bias level. The real-world experiment tests ReMatch under real ERA5-to-HRRR wind-field distribution shifts. We compare downscaling fidelity and probabilistic calibration, and show downstream impact through a particle-advection experiment, where ensemble trajectories reveal how calibration affects predicted motion through a flow field.

Baselines and Variants. We use two versions of ReMatch: $\text{ReMatch}_{\text{U}}$ with a UNet mean predictor and $\text{ReMatch}_{\text{S}}$ with a SwinIR (Liang et al., 2021) mean predictor. We compare ReMatch against deterministic baselines, including UNet (Song et al., 2020), Conv-FNO (Wen et al., 2022), and SwinIR (Liang et al., 2021), and probabilistic CorrDiff-based baselines. The latter include the original CorrDiff (Mardani et al., 2025) conditional diffusion model, classifier-free guidance on residual diffusion (CFG (Ho & Salimans, 2022)), finetuning baseline adapted from (Couairon et al., 2026) and uncertainty-conditioned variants (UC_q and UC_μ).

Except for SwinLR and Conv-FNO, all regression models are UNet-based (Song et al., 2020), and all diffusion models use an EDM-style residual diffusion formulation (Karras et al., 2022) with a UNet-based score network. We implement conditioning by channel-wise concatenation. Details of each baseline and implementation are provided in Appendix A.10.

Evaluation Metrics. We report RMSE, SSR, and Continuous Ranked Probability Score (CRPS) as primary metrics. RMSE measures reconstruction accuracy, while SSR and CRPS evaluate probabilistic quality; SSR is particularly useful for diagnosing ensemble under-dispersion. We report LPIPS (Zhang et al., 2018) as an auxiliary perceptual metric, following prior use in atmospheric-field prediction (Yu et al., 2024; Hokson et al., 2026). We also report ACC (Nai et al., 2025), computed between predicted and ground-truth (GT) deviations from the training GT mean across channels and times. Metrics in Table 1 are computed using 600 random test samples. We normalize LPIPS channelwise with quantile extrema, and average over ensembles for stochastic-model LPIPS.

5.1 Bias-Controlled Synthetic Benchmark

To validate our theoretical analysis, we conduct a controlled experiment on the BLASTNet dataset (Chung et al., 2023). We inject structured downscaling bias of increasing severity into the low-resolution inputs while keeping the high-resolution targets fixed (details and bias generation procedure in Appendix A.7). We use CorrDiff as a baseline method and ReMatch$_{\mathrm{U}}$ as our proposed method. In all cases, 12 ensemble predictions are generated to calculate spreads and RMSE.

Dataset. We use the BLASTNet Momentum128 3D super-resolution dataset (Chung et al., 2023) as a controlled testbed. The original dataset provides clean paired low- and high-resolution fields with 4 channels. We use the $8\times$ super-resolution setting, with 16×16 LR inputs and 128×128 HR targets. We define four bias levels (Level 0 to 3), where Level 0 uses the original clean interpolated low-resolution field and Levels 1–3 introduce progressively stronger structured mismatch by mixing 2D fields along the z axis with Gaussian weights (full procedure in Appendix A.7). Figure 6 illustrates how the input changes across the four bias levels under same HR target.

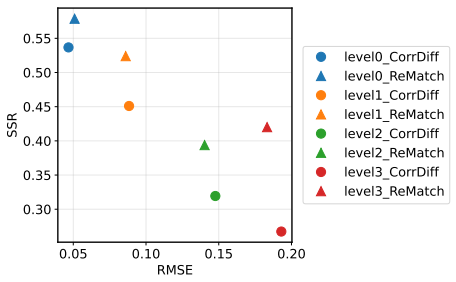


Figure 1: **Performance under increasing LR input bias.** ReMatch consistently improves SSR and reduces RMSE in biased settings, with the performance gap becoming more pronounced as the bias level increases.(SSR$\rightarrow$ 1, RMSE$\downarrow$)

5.1.1 Results

Does residual misspecification explain under-dispersion? Figure 1 shows the RMSE–SSR scatter across increasing LR–HR bias levels. As the bias level increases, the baseline exhibits a clear degradation: RMSE increases while SSR decreases. This trend indicates that stronger LR–HR mismatch induces a more difficult residual correction regime and leads to increasingly under-dispersive ensembles.

Does ReMatch mitigate residual target misspecification? ReMatch improves RMSE in all biased settings except Level 0, where the LR input is not biased. It also consistently improves SSR, which becomes more pronounced at higher bias levels. These results support our central hypothesis: when paired low- and high-resolution data contain structured bias, the mean–residual decomposition can induce residual target misspecification, and residual distribution matching can mitigate this effect. Channel-wise RMSE and SSR results are provided in Appendix A.7.

5.2 Real-world wind field downscaling benchmark

Our main experiment evaluates ReMatch on a real-world atmospheric downscaling task with realistic distribution shift. Deterministic baselines (UNet, ConvFNO, and SwinIR) produce a single prediction, whereas stochastic methods (Diffusion, CorrDiff, CFG, UC$_{\mu}$, UC$_{\mathrm{q}}$, Finetuning, ReMatch$_{\mathrm{U}}$, and ReMatch$_{\mathrm{S}}$) generate 12 ensemble members. Results for UC$_{\mathrm{q}}$ are provided in Appendix A.10.

Dataset. We use ERA5 (Hersbach et al., 2020) as the coarse-resolution input and HRRR (Dowell et al., 2022) as the high-resolution target over the northeast CONUS region. The input contains u- and v-wind fields at five stratospheric pressure levels, giving 10 channels in total and 4 temporal channels for sinusoidal embeddings of sample day and month on a 21×21 grid. The target consists of the same 10 wind-component channels on a 168×168 grid. The dataset contains hourly fields from 2018 to 2021. We use 2018–2020 for training and 2021 for testing; for methods requiring calibration data, we hold out 2020 as a calibration set. Further details are in Appendix A.9.

5.2.1 Results

What does realistic LR–HR mismatch look like? The first column of Fig. 2 illustrates LR–HR mismatch: even for the same atmospheric scene, the LR input and the corresponding HR target differ substantially. In the appendix, Fig. 8 shows more examples where RMSE can quantify this

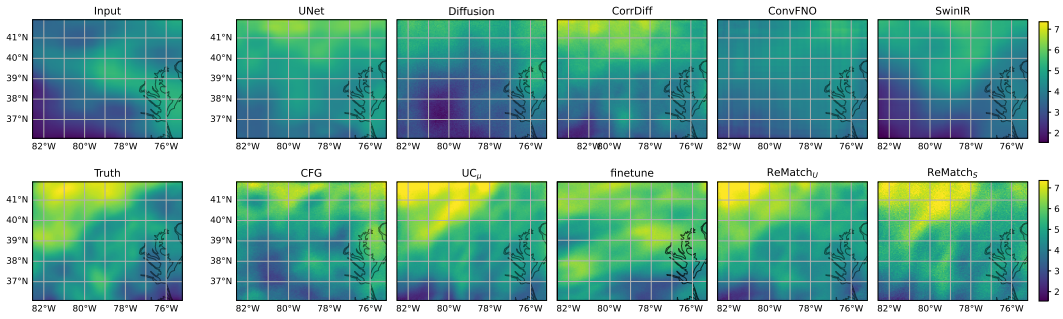


Figure 2: **Qualitative comparison of downscaled 50 hPa v-wind fields on 2021-06-05 10:00:00.** The top-left panel shows the upsampled ERA5 input, and the bottom-left panel shows the corresponding HRRR GT. Deterministic baselines produce a single prediction. Diffusion-based methods generate 12 ensemble members – we only show the 7th ensemble member for stochastic methods.

Table 1: **HRRR-ERA5 pressure-level super-resolution comparison over all channels.**

Metric	UNet	ConvFNO	SwinIR	Diffusion	CFG	UQ$_\mu$	Finetune	CorrDiff	**ReMatch$_U$**	**ReMatch$_S$**
RMSE ↓	1.8517	1.9213	<u>1.7642</u>	1.8570	1.8309	1.8584	2.015	1.8342	1.8102	**1.7194**
SSR → 1	–	–	–	0.6419	0.3418	**0.8745**	0.6499	0.3241	0.6347	<u>0.7150</u>
CRPS/MAE ↓	1.5420	1.6046	1.4538	1.1908	1.2740	1.1545	<u>1.1440</u>	1.2859	1.1567	**1.0976**
LPIPS ↓	0.0737	<u>0.0715</u>	**0.0641**	0.0943	0.0793	0.0871	0.0959	0.0904	0.0841	<u>0.0715</u>
ACC → 1	0.8980	0.8957	0.9016	0.8973	0.8994	0.8926	<u>0.9046</u>	0.8991	0.9017	**0.9051**

Notes: For deterministic models, CRPS is replaced with MAE and SSR is omitted. Best results are bolded, and second-best results are underlined.

mismatch, and Table 4 shows that the amount of bias realistically depends on channel. The cross-source mismatch makes real-world downscaling more challenging than idealized super-resolution.

Do the proposed methods improve downscaling performance? Figure 2 shows that UNet remains biased in overall scale. In contrast, mean–residual methods using UNet as the mean predictor, including CorrDiff, CFG, UC$_\mu$, and ReMatch$_\text{U}$, produce predictions that are closer to the HR target in both spatial structure and magnitude, suggesting that residual generation helps correct errors in the deterministic mean estimate. A similar trend is observed for ReMatch$_\text{S}$, which uses SwinIR as the mean predictor. Overall, ReMatch$_\text{U}$ and ReMatch$_\text{S}$ produce visually competitive reconstructions while better recovering the large-scale magnitude and small-scale spatial patterns of the HR field.

Table 1 reports results on 600 test samples. ReMatch$_\text{S}$ achieves the best RMSE, CRPS, and ACC, consistent with the qualitative improvements in Figure 2. Although UC$_\mu$ achieves the best SSR, it sacrifices performance on other metrics, particularly RMSE, while ReMatch$_\text{S}$ maintains a strong performance across all metrics. SwinIR achieves the best LPIPS but not on other metrics. ReMatch$_\text{U}$, while not the best overall, consistently outperforms the baseline mean–residual model, CorrDiff, across all metrics, demonstrating the benefit of residual distribution matching. Overall, these results show that ReMatch improves the accuracy–calibration tradeoff in real-world wind-field downscaling, with further gains when paired with a more expressive regression model.

How much does the deterministic mean predictor matter? Among deterministic baselines, SwinIR achieves the strongest performance, showing that a stronger mean predictor can improve reconstruction accuracy. However, purely deterministic estimators, including UNet, ConvFNO, and SwinIR cannot quantify uncertainty. This limits their use in downstream tasks where calibrated ensemble spread is essential. ReMatch$_\text{S}$, which adds residual generation on top of the SwinIR mean predictor, further improves both RMSE and CRPS/MAE, suggesting that ReMatch excels with stronger deterministic backbones while providing calibrated stochastic predictions.

6 Discussion

Role of calibration data. ReMatch relies on a held-out calibration set because residual target adaptation is only meaningful when the calibration residuals reflect the correction regime expected at test time. This is a weaker and more practical assumption than requiring access to test labels, but it is

still an important one: ReMatch is not intended to solve arbitrary distribution shift without representative held-out data. In our main experiments, we use a train–calibration–test split of approximately 2:1:1 to provide sufficient calibration samples to estimate the target residual distribution while retaining enough data for training the mean predictor and residual generator. The appropriate split is task-dependent: a larger calibration set can better represent the test residual regime and stabilize residual transport, but it reduces model training data given a fixed total dataset size and increases the computational cost of optimal transport. We empirically examine this trade-off by varying the train-to-calibration ratio on BLASTNet. We found that using more calibration data generally improves SSR while maintaining comparable RMSE, whereas an extremely small calibration set can substantially reduce SSR due to an insufficient representation of the target residual distribution. Detailed discussion of the calibration-set sensitivity are provided in Appendix A.8.

Does ensemble calibration matter for downstream tasks? Predicted physical fields are often used as environments for downstream decision making and agent motion planning. To examine how residual calibration may affect such downstream use, we advect particles through 48 consecutive hourly wind fields.

Figure 3 compares trajectories from the HR ground-truth field, CorrDiff ensembles, and ReMatch$_{\mathrm{U}}$ ensembles. CorrDiff trajectories deviate from the ground-truth path as the particle moves away from the initial position, whereas ReMatch$_{\mathrm{U}}$ remains better centered around the truth while preserving ensemble spread. While this example alone does not establish a general downstream advantage, it suggests that better-calibrated spread can improve the usefulness of ensemble forecasts for downstream uncertainty-aware tasks.

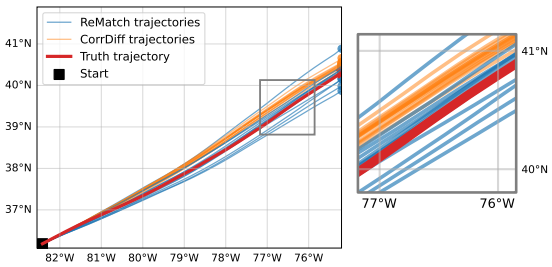


Figure 3: **Particle trajectory comparison using ensemble wind-field predictions over 48 consecutive hourly fields.** The red line denotes the trajectory advected by the HR ground-truth field, orange lines denote CorrDiff ensemble trajectories, and blue lines denote ReMatch$_{\mathrm{U}}$ ensemble trajectories.

Limitations and future directions. ReMatch performs residual matching in a low-dimensional PCA space, which improves tractability but may not fully preserve nonlinear or localized residual structures. Learned or physics-aware representations could provide richer transport spaces. Our current implementation also performs transport independently across output channels, which does not explicitly preserve cross-channel dependencies; joint or structure-aware transport is an important direction for preserving coupled physical structure. Finally, extending ReMatch beyond the mean–residual frameworks studied here, such as to deterministic-guided diffusion, flow matching, or quantile regression, is an interesting direction for future work.

Conclusion. We studied probabilistic downscaling under real-world bias through the lens of residual target misspecification, where a fitted mean predictor induces residual targets whose training distribution can differ from the correction regime required at test time. We showed how this mismatch can lead to underdispersion through residual-energy and conditional residual-mean mismatch, and established that residual distribution matching jointly controls these two sources of miscalibration. Motivated by this connection, we introduced ReMatch, which adapts training residuals toward a held-out calibration regime using optimal transport. Across controlled bias experiments, real-world ERA5-to-HRRR wind downscaling, and a downstream particle-advection example, ReMatch improves the accuracy–calibration tradeoff over strong baselines.

More broadly, our results suggest that calibration failures in residual generative pipelines should not be viewed only as deficiencies of the stochastic generator; they can originate from the predictor-induced targets on which it is trained. This perspective motivates addressing calibration at the target-distribution level, rather than only improving the generative model itself.

AI USE STATEMENT

We used generative AI tools to assist with writing and developing experimental code. Specifically, we used AI tools to edit the manuscript for clarity, identify relevant prior work, implement baseline methods, develop visualization code, and debug experimental code. For the theoretical analysis in Section 3, generative AI tools were also used to discuss mathematical connections and intermediate proof steps.

Generative AI tools were not used to formulate the problem, design the proposed method, or interpret the experimental results. All AI-assisted code, mathematical arguments, citations, and manuscript edits were reviewed and verified by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the datasets, baseline methods, and experimental configurations in the appendix. Due to the large size of the ERA5 and HRRR datasets, we do not distribute the data directly; instead, we describe the data sources and preprocessing procedures required to reproduce the experiments. Detailed descriptions of the datasets and baseline methods are provided in Appendix A.9 and Appendix A.10, respectively. The implementation will be made publicly available after the review process.

REFERENCES

Eirikur Agustsson, Alexander Sage, Radu Timofte, and Luc Van Gool. Optimal transport maps for distribution preserving operations on latent spaces of generative models. *arXiv preprint arXiv:1711.01970*, 2017.

Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pp. 717–730. PMLR, 2022.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

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.

Fuhao Chen and Linyue Gao. Learning residual distributions with diffusion models for probabilistic wind power forecasting. *Energies*, 18(16):4226, 2025. doi: 10.3390/en18164226.

Zehua Chen, Yihan Wu, Yichong Leng, Jiawei Chen, Haohe Liu, Xu Tan, Yang Cui, Ke Wang, Lei He, Sheng Zhao, et al. Resgrad: Residual denoising diffusion probabilistic models for text to speech. *arXiv preprint arXiv:2212.14518*, 2022.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Generative modeling through the semi-dual formulation of unbalanced optimal transport. *Advances in Neural Information Processing Systems*, 36:42433–42455, 2023.

Wai Tong Chung, Bassem Akoush, Pushan Sharma, Alex Tamkin, Ki Sung Jung, Jacqueline Chen, Jack Guo, Davy Brouzet, Mohsen Talei, Bruno Savard, et al. Turbulence in focus: Benchmarking scaling behavior of 3d volumetric super-resolution with blastnet 2.0 data. *Advances in Neural Information Processing Systems*, 36:77430–77484, 2023.

Guillaume Couairon, Renu Singh, Anastase Charantonis, Christian Lessig, and Claire Monteleoni. Archesweathergen: Skillful and compute-efficient probabilistic weather forecasting with machine learning. *Science Advances*, 2026. doi: 10.1126/sciadv.adx2372.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30, 2017a.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017b. doi: 10.1109/TPAMI.2016.2615921.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10648–10656, 2019.

David C Dowell, Curtis R Alexander, Eric P James, Stephen S Weygandt, Stanley G Benjamin, Geoffrey S Manikin, Benjamin T Blake, John M Brown, Joseph B Olson, Ming Hu, et al. The high-resolution rapid refresh (hrrr): An hourly updating convection-allowing forecast model. part i: Motivation and system description. *Weather and Forecasting*, 37(8):1371–1395, 2022.

Bradley Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American statistical Association*, 81(394):461–470, 1986.

Tobias Sebastian Finn, Charlotte Durand, Alban Farchi, Marc Bocquet, Pierre Rampal, and Alberto Carrassi. Generative diffusion for regional surrogate models from sea-ice simulations. *Journal of Advances in Modeling Earth Systems*, 16(10):e2024MS004395, 2024. doi: 10.1029/2024MS004395.

Stathi Fotiadis, Noah D Brenowitz, Tomas Geffner, Yair Cohen, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Adaptive flow matching for resolving small-scale physics. In *Forty-second International Conference on Machine Learning*, 2025.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738, 2015. doi: 10.1007/s00440-014-0583-7.

Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *arXiv preprint arXiv:2402.04290*, 2024.

Zhitao Han and Wenhui Huang. Discrete residual diffusion model for high-resolution prostate mri synthesis. *Physics in Medicine & Biology*, 69(5):055024, 2024.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jose Angelo Hokson, Shinjiro Kanae, and Yusuke Hiraga. Predicting tropical cyclone rainfall through perceptual-based identification of similar atmospheric fields. *Weather and Climate Extremes*, pp. 100896, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. In *International Conference on Learning Representations*, 2019.

Chih-Yu Lai, Yu-Chien Ning, and Duane S Boning. Rdit: Residual-based diffusion implicit models for probabilistic time series forecasting. *arXiv preprint arXiv:2509.02341*, 2025.

Guowen Li, Yang Liu, Shilei Cao, Mengxuan Chen, Xintong Liu, Xuehe Wang, Juepeng Zheng, Yuxiang Li, Zhiyu Ye, Haoyuan Liang, et al. Rescast: Enhancing global medium-range precipitation forecasting with residual diffusion model.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Tianyi Lin, Chenyou Fan, Nhat Ho, Marco Cuturi, and Michael I. Jordan. Projection robust wasserstein distance and riemannian optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9383–9397, 2020.

Huidong Liu, Yang Guo, Na Lei, Zhixin Shu, Shing-Tung Yau, Dimitris Samaras, and Xianfeng Gu. Latent space optimal transport for generative models. *arXiv preprint arXiv:1809.05964*, 2018.

Ignacio Lopez-Gomez, Zhong Yi Wan, Leonardo Zepeda-Núñez, Tapio Schneider, John Anderson, and Fei Sha. Dynamical-generative downscaling of climate model ensembles. *Proceedings of the National Academy of Sciences*, 122(17):e2420288122, 2025.

Douglas Maraun and Martin Widmann. *Statistical downscaling and bias correction for climate research*. Cambridge University Press, 2018.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1):124, 2025.

Roberto Molinaro, Niall Siegenheim, Henry Martin, Mark Frey, Niels Poulsen, Philipp Seitz, and Marvin Vincent Gabler. Universal diffusion-based probabilistic downscaling. *arXiv preprint arXiv:2602.11893*, 2026.

Congyi Nai, Xi Chen, Shangshang Yang, Ziniu Xiao, and Baoxiang Pan. Boosting weather forecast via generative superensemble. *npj Climate and Atmospheric Science*, 8(1):377, 2025.

Thao Nguyen, Jiaqi Ma, Fahad Shahbaz Khan, Souhaib Ben Taieb, and Salman Khan. Raindiff: End-to-end precipitation nowcasting via token-wise attention diffusion. *arXiv preprint arXiv:2510.14962*, 2025.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5072–5081. PMLR, 2019.

Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

Neelesh Rampal, Sanaa Hobeichi, Peter B Gibson, Jorge Baño-Medina, Gab Abramowitz, Tom Beucler, Jose González-Abad, William Chapman, Paula Harder, and José Manuel Gutiérrez. Enhancing regional climate downscaling through advances in machine learning. *Artificial Intelligence for the Earth Systems*, 3(2):230066, 2024.

Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 849–858. PMLR, 2019.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Maximilian Springenberg, Noelia Otero, Yuxin Xue, and Jackie Ma. Diffscale: Continuous downscaling and bias correction of subseasonal wind speed forecasts using diffusion models. *Journal of Advances in Modeling Earth Systems*, 18(3):e2025MS005282, 2026.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, 2012.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Zhong Yi Wan, Ricardo Baptista, Anudhyan Boral, Yi-Fan Chen, John Anderson, Fei Sha, and Leonardo Zepeda-Núñez. Debias coarsely, sample conditionally: Statistical downscaling through optimal transport and probabilistic diffusion models. *Advances in Neural Information Processing Systems*, 36:47749–47763, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. In *Bernoulli*, volume 25, pp. 2620–2648, 2019. doi: 10.3150/18-BEJ1065.

Gege Wen, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M Benson. U-fno—an enhanced fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, 163:104180, 2022.

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27758–27767, 2024.

Yubai Yuan, Babak Shahbaba, Norbert Fortin, Keiland Cooper, Qing Nie, and Annie Qu. Optimal transport for latent integration with an application to heterogeneous neuronal activity data. *arXiv preprint arXiv:2407.00099*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A Appendix

A.1 Related Work

Statistical and deep learning downscaling. Statistical downscaling models high-resolution fields from coarse inputs using learned relationships between resolutions (Maraun & Widmann, 2018). Classical approaches such as bias correction and quantile mapping are effective for marginal statistics, but often struggle to preserve spatial coherence and multivariate dependencies. Deep neural networks have improved deterministic downscaling accuracy through convolutional, transformer-based, and operator-learning architectures, including SwinIR (Liang et al., 2021) and U-FNO (Wen et al., 2022). However, deterministic predictors produce only a single estimate and therefore cannot represent the spread of physically plausible high-resolution states. This has motivated probabilistic downscaling methods based on modern generative models (Rampal et al., 2024), including flow matching (Fotiadis et al., 2025) and diffusion-based approaches (Wan et al., 2023; Springenberg et al., 2026), which aim to generate calibrated ensembles for uncertainty-aware downstream applications.

Diffusion-based mean–residual downscaling. Diffusion models have recently become a strong framework for probabilistic atmospheric downscaling because they can model complex high-dimensional conditional distributions. CorrDiff (Mardani et al., 2025) is a representative mean–residual approach: a deterministic regression model first predicts a high-resolution mean field, and a conditional diffusion model then generates stochastic residual corrections. Similar residual or two-stage formulations appear in precipitation nowcasting and forecasting models such as DiffCast (Yu et al., 2024), RainDiff (Nguyen et al., 2025), ResCast (Li et al.), and CasCast (Gong et al., 2024). While this decomposition reduces the variance that the diffusion model must learn, prior work reports persistent under-dispersion in generated ensembles (Mardani et al., 2025; Fotiadis et al., 2025). Our work studies this failure mode directly and argues that it arises from predictor-induced residual target misspecification, rather than only from limitations of the stochastic sampler.

Cross-source/model mismatch vs. predictor-induced target mismatch. Recent work has also considered robustness across heterogeneous climate and weather data sources. R2-D2 (Lopez-Gomez et al., 2025) combines dynamical and generative downscaling, using an intermediate WRF simulation to map outputs from different Earth-system models to a common effective representation before learning a residual diffusion model. Molinaro et al. (2026) instead train a model-agnostic conditional diffusion downscaler that generalizes zero-shot across heterogeneous upstream forecasting systems. These approaches demonstrate that generative downscaling can generalize across source models, but address source mismatch through common physical preprocessing or source-agnostic conditional generation. ReMatch targets a complementary failure mode: the mismatch between the residual distribution induced on samples used to fit an upstream mean predictor and the correction distribution required on held-out samples. Rather than removing source dependence from the input representation, ReMatch explicitly aligns these predictor-induced residual targets using held-out calibration data.

Distribution matching and domain adaptation. Our approach is related to domain adaptation, where source and target distributions are aligned to improve transfer under distribution shift (Ben-David et al., 2010; Pan & Yang, 2010). Classical covariate-shift and importance-weighting methods reweight source samples to match the target input distribution under the assumption that the conditional target remains unchanged (Shimodaira, 2000; Sugiyama & Kawanabe, 2012), while OT-based domain adaptation aligns source and target feature-label distributions through a learned transport plan (Courty et al., 2017b;a; Redko et al., 2019). In downscaling, prior work uses OT to debias the coarse input before conditional diffusion sampling (Wan et al., 2023). ReMatch differs in that it does not primarily debias the input distribution. Instead, it aligns the predictor-induced residual target distribution that the stochastic generator is trained to model.

Distribution matching with projected transport. Optimal transport provides a principled framework for comparing and aligning probability distributions (Villani et al., 2009; Peyré & Cuturi, 2019). Entropic regularization improves the computational scalability of empirical OT (Cuturi, 2013). However, estimating Wasserstein distances in high-dimensional spaces is statistically chal-

lenging, with convergence rates that deteriorate with dimension (Fournier & Guillin, 2015; Weed & Bach, 2019). A common remedy is to compare distributions through lower-dimensional projections or compact representations, as in sliced Wasserstein distances (Bonneel et al., 2015), max-sliced and subspace-robust variants (Deshpande et al., 2019; Paty & Cuturi, 2019; Lin et al., 2020), and latent-space transport or regularization methods (Tolstikhin et al., 2018; Kolouri et al., 2019). ReMatch follows this projection-based perspective by computing transport costs in a latent representation of residuals and conditioning inputs, rather than directly in the full residual field. This provides a tractable way to match residual correction targets, while suggesting that richer physics-aware representations or scalable projected distribution-matching objectives could further improve residual alignment in future work.

A.2 Additional Theory for Residual Target Misspecification

A.2.1 Residual Optimism under Empirical Risk Minimization

Let $S = \{Z_i\}_{i=1}^n$, where $Z_i = (X_i, Y_i) \overset{\text{iid}}{\sim} P$, and define the squared loss, empirical risk, and population risk as

$$\ell(\phi; Z) := \|Y - \mu_\phi(X)\|_2^2, \qquad L_S(\phi) := \frac{1}{n}\sum_{i=1}^n \ell(\phi; Z_i), \qquad L_P(\phi) := \mathbb{E}_{Z\sim P}[\ell(\phi; Z)].$$

Suppose the mean predictor is obtained by exact empirical risk minimization,

$$\hat{\phi}_S \in \arg\min_{\phi\in\Phi} L_S(\phi).$$

Under the residual-energy definition in equation 2, the in-sample and out-of-sample residual energies for a realized fitting sample $\mathcal{S}$ correspond to

$$M_{\texttt{in}} = L_S(\hat{\phi}_S), \qquad M_{\texttt{out}} = L_P(\hat{\phi}_S).$$

where, as in the main text, dependence on $\mathcal{S}$ is suppressed. Averaging over the randomness of the fitting sample, we obtain the following result.

Proposition. If $\hat{\phi}_S$ is an empirical risk minimizer trained on $S \sim P^n$, then

$$\mathbb{E}_{\mathcal{S}}[M_{\texttt{in}}] \le \mathbb{E}_{\mathcal{S}}[M_{\texttt{out}}]. \tag{10}$$

Proof. Let $S' \sim P^n$ be an independent copy of S, and let $\hat{\phi}_{S'}$ be the corresponding empirical risk minimizer. Since $\hat{\phi}_S$ minimizes empirical risk on S,

$$L_S(\hat{\phi}_S) \le L_S(\hat{\phi}_{S'}).$$

Taking expectation over S and S' yields

$$\mathbb{E}_{S,S'}[L_S(\hat{\phi}_S)] \le \mathbb{E}_{S,S'}[L_S(\hat{\phi}_{S'})].$$

Since $L_S(\hat{\phi}_S)$ does not depend on S', the left-hand side becomes

$$\mathbb{E}_{S,S'}[L_S(\hat{\phi}_S)] = \mathbb{E}_S[L_S(\hat{\phi}_S)] = \mathbb{E}_{\mathcal{S}}[M_{\texttt{in}}].$$

For the right-hand side, conditional on S', the predictor $\hat{\phi}_{S'}$ is independent of S. Since the observations in S are i.i.d. draws from P,

$$\mathbb{E}_S[L_S(\hat{\phi}_{S'}) \mid S'] = L_P(\hat{\phi}_{S'}).$$

Taking expectation over S' then gives

$$\mathbb{E}_{S,S'}[L_S(\hat{\phi}_{S'})] = \mathbb{E}_{S'}[L_P(\hat{\phi}_{S'})].$$

Because S and S' are identically distributed,

$$\mathbb{E}_{S'}[L_P(\hat{\phi}_{S'})] = \mathbb{E}_S[L_P(\hat{\phi}_S)] = \mathbb{E}_{\mathcal{S}}[M_{\texttt{out}}].$$

Combining the two sides gives

$$\mathbb{E}_{\mathcal{S}}[M_{\texttt{in}}] \le \mathbb{E}_{\mathcal{S}}[M_{\texttt{out}}].$$

□ □

The result requires no distribution shift: both the fitting sample and the independent observation defining L_P are drawn from the same distribution P. It requires neither linearity nor Gaussianity. The inequality is non-strict because equality is possible when fitting does not benefit from adaptation to the observed sample. The result assumes exact ERM; predictors obtained by approximate or finite-step optimization need not satisfy the inequality exactly.

A.2.2 A LINEAR-GAUSSIAN ILLUSTRATION

To illustrate a setting in which the inequality is strict, consider

$$Y = X\beta^\star + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

where $X \in \mathbb{R}^{n\times p}$ has full column rank. Ordinary least squares gives

$$\hat{Y} = HY, \qquad H = X(X^\top X)^{-1}X^\top.$$

The training residual is

$$R_{\texttt{in}} = (I - H)\epsilon.$$

Conditional on X,

$$\mathbb{E}\left[\|R_{\texttt{in}}\|_2^2 \mid X\right] = \sigma^2 \operatorname{tr}(I - H) = \sigma^2(n - p),$$

where $\operatorname{tr}(H) = p$. Therefore, the expected residual energy per observation is

$$\mathbb{E}_{\mathcal{S}}[M_{\texttt{in}}] = \sigma^2 \left(1 - \frac{p}{n}\right). \tag{11}$$

For a fresh observation $y_0 = x_0^\top \beta^\star + \epsilon_0$ independent of the training sample,

$$y_0 - x_0^\top \hat{\beta} = \epsilon_0 - x_0^\top (X^\top X)^{-1} X^\top \epsilon,$$

and hence

$$\mathbb{E}\left[(y_0 - x_0^\top \hat{\beta})^2 \mid X, x_0\right] = \sigma^2 \left[1 + x_0^\top (X^\top X)^{-1} x_0\right].$$

If the rows of X and x_0 are i.i.d. $\mathcal{N}(0, \Sigma)$, then

$$X^\top X \sim \mathcal{W}_p(\Sigma, n),$$

and, for $n > p + 1$,

$$\mathbb{E}[(X^\top X)^{-1}] = \frac{\Sigma^{-1}}{n - p - 1}.$$

Since $\mathbb{E}[x_0 x_0^\top] = \Sigma$,

$$\mathbb{E}\left[x_0^\top (X^\top X)^{-1} x_0\right] = \operatorname{tr}\left(\mathbb{E}[(X^\top X)^{-1}]\Sigma\right) = \frac{p}{n - p - 1}.$$

Therefore,

$$\mathbb{E}_{\mathcal{S}}[M_{\texttt{out}}] = \sigma^2 \left(1 + \frac{p}{n - p - 1}\right). \tag{12}$$

For $p > 0$ and $n > p + 1$,

$$\mathbb{E}_{\mathcal{S}}[M_{\texttt{in}}] = \sigma^2 \left(1 - \frac{p}{n}\right) < \sigma^2 < \sigma^2 \left(1 + \frac{p}{n - p - 1}\right) = \mathbb{E}_{\mathcal{S}}[M_{\texttt{out}}].$$

Thus, fitting alone can produce a strict residual-energy gap even when fitting and unseen observations follow exactly the same data-generating distribution. This example is illustrative only; equation 10 itself requires neither linearity nor Gaussianity.

A.2.3 PROOF OF THE SPREAD–SKILL RELATION

Consider an out-of-sample input–target pair (X, Y) and let

$$\hat{Y} = \mu_\phi(X) + \hat{R}, \qquad \hat{R} \sim q_\theta(\cdot \mid X, \mu_\phi(X)),$$

with the conditional mean and variance of the predicted residual defined as

$$\hat{m}(x) = \mathbb{E}_{q_\theta}[\hat{R} \mid X = x], \qquad \hat{V}(x) = \operatorname{tr} \operatorname{Cov}_{q_\theta}(\hat{R} \mid X = x).$$

Because $\mu_\phi(X)$ is deterministic conditional on X,

$$\operatorname{tr} \operatorname{Cov}_{q_\theta}(\hat{Y} \mid X = x) = \operatorname{tr} \operatorname{Cov}_{q_\theta}(\hat{R} \mid X = x) = \hat{V}(x).$$

Therefore,

$$\mathrm{Spread}^2 = \mathbb{E}[\hat{V}(X_{\mathtt{out}})].$$

The conditional ensemble mean is

$$\mathbb{E}[\hat{Y} \mid X] = \mu_\phi(X) + \hat{m}(X).$$

For an out-of-sample target,

$$Y_{\mathtt{out}} = \mu_\phi(X_{\mathtt{out}}) + R_{\mathtt{out}},$$

and hence

$$Y_{\mathtt{out}} - \mathbb{E}[\hat{Y} \mid X_{\mathtt{out}}] = R_{\mathtt{out}} - \hat{m}(X_{\mathtt{out}}).$$

Applying the conditional bias–variance decomposition gives

$$\mathbb{E}\left[\|Y_{\mathtt{out}} - \mathbb{E}[\hat{Y} \mid X_{\mathtt{out}}]\|_2^2 \mid X_{\mathtt{out}} = x\right] = V_{\mathtt{out}}(x) + \|m_{\mathtt{out}}(x) - \hat{m}(x)\|_2^2.$$

Taking expectation over $X_{\mathtt{out}}$ gives

$$\mathrm{Skill}^2 = \mathbb{E}\left[V_{\mathtt{out}}(X_{\mathtt{out}}) + \|m_{\mathtt{out}}(X_{\mathtt{out}}) - \hat{m}(X_{\mathtt{out}})\|_2^2\right].$$

Therefore,

$$\mathrm{SSR}^2 = \frac{\mathbb{E}[\hat{V}(X_{\mathtt{out}})]}{\mathbb{E}\left[V_{\mathtt{out}}(X_{\mathtt{out}}) + \|m_{\mathtt{out}}(X_{\mathtt{out}}) - \hat{m}(X_{\mathtt{out}})\|_2^2\right]}.$$

Since $\mathrm{SSR} \le 1$ if and only if $\mathrm{Skill}^2 - \mathrm{Spread}^2 \ge 0$, we have

$$\begin{aligned}\mathrm{Skill}^2 - \mathrm{Spread}^2 &= \mathbb{E}\left[V_{\mathtt{out}} - \hat{V} + \|m_{\mathtt{out}} - \hat{m}\|_2^2\right] \\ &= \mathbb{E}\left[V_{\mathtt{out}} - \hat{V} + \|m_{\mathtt{out}}\|_2^2 - \|\hat{m}\|_2^2\right] + 2\mathbb{E}\left[\|\hat{m}\|_2^2 - \hat{m}^\top m_{\mathtt{out}}\right], \qquad (13)\end{aligned}$$

where all quantities inside the expectations are evaluated at $X_{\mathtt{out}}$.

Define the predicted residual energy as

$$\hat{M} := \mathbb{E}\left[\hat{V}(X_{\mathtt{out}}) + \|\hat{m}(X_{\mathtt{out}})\|_2^2\right].$$

Since

$$M_{\mathtt{out}} = \mathbb{E}\left[V_{\mathtt{out}}(X_{\mathtt{out}}) + \|m_{\mathtt{out}}(X_{\mathtt{out}})\|_2^2\right],$$

we define the *predictive residual-energy gap* as

$$\Delta\hat{M} := M_{\mathtt{out}} - \hat{M}.$$

Using this notation, Eq. equation 13 becomes

$$\mathrm{Skill}^2 - \mathrm{Spread}^2 = \Delta\hat{M} + 2\mathbb{E}\left[\hat{m}^\top(\hat{m} - m_{\mathtt{out}})\right].$$

Thus, the exact condition for underdispersion is

$$\boxed{\mathrm{SSR} \le 1 \quad \Longleftrightarrow \quad \mathbb{E}\left[\hat{m}^\top(m_{\mathtt{out}} - \hat{m})\right] \le \frac{\Delta\hat{M}}{2}.}$$

Connection to ERM residual optimism. Equation equation 4 establishes an optimism gap in the residual targets themselves,

$$\mathbb{E}_{\mathcal{S}}[M_{\mathtt{in}}] \le \mathbb{E}_{\mathcal{S}}[M_{\mathtt{out}}].$$

The predictive residual energy $\hat{M}$, in contrast, is defined by the learned residual generator evaluated on out-of-sample inputs and therefore need not equal $M_{\mathtt{in}}$. The ERM result thus identifies a mechanism by which the residual target available during fitting can systematically underestimate the correction energy required out of sample. For a learned residual generator that inherits this target-level optimism, $\Delta\hat{M} > 0$, the exact SSR condition above characterizes when this predictive energy gap leads to underdispersion.

Thus, after removing residual-predictor approximation error, the predictive residual-energy gap reduces to the fitting-induced residual-energy optimism established by ERM. For a fixed fitting sample with $\Delta\hat{M} \geq 0$, the exact underdispersion condition becomes

$$\boxed{\mathbb{E}\left[\hat{m}^\top(m_{\text{out}} - \hat{m})\right] \leq \frac{\Delta\hat{M}}{2}.}$$

This condition includes several useful regimes in which the fitting-induced residual-energy gap leads to underdispersion. First, if $\hat{m}(x) = 0$, the mean-dependent term vanishes and $\Delta\hat{M} \geq 0$ is sufficient for $\text{SSR} \leq 1$. Second, if $\hat{m}(x) = m_{\text{out}}(x)$, the mean-dependent term also vanishes and $\Delta\hat{M} \geq 0$ is again sufficient for $\text{SSR} \leq 1$. More generally, neither equality is necessary, and underdispersion holds whenever

$$\mathbb{E}\left[\hat{m}^\top m_{\text{out}}\right] \leq \mathbb{E}\left[\|\hat{m}\|_2^2\right] + \frac{\Delta\hat{M}}{2}.$$

Thus, the in-sample and out-of-sample residual means may be positively aligned and need not match exactly, as long as their cross second moment does not exceed the in-sample residual-mean second moment by more than half of the residual-energy gap.

Special case 1: negligible in-sample residual mean. Consider first the case in which the predicted residual mean is negligible, $\hat{m}(x) \approx 0$. The residual-energy decomposition gives

$$\hat{M} = \mathbb{E}\left[\hat{V}(X_{\text{out}}) + \|\hat{m}(X_{\text{out}})\|_2^2\right] \approx \mathbb{E}[\hat{V}(X_{\text{out}})],$$

so the numerator of the SSR expression is approximately $\hat{M}$. Under the same assumption,

$$\begin{aligned}\mathbb{E}\left[V_{\text{out}}(X_{\text{out}}) + \|m_{\text{out}}(X_{\text{out}}) - \hat{m}(X_{\text{out}})\|_2^2\right] \\ \approx \mathbb{E}\left[V_{\text{out}}(X_{\text{out}}) + \|m_{\text{out}}(X_{\text{out}})\|_2^2\right] = M_{\text{out}}.\end{aligned}$$

Therefore,

$$\text{SSR} \approx \sqrt{\frac{\hat{M}}{M_{\text{out}}}},$$

which recovers Eq. equation 8 in Section 3. This case shows that when the fitted mean predictor leaves negligible systematic residual bias on its fitting samples, the spread–skill relation is governed directly by the train–test residual-energy gap.

Special case 2: matched conditional residual means. Consider instead the case in which the conditional residual mean generalizes across the fitting and out-of-sample regimes, $\hat{m}(x) \approx m_{\text{out}}(x)$. The mean-mismatch contribution to the skill then approximately vanishes, giving

$$\text{SSR}^2 \approx \frac{\mathbb{E}[\hat{V}(X_{\text{out}})]}{\mathbb{E}[V_{\text{out}}(X_{\text{out}})]}.$$

Thus, when the conditional residual means are matched across the two regimes, underdispersion is governed directly by the conditional residual-variance gap. In the exact case $\hat{m}(x) = m_{\text{out}}(x)$, the mean-energy terms also cancel from the predictive residual-energy gap,

$$\Delta\hat{M} = M_{\text{out}} - \hat{M} = \mathbb{E}\left[V_{\text{out}}(X_{\text{out}}) - \hat{V}(X_{\text{out}})\right].$$

Hence, for a fixed fitting sample,

$$\text{SSR} \leq 1 \quad \Longleftrightarrow \quad \mathbb{E}[V_{\text{out}}(X_{\text{out}})] \geq \mathbb{E}[\hat{V}(X_{\text{out}})].$$

Averaged over fitting samples, the ERM residual-energy optimism therefore reduces to a residual-variance optimism in this special case.

Table 2: In–out distribution mismatch of the target residuals and upstream mean predictions on the ERA5–HRRR dataset 5.2. **Parentheses ( )** report the corresponding values after transporting the training residuals toward the held-out calibration residual regime. Distributional differences are measured using Wasserstein (W_2), normalized Wasserstein, Jensen–Shannon (JS), and Kolmogorov–Smirnov (KS) distances. Out-of-sample statistics are unchanged by transport.

Variable	$\text{Mean}_{\texttt{in}}$	$\text{Mean}_{\texttt{out}}$	$\text{Std}_{\texttt{in}}$	$\text{Std}_{\texttt{out}}$	W_2	Norm. W_2	JS	KS
$r_\phi(x, y)$	0.73 (1.43)	2.01	0.13 (0.38)	0.50	1.28 (0.58)	0.94 (0.34)	0.98 (0.61)	0.99 (0.59)
$\mu_\phi(x)$	14.98 (15.03)	14.22	7.74 (7.72)	7.19	0.89 (0.90)	0.06 (0.06)	0.23 (0.23)	0.09 (0.08)

A.3 Empirical Characterization of Residual Target Misspecification

We complement the theoretical analysis in Section 3 by directly characterizing the mismatch between the residual targets induced on the fitting and held-out samples. This analysis concerns the target residuals $R_{\texttt{in}}$ and $R_{\texttt{out}}$

For each sample, we compute the magnitude of the target residual $r_\phi(x, y) = y - \mu_\phi(x)$ and report its empirical mean and standard deviation across samples. Because the residual energy is defined as $M_d = \mathbb{E}[\|R_d\|_2^2]$, the corresponding empirical second moment can be recovered from these statistics as

$$M_d = \mathbb{E}[\|R_d\|_2^2] = \text{Mean}(\|R_d\|_2)^2 + \text{Std}(\|R_d\|_2)^2. \tag{14}$$

As shown in Table 2, the target residual magnitude increases from 0.73 ± 0.13 on the `in` samples to 2.01 ± 0.50 on the `out` samples. These statistics correspond to

$$M_{\texttt{in}} = 0.73^2 + 0.13^2 = 0.55, \qquad M_{\texttt{out}} = 2.01^2 + 0.50^2 = 4.29,$$

yielding a realized residual-energy gap of

$$\Delta M = M_{\texttt{out}} - M_{\texttt{in}} = 3.74. \tag{15}$$

The direction of this realized gap is consistent with the fitting-induced residual-energy optimism derived in Section 3, which establishes the corresponding inequality in expectation over fitting samples. The mismatch is not limited to residual energy. The full residual target distributions exhibit a Wasserstein distance of 1.28, corresponding to a normalized Wasserstein discrepancy of 0.94, together with large Jensen–Shannon and Kolmogorov–Smirnov distances. In comparison, the upstream mean predictions exhibit substantially smaller distributional differences, including a normalized Wasserstein discrepancy of only 0.06. These observations indicate that the residual targets exhibit distributional differences beyond the second-moment mismatch characterized by our theoretical analysis.

ReMatch addresses this broader mismatch by transporting the training residual targets toward residuals obtained from the held-out calibration regime. After transport, the residual magnitude becomes 1.43 ± 0.38, corresponding to an empirical residual energy of

$$\tilde{M}_{\texttt{in}} = 1.43^2 + 0.38^2 = 2.19.$$

Thus, transport moves the residual energy from 0.55 toward the held-out value of 4.29 while simultaneously reducing the normalized Wasserstein discrepancy from 0.94 to 0.34. The Wasserstein, Jensen–Shannon, and Kolmogorov–Smirnov distances all decrease after transport, showing that ReMatch does not merely rescale residual energy but moves the training residual targets toward the held-out residual distribution more broadly.

A.4 Distribution matching for residual target misspecification

Wasserstein distance controls the conditional residual-mean mismatch. For any coupling $\gamma_x \in \Pi(Q_x, P_x)$,

$$\begin{aligned}
\|\hat{m}(x) - m_{\texttt{out}}(x)\|_2 &= \left\| \mathbb{E}_{\gamma_x}[\hat{R} - R_{\texttt{out}}] \right\|_2 \\
&\le \mathbb{E}_{\gamma_x} \|\hat{R} - R_{\texttt{out}}\|_2 && \text{(Jensen's inequality)} \\
&\le \left(\mathbb{E}_{\gamma_x} \|\hat{R} - R_{\texttt{out}}\|_2^2 \right)^{1/2} && \text{(Cauchy–Schwarz inequality).}
\end{aligned}$$

Taking the infimum over all couplings, by the definition of the 2-Wasserstein distance,

$$W_2^2(Q_x, P_x) := \inf_{\gamma_x \in \Pi(Q_x, P_x)} \mathbb{E}_{\gamma_x} \|\hat{R} - R_{\text{out}}\|_2^2,$$

gives

$$\|\hat{m}(x) - m_{\text{out}}(x)\|_2 \le W_2(Q_x, P_x).$$

Averaging the squared inequality over X_{out} and using the definition

$$\mathcal{W}_{\text{cond}}^2 := \mathbb{E}_{X_{\text{out}}} \left[W_2^2(Q_{X_{\text{out}}}, P_{X_{\text{out}}}) \right]$$

therefore gives

$$\boxed{\mathbb{E}_{X_{\text{out}}} \|\hat{m}(X_{\text{out}}) - m_{\text{out}}(X_{\text{out}})\|_2^2 \le \mathcal{W}_{\text{cond}}^2.}$$

Hence, $\mathcal{W}_{\text{cond}} \to 0$ implies convergence of the predicted conditional residual mean to the out-of-sample conditional residual mean in L^2.

Wasserstein distance controls the residual-energy mismatch. Consider a joint coupling obtained by coupling the conditional residual distributions Q_x and P_x for each x. The reverse triangle inequality in L^2 gives

$$\begin{aligned} \left| \sqrt{\hat{M}} - \sqrt{M_{\text{out}}} \right| &= \left| \sqrt{\mathbb{E}\|\hat{R}\|_2^2} - \sqrt{\mathbb{E}\|R_{\text{out}}\|_2^2} \right| \\ &\le \left(\mathbb{E}\|\hat{R} - R_{\text{out}}\|_2^2 \right)^{1/2}. \end{aligned}$$

Taking the optimal conditional couplings yields

$$\boxed{\left| \sqrt{\hat{M}} - \sqrt{M_{\text{out}}} \right| \le \mathcal{W}_{\text{cond}}.}$$

Thus, $\mathcal{W}_{\text{cond}} \to 0$ also implies $\Delta\hat{M} \to 0$.

Exact conditional residual distribution matching,

$$Q_x = P_x \qquad \forall x,$$

simultaneously gives

$$\hat{m}(x) = m_{\text{out}}(x), \qquad \hat{V}(x) = V_{\text{out}}(x), \qquad \Delta\hat{M} = 0.$$

Substituting these relations into Eq. equation 5 gives

$$\text{SSR}^2 = \frac{\mathbb{E}[V_{\text{out}}(X_{\text{out}})]}{\mathbb{E}[V_{\text{out}}(X_{\text{out}})]} = 1.$$

Thus, conditional residual distribution matching provides a stronger calibration target than correcting residual energy alone: it jointly controls the residual-energy gap and conditional residual-mean mismatch appearing in the general SSR condition. ReMatch approximates this target using held-out calibration residuals as an observable proxy for the out-of-sample correction distribution.

A.5 Algorithm Overview

A.6 Residual Distribution Matching via Optimal Transport

Directly computing optimal transport over full residual fields is computationally expensive and can be overly sensitive to grid-scale noise. We therefore compute the transport cost in compact PCA representations of the residual targets and low-resolution conditioning inputs Agustsson et al. (2017); Yuan et al. (2024); Liu et al. (2018). The PCA bases are fitted using the concatenated training and calibration samples, and each field is projected onto the leading components before transport.

For PCA truncation, we choose the number of modes so that the reconstruction fidelity exceeds 99% for all channels. For the BlastNet dataset Chung et al. (2023), we use 200 modes for the high-resolution residual field and 70 modes for the low-resolution conditioning field. For the HRRR

Algorithm 1 ReMatch training pipeline

Input: $\mathcal{D}_{\mathtt{tr}} = \{(x_i, y_i)\}_{i=1}^{n_{\mathtt{tr}}}$, $\mathcal{D}_{\mathtt{cal}} = \{(x_j, y_j)\}_{j=1}^{n_{\mathtt{cal}}}$
Stage 1. Mean training:
  $\phi^\star \leftarrow \arg\min_\phi \mathcal{L}_{\mathrm{reg}}(\phi; \mathcal{D}_{\mathtt{tr}})$
Stage 2. Residual target alignment:
  Compute residuals $r_i^{\mathtt{tr}} = y_i - \mu_{\phi^\star}(x_i)$ and $r_j^{\mathtt{cal}} = y_j - \mu_{\phi^\star}(x_j)$
  Project onto PCA: $z_i^{\mathtt{tr}} = U_K^\top r_i^{\mathtt{tr}}, \quad z_j^{\mathtt{cal}} = U_K^\top r_j^{\mathtt{cal}}, \quad \ell_i^{\mathtt{tr}} = V^\top x_i, \ell_j^{\mathtt{cal}} = V^\top x_j$
  Define cost $c_{ij} = \alpha \|z_i^{\mathtt{tr}} - z_j^{\mathtt{cal}}\|_2^2 + \lambda_{\mathrm{cond}} \|\ell_i^{\mathtt{tr}} - \ell_j^{\mathtt{cal}}\|_2^2$
  Compute sparse transport plan π
  Transport: select $\mathcal{T}_i$, $\tilde{z}_i^{\mathtt{tr}} = \sum_{j \in \mathcal{T}_i} \tilde{\pi}_{ij} z_j^{\mathrm{cal}}$, $\tilde{r}_i = U_K \tilde{z}_i^{\mathtt{tr}}, \quad \tilde{\mu}_i = y_i - \tilde{r}_i$
  Define residual-learning dataset $\mathcal{D}_{\mathrm{res}} = \{(x_i, \tilde{\mu}_i, \tilde{r}_i)\}_{i=1}^{n_{\mathtt{tr}}} \cup \{(x_j, \mu_{\phi^\star}(x_j), r_j^{\mathtt{cal}})\}_{j=1}^{n_{\mathtt{cal}}}$.
Stage 3. Residual learning:
  $\theta^\star \leftarrow \arg\min_\theta \mathcal{L}_{\mathrm{res}}(\theta; \mathcal{D}_{\mathrm{res}})$
Return: $\mu_{\phi^\star}$, $q_{\theta^\star}$

dataset Dowell et al. (2022), we use 100 residual modes and for ERA5 Hersbach et al. (2020), we use 30 modes. BlastNet requires more modes because it contains a wider variety of spatial patterns induced by varying PDE coefficients and flow conditions.

For a residual sample r_ϕ, let

$$z = U_K^\top (r_\phi - \bar{r}) \in \mathbb{R}^K$$

denote its residual PCA representation, where U_K is the residual PCA basis and $\bar{r}$ is the residual mean used for centering. Similarly, for a low-resolution input x, let

$$\ell = V_{K_{\mathrm{cond}}}^\top (x - \bar{x}) \in \mathbb{R}^{K_{\mathrm{cond}}}$$

denote its conditioning representation, where $V_{K_{\mathrm{cond}}}$ is the PCA basis for the low-resolution input. Let

$$\{z_i^{\mathtt{tr}}, \ell_i^{\mathtt{tr}}\}_{i=1}^{n_{\mathtt{tr}}} \quad \text{and} \quad \{z_j^{\mathtt{cal}}, \ell_j^{\mathtt{cal}}\}_{j=1}^{n_{\mathtt{cal}}}$$

denote the resulting low-dimensional representations for the training and calibration sets, respectively.

We compute transport separately for each output channel. For each channel, we define the pairwise cost between training sample i and calibration sample j as

$$c_{ij} = \alpha \left\| z_i^{\mathtt{tr}} - z_j^{\mathtt{cal}} \right\|_2^2 + \lambda_{\mathrm{cond}} \left\| \ell_i^{\mathtt{tr}} - \ell_j^{\mathtt{cal}} \right\|_2^2 .$$

The first term encourages residual similarity, while the second term preserves consistency with the low-resolution conditioning input. This prevents the transported residual from being matched to a calibration residual whose associated low-resolution field is incompatible with the source input. In all experiments, we set $\alpha = 1.0$ and $\lambda_{\mathrm{cond}} = 1.0$.

To make the transport computation scalable, we construct a sparse k-nearest-neighbor graph in the augmented PCA feature space and solve a balanced entropic Sinkhorn problem only over the resulting candidate pairs Cuturi (2013). Specifically, we use $k = 32$, entropic regularization 0.1, and 1500 Sinkhorn iterations. The sparse transport plan π is computed with uniform source and target marginals. When the number of training and calibration samples differs, the target marginal simply assigns a larger mass to each calibration sample.

Because a sparse k-nearest-neighbor graph can leave some calibration samples with too few incoming source candidates, we augment the graph before Sinkhorn normalization. In particular, we enforce a minimum target in-degree proportional to the source-to-target sample ratio by adding nearest reverse-neighbor edges for under-covered calibration samples. This improves feasibility of the balanced transport constraints while keeping the sparse graph size fixed.

After computing the sparse plan, we construct transported residual coefficients using a weighted top-m selection rather than averaging over all k neighbors. For each training sample, we select the top $m = 3$ calibration candidates according to the transport weights, with a capacity constraint limiting how often each calibration sample can be reused. This reduces collapse onto a small subset

of calibration residuals while avoiding overly diffuse barycentric averages. The selected weights are renormalized within the top-m set.

For each training sample i, the transported residual representation is therefore computed as

$$\tilde{z}_i^{\mathtt{tr}} = \sum_{j \in \mathcal{T}_i} \tilde{\pi}_{ij} z_j^{\mathtt{cal}},$$

where $\mathcal{T}_i$ denotes the selected top-m calibration candidates and $\tilde{\pi}_{ij}$ denotes the corresponding normalized transport weight. We then reconstruct the transported residual in pixel space as

$$\tilde{r}_i = U_K \tilde{z}_i^{\mathtt{tr}} + \bar{r}.$$

Finally, we define the corresponding pseudo-mean by enforcing exact consistency with the original high-resolution target:

$$\tilde{\mu}_i := y_i - \tilde{r}_i.$$

This yields adapted pseudo training samples

$$(x_i, \tilde{\mu}_i, \tilde{r}_i)$$

that preserve the mean–residual decomposition

$$y_i = \tilde{\mu}_i + \tilde{r}_i$$

by construction. Thus, ReMatch shifts the residual target distribution toward the calibration residual regime while maintaining sample-wise consistency among the input, pseudo-mean, residual, and ground-truth target.

A.7 Synthetic Bias-Controlled Benchmark Details

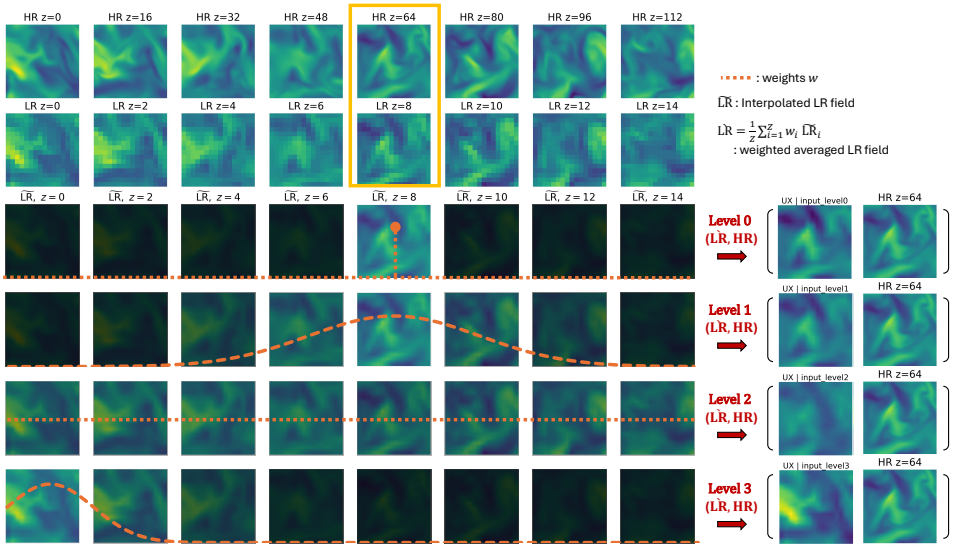


Figure 4: Synthetic LR bias levels used in the BlastNet benchmark. Level 0 uses the original interpolated LR input. Levels 1–3 introduce progressively stronger structured mismatch by mixing slices along the z-axis with Gaussian weights.

We use the BLASTNet Momentum128 3D super-resolution dataset Chung et al. (2023) as a controlled synthetic benchmark. The dataset provides clean paired low- and high-resolution turbulent flow fields, where the low-resolution inputs are generated from the corresponding high-resolution DNS fields through filtering and downsampling. Each high-resolution sample is a three-dimensional field of size $128 \times 128 \times 128$, with corresponding low-resolution fields provided at multiple downsampling scales. Both the input and target fields contain four physical channels: density `RHO` and the three velocity components `UX`, `UY`, and `UZ`. We use clusters 4, 11, 14, and 17, and take the $8\times$ downsampled inputs (spatial resolution 16×16). Since our implementation operates on 2D inputs, we extract slices along the z-axis. This also enables controlled bias injection along the depth direction.

We define four bias levels. Level 0 uses the original interpolated low-resolution field (clean interpolation). For levels 1–3, we replace each low-resolution slice with a weighted mixture of slices along the z-axis using Gaussian weights

$$w(z) = \exp\left(-\frac{1}{2}\left(\frac{z-c}{\sigma}\right)^2\right),$$

where the center c and bandwidth σ are chosen to produce progressively stronger structured mismatch: local mixture ($\sigma = 2$) for level 1, broad averaging for level 2, and sharp distant centering ($\sigma = 1$) for level 3.

Figure 4 illustrates the four bias levels and the corresponding weighting patterns. The baseline method Mardani et al. (2025) is trained using the entire training set of 6064 samples. For ReMatch, we partition this training set into a training subset ($n = 4048$) and a calibration set ($n = 2016$). The test set contains 2032 samples.

A.7.1 Controlled Synthetic Study of Bias-induced Residual Misspecification

Figure 5 shows channel-wise RMSE and SSR across four bias levels. Across all channels, ReMatch yields increasingly larger SSR improvements as the bias level increases, with only minor RMSE degradation or, in some cases, RMSE improvement. Figure 6 provides a qualitative comparison between the baseline method and ReMatch across the four input bias levels. Qualitatively, ReMatch better preserves high-frequency features than CorrDiff, particularly as the bias level increases.

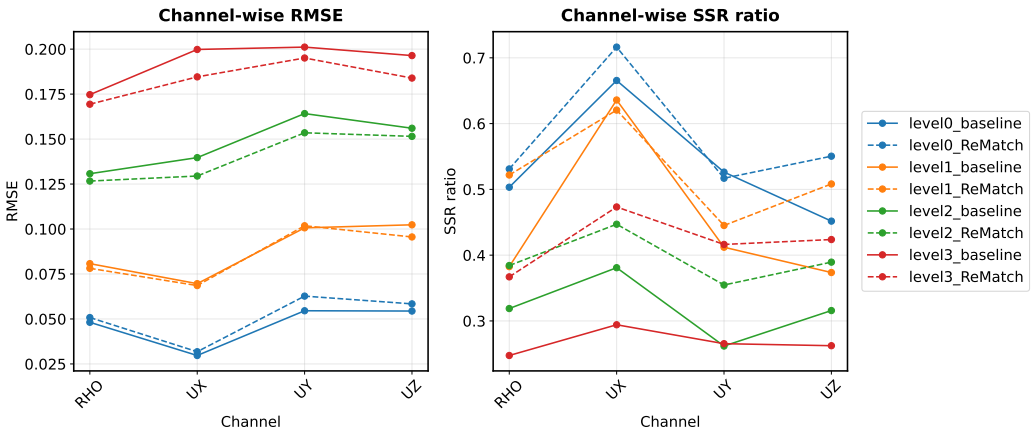


Figure 5: Channel-wise RMSE and SSR ratio over different bias levels.

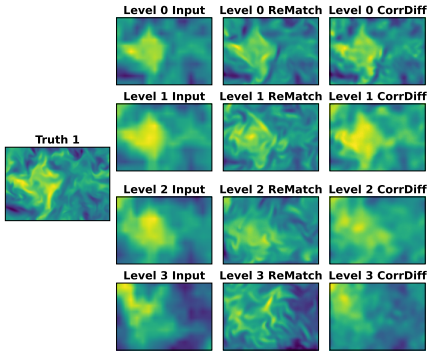


Figure 6: **Qualitative comparison across LR input bias levels.** Rows correspond to LR bias levels. Columns show the GT HR field, LR input, ReMatch and CorrDiff prediction.

A.8 Calibration set sensitivity experiment

ReMatch relies on a held-out calibration set to estimate the target residual distribution. In our main experiments, we use a 2:1 train-to-calibration split. To examine the sensitivity of ReMatch to the calibration set size, we vary the train-to-calibration ratio while keeping the overall available training data fixed. Specifically, we consider ratios of 1:1, 2:1, 5:1, 10:1, and 100:1. CorrDiff corresponds to the special case with no calibration set and therefore no residual distribution matching.

We conduct this experiment on the BLASTNet dataset with bias level 2. For each train-to-calibration ratio, ReMatch is trained using the corresponding split and evaluated over 20 test subsets, each containing 100 samples. Table 3 reports the RMSE and SSR averaged across four channels. Overall, increasing the calibration set size tends to improve SSR while maintaining comparable or improved RMSE. Figure 7 further shows the RMSE and SSR for the UX channel, with error bars indicating the standard deviation across the 20 test subsets.

A notable observation is the substantial drop in SSR for the 100:1 split. ReMatch transports training residuals toward the residual distribution estimated from the calibration set. Although the transport does not rely on one-to-one matching and instead uses a weighted combination of the top-k transport components, an extremely small calibration set provides only a sparse representation of the target residual distribution. Consequently, the transported residuals can become concentrated around a limited set of calibration residuals, resulting in reduced predictive spread and lower SSR. These results suggest that the calibration set should be sufficiently large to represent the target residual distribution. However, allocating more samples to calibration leaves fewer samples for regression training and increases the computational cost of optimal transport. Therefore, the calibration split

	CorrDiff	100:1	10:1	5:1	2:1	1:1
RMSE↓	0.1757	0.17370	0.17491	0.170417	0.17303	0.17190
SSR↑	0.722	0.530632	0.73590	0.86210	0.84919	0.9038

Table 3: Sensitivity of ReMatch to the train-to-calibration split on the BLASTNet dataset on bias level 2. RMSE and SSR are averaged across four output channels over 20 test subsets of 100 samples each. CorrDiff corresponds to the case without residual distribution matching.

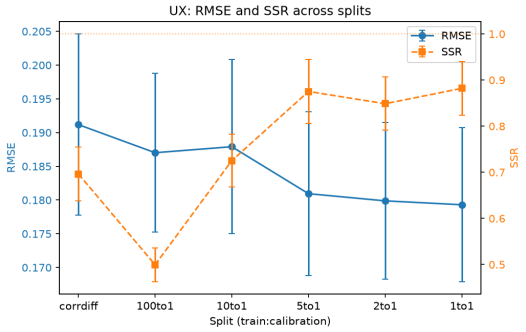


Figure 7: Sensitivity of ReMatch to the train-to-calibration split for the UX channel of the BLASTNet dataset. The left and right y-axes show RMSE and SSR, respectively. Each point reports the mean over 20 test subsets of 100 samples, with error bars indicating one standard deviation.

introduces a practical trade-off between regression accuracy, residual calibration, and computational cost.

A.9 ERA5-HRRR Dataset Details

We use the u- and v-components of wind at five lower-stratospheric pressure levels: 50, 75, 100, 125, and 150 hPa. The experiments are conducted over a $500\,\mathrm{km} \times 500\,\mathrm{km}$ region in the northeastern CONUS, defined by $\mathrm{lat} \in [36.0795, 41.8914]^\circ$ and $\mathrm{lon} \in [-82.5501, -75.2087]^\circ$.

ERA5 Hersbach et al. (2020) We downloaded $25 \times 25\,\mathrm{km}$ scale ERA5 reanalysis data for every hour from 2018-2021. We note that ERA5 only has 70 hPa instead of 75 hPa pressure level data, but we consider them as the same channel since it does not detract from our setting of input/output mismatch. We use the direct low-resolution input to train the SwinIR models, but to create the low-resolution input for all our other experiments, we follow the default PhysicsNeMo CorrDiff pipeline and use bilinear interpolation to upsample data to match the $3 \times 3\,\mathrm{km}$ resolution of HRRR data. We used 2018-2020 as our train set and 2021 as our test set. For methods that need a calibration set (UC, ReMatch), we used 2018-2019 as the train set, 2020 as the calibration set, and 2021 as the test set.

HRRR Dowell et al. (2022) We used $3 \times 3\,\mathrm{km}$ scale HRRR pressure data for every hour from 2018-2021 using lead times and forecasts of up to 6 hours to fill gaps.

To create the combined dataset, we also added four sinusoidal-embedding temporal input channels for the day and month. Fig. 8 indicates the main challenge of the real-world downscaling task with atmospheric data. There is an inherent mismatch between the LR and HR data over the same region since they come from two different sources. We match the spatial dimensions of the two datasets but do not explicitly reproject ERA5 onto the native HRRR grid. Consequently, the ERA5–HRRR residual can contain a systematic contribution associated with differences in spatial sampling and grid geometry, differences in model physics and data assimilation from cross-sensor sources, systematic pressure-level misalignment for the 75 hPa channel, and stochastic finer scale atmospheric variability.

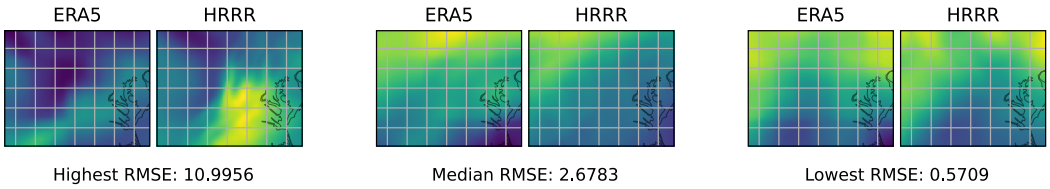


Figure 8: **Downscaling bias in ERA5 and HRRR has a positive relationship with RMSE.** We show three samples of upsampled ERA5 data and HRRR 100hPa u-wind from our 2018-2020 training set. The left sample with high RMSE shows a clear mismatch in the center while the right sample with low RMSE generally looks the same, indicating that RMSE is a good estimator of input/output mismatch.

Table 4: **Channel-wise statistics of downscaling bias.**

Channel	50u	50v	75u	75v	100u	100v	125u	125v	150u	150v
Mean	1.99	2.59	2.77	3.52	2.68	4.27	3.08	5.19	3.62	6.19
Std. Dev.	0.81	1.39	1.18	1.93	1.08	2.21	1.27	2.56	1.55	2.30

Table 4 shows higher bias levels for higher pressure levels, especially in the v-wind channel. We hypothesize this is due to higher pressure levels being closer to the jet stream (250 hPa) which has strong wind magnitudes that could cause more sensor inaccuracies.

A.10 Implementation Details

ReMatch is not tied to a specific architecture. It only assumes a deterministic mean predictor followed by a stochastic residual generator, so the mean predictor can be instantiated with any suitable pointwise estimator and the residual stage with any stochastic or ensemble-based generative model. With this decomposition, we can model predictable biases from the ERA5 input with the deterministic regression component, including systematic or state-dependent differences between the two datasets. We can model variability that is not determined by the available low-resolution input with the stochastic residual diffusion component, allowing the model to represent multiple plausible HRRR realizations for a given ERA5 input. In our experiments, we use U-Net regression and SwinIR Liang et al. (2021) as deterministic mean predictors, trained with ℓ_2 and ℓ_1 losses, respectively. For stochastic residual prediction, we use EDM-style diffusion preconditioning and sampling Karras et al. (2022) with a U-Net-based score network Song et al. (2020). The low-resolution input x and mean field μ, either transported pseudo-mean from training set $\tilde{\mu}_i$ or mean predictor output in calibration set $\mu_\phi(x_j)$, are concatenated channel-wise as the conditioning input. Other stochastic residual generators can be used in the same framework.

Unless otherwise noted, all UNet-based regression and diffusion models are trained under the same implementation settings to ensure a fair comparison. We use data-parallel training on four NVIDIA A100 GPUs with a per-GPU batch size of 32 for experiments, in which training takes about 8-10 hours. All regression UNets and diffusion models are trained for 8M processed samples, following the training-duration convention used in the CorrDiff implementation at `https://github.com/NVIDIA/physicsnemo/blob/main/examples/weather/corrdiff/README.md`.

For the CorrDiff diffusion backbone, we instantiate `EDMPrecondSuperResolution` with a `SongUNetPosEmbd` denoising network. The `mini` model-size preset controls the capacity of the underlying U-Net through three parameters: `model_channels`, `channel_mult`, and `attn_resolutions`. Specifically, `model_channels`=64 sets the base channel width; `channel_mult`=$[1, 2, 2]$ defines a three-level encoder–decoder hierarchy with feature widths $64, 128, 128$; and `attn_resolutions`=$[16]$ inserts self-attention blocks at the 16×16 feature-map resolution. All other architectural choices follow the default PhysicsNeMo CorrDiff settings.

CorrDiff Mardani et al. (2025) A two-stage mean–residual baseline consisting of a regression mean predictor followed by a diffusion model for stochastic residual prediction.

UNet Song et al. (2020) A deterministic regression baseline to map the LR input directly to the HR output.

Conditional Diffusion Model (CDM) A diffusion model conditioned on the LR input without an upstream mean predictor.

CFG Ho & Salimans (2022) Classifier-free guidance applied only to the residual diffusion model in CorrDiff to sharpen residual samples and mitigate small-scale underdispersion. We implement classifier-free guidance (CFG) on top of the CorrDiff architecture as a baseline. CFG allows conditional sampling without training an additional classifier. During training for diffusion, we zero out the conditioning on the LR input 10% of the time so the model learns both conditional and unconditional inputs. During inference, we use 18 denoising steps in the stochastic sampler with a cosine guidance weighting schedule from $w = 0$ to 8. Comparing this to fixed weights from 0 to 8, we found that scheduling increasing guidance helped maintain RMSE skill while improving spread. The sample at each time step is then a linear combination of the conditional and unconditional output. This encourages the model to learn features induced specifically by the conditioning.

Finetuning Couairon et al. (2026). We construct a fine-tuning baseline following the held-out residual adaptation strategy of ArchesWeatherGen Couairon et al. (2026). We first train the deterministic regression model and residual diffusion model on the 2018–2019 training set for 8M processed samples each. We then freeze the regression model and fine-tune only the residual diffusion model for an additional 1M processed samples on the 2020 calibration set. Unlike the original method, we do not use inference-time noise rescaling and retain the same diffusion architecture as the other baselines.

UC Kendall & Gal (2017) A CorrDiff variant that conditions the residual diffusion model on an uncertainty estimate of the regression mean, predicted by a lightweight UNet. In addition to UC_μ that was discussed in Section 5, we also tested using error quantiles as an input, which we call UC_q. Specifically, UC_μ conditions on the predicted MAE, and UC_q conditions on the predicted lower and upper error quantiles, q_5 and q_{95}. To construct these baselines, we first train a deterministic mean predictor on the 2018–2019 training set. We then train a separate uncertainty estimator on the 2020 calibration set to predict either the MAE or the q_5/q_{95} error quantiles of the mean predictor. The uncertainty estimator uses the same UNet as the regression mean and is trained for 8M samples. Using the calibration set as a proxy for the test regime allows the uncertainty network to learn held-out error statistics, rather than relying on training-time errors that may be overconfident.

- UC_μ: The network predicts a scalar MAE value of the residuals between the ground truth and the mean predictor. Uses L2 loss.
- UC_q: The network predicts scalar quantile values for Q5 and Q95. Uses Pinball loss.

After training the mean predictor and uncertainty estimator sequentially, we train the residual diffusion model conditioned on the LR input, the predicted mean, and an uncertainty measure. During diffusion training, we use the ground-truth uncertainty measure computed from the mean-prediction error, while at test time we replace it with the output of the uncertainty estimator. This choice stabilizes diffusion training by providing accurate uncertainty conditioning during optimization, while preserving a fully predictive pipeline at inference. The residual diffusion model is trained on the combined 2018–2020 data, including both the training and calibration splits.

Additionally, Figure 5 displays the results of UC_q for the same 600 test samples as Table 1. While it has a slightly better RMSE, LPIPS, and ACC compared to the UC_μ variant, it performs worse on calibration metrics, which are essential for good probabilistic models.

SwinIR Liang et al. (2021) A deterministic transformer-based super-resolution baseline, which directly maps the LR input to the HR field. We implement a SwinIR-based super-resolution baseline following the network configuration used in the original SwinIR model at

Table 5: **UC_q quantitative results on HRRR-ERA5 pressure-level super-resolution over all channels.**

Metric	UQ_q	UQ_μ	$Rematch_U$
RMSE ↓	1.8551	1.8584	1.8102
SSR → 1	0.5950	0.8745	0.6347
CRPS/MAE ↓	1.1946	1.1545	1.1567
LPIPS ↓	0.0842	0.0871	0.0841
ACC → 1	0.8952	0.8926	0.9017

https://github.com/JingyunLiang/SwinIR. For the network architecture, we use the main SwinIR configuration from the original paper: the embedding dimension is set to 180, the model contains six Residual Swin Transformer Blocks (RSTBs), and each RSTB contains six Swin Transformer layers. The number of attention heads is set to 6 in each RSTB, the MLP ratio is set to 2, and the residual connection inside each RSTB uses a single 3×3 convolution. We also use the standard multi-step PixelShuffle reconstruction head with an intermediate feature dimension of 64. The only modification from the original SwinIR parameter setting is the window size. While the original model uses a window size of 8 for super-resolution, we set the window size to 7 so that the self-attention windows exactly divide the low-resolution input size of 21×21. The model is trained using a ℓ_1 loss between the predicted and ground-truth high-resolution fields. We train the SwinIR baseline for 100k processed samples, at which point the validation loss is observed to have converged.

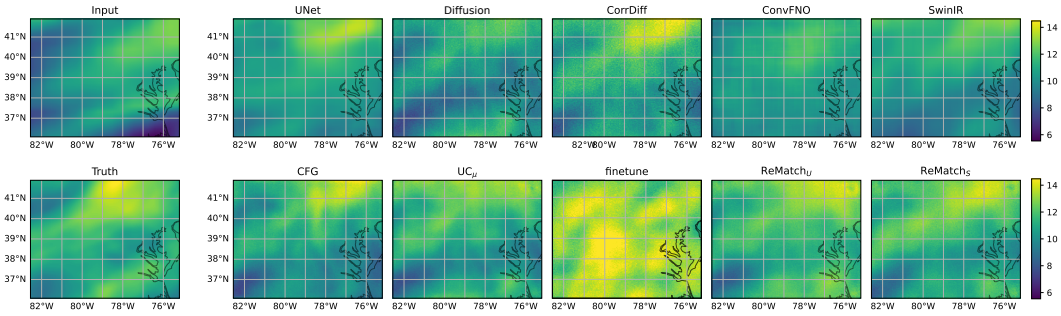
Conv-FNO Wen et al. (2022) A deterministic neural-operator baseline, which combines Fourier neural operator layers with convolutional blocks for direct LR-to-HR super-resolution. We implement an FNO-based super-resolution baseline using a 2D FNO model from NVIDIA's PhysicsNeMo at `https://github.com/NVIDIA/physicsnemo/blob/main/physicsnemo/models/fno/fno.py` and modifying the architecture and parameters according to the BlastNet Conv-FNO model at `https://github.com/blastnet/blastnet2_sr_benchmark/blob/ae2a3a6b1c1c8fb9dcf478aa47ee254159c42f1b/models/convfno.py#L12`, that modified a 3D FNO model to be used for super-resolution.

Our main modifications were replacing FNO's decoder with an Upsampler block and adding a ResidualBlock to the spectral layers in FNO2DEncoder. We use 34 features, 34 latent channels, 32 spectral and residual layers, kernel size of 3, 2 modes, a padding of 6, and a 8x upscaling factor. These parameters were chosen to match the implementation in BlastNet to remain similar to an established super-resolution baseline. The model was trained using MSE loss between the predicted and ground-truth high-resolution fields. We train the Conv-FNO baseline for 8 million processed samples, similar to the CorrDiff regression capacity.

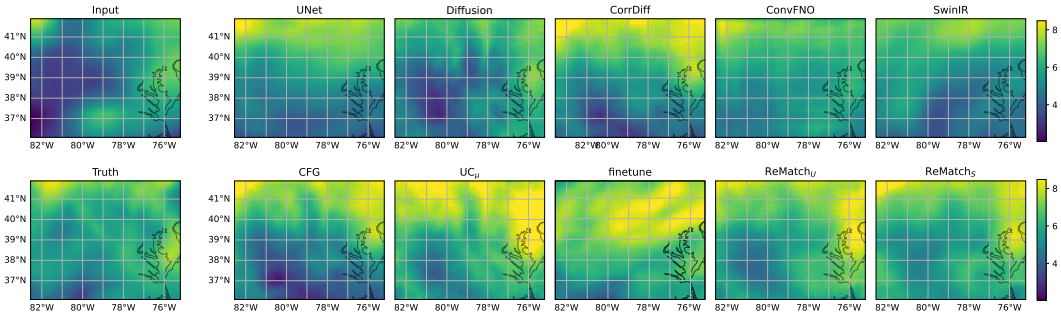
A.11 Additional Analysis on Real-World Wind Field Downscaling Task

A.11.1 Method Comparisons

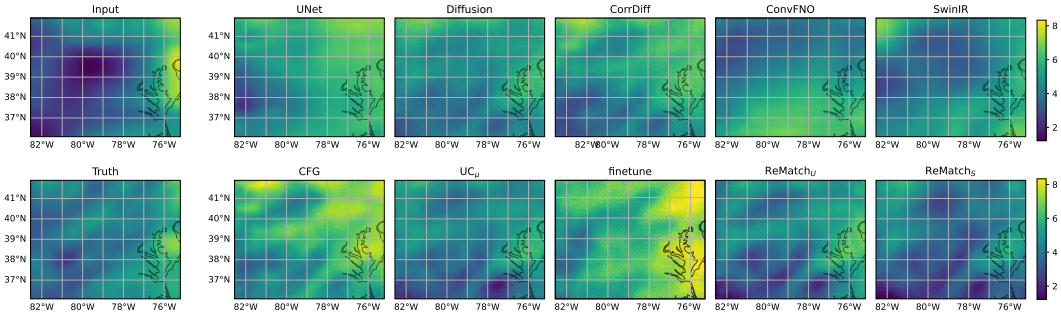
We provide additional qualitative examples of downscaling results for all baseline models and the proposed method.



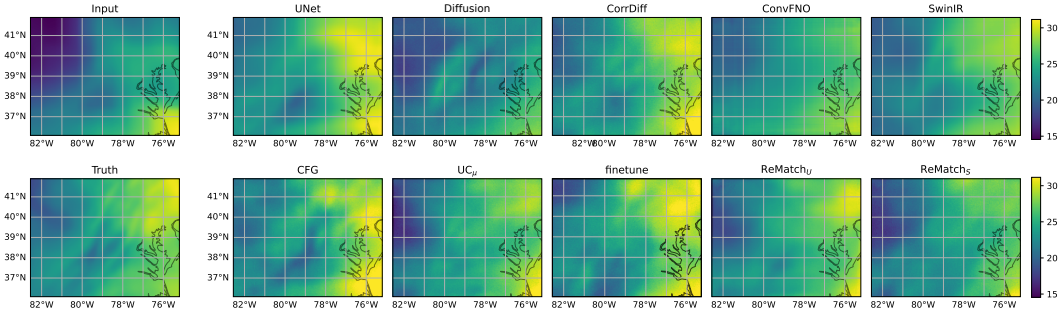
(a) Qualitative comparison of downscaled 125 hPa u-wind fields on 2021-07-20 00:00:00.



(b) Qualitative comparison of downscaled 75 hPa v-wind fields on 2021-09-14 01:00:00.



(c) Qualitative comparison of downscaled 125 hPa u-wind fields on 2021-07-06 19:00:00.

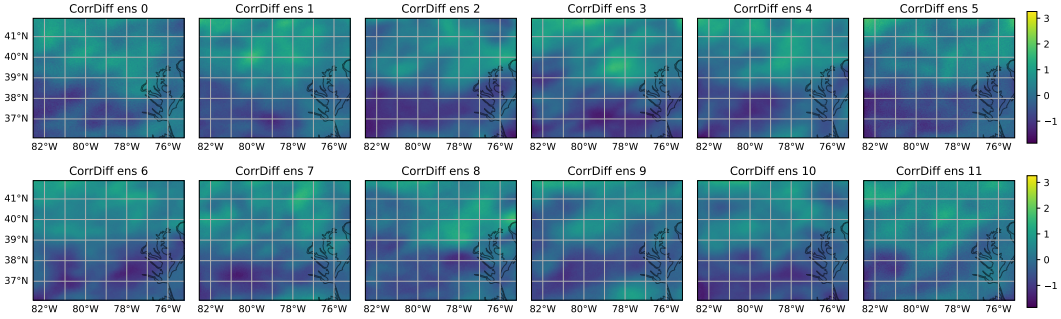


(d) Qualitative comparison of downscaled 150 hPa u-wind fields on 2021-10-19 7:00:00.

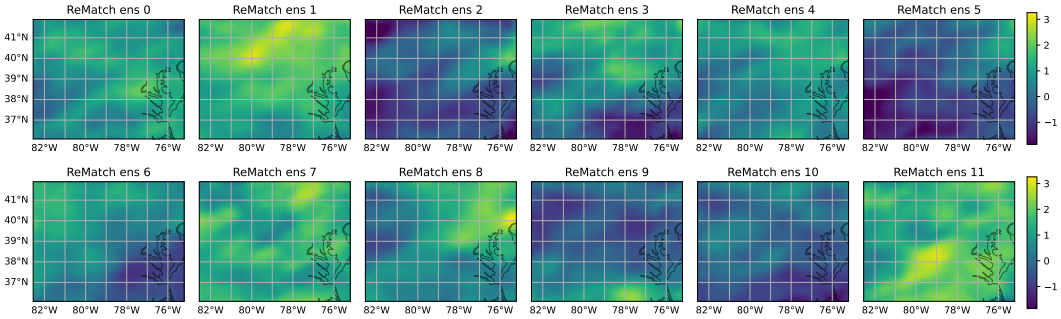
Figure 9: Qualitative comparison of all baseline and proposed methods across randomly selected channels and dates.

A.12 Ensemble Samples Comparison

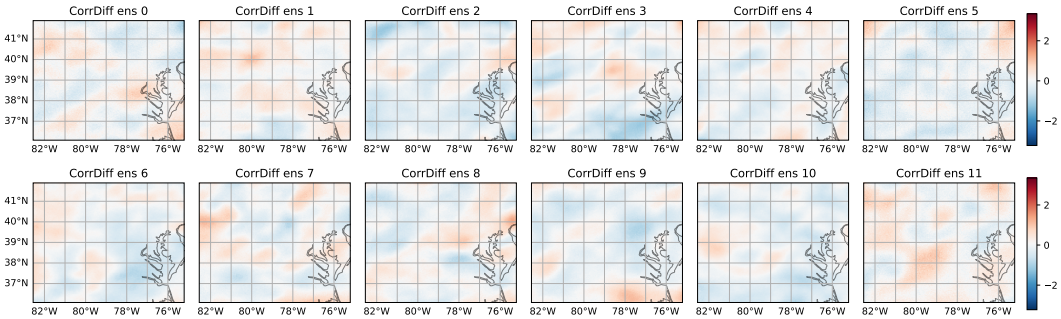
We further analyze individual ensemble samples. Figure 2 shows downscaling results for the 50 hPa v-wind field on 2021-09-05 at 10:00:00. In Figure 10, we show all 12 ensemble samples of the final predictions, $\mu_\phi(x) + r_\theta(x, \mu_\phi)$, together with their predicted residuals $r_\theta(x, \mu_\phi)$.



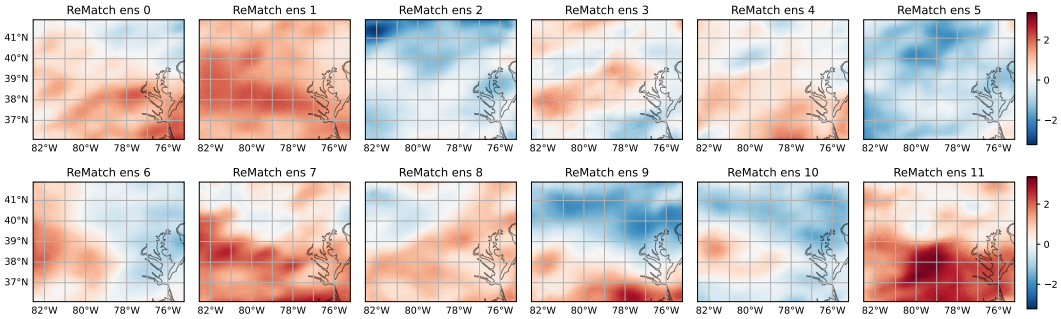
(a) CorrDiff all prediction ensemble samples.



(b) ReMatch$_{\text{U}}$ all prediction ensemble samples.



(c) CorrDiff all residual ensemble samples.



(d) ReMatch$_{\text{U}}$ all residual ensemble samples.

Figure 10: Ensemble predictions from Fig. 2.

Consistent with prior observations Mardani et al. (2025); Fotiadis et al. (2025) and our quantitative results, CorrDiff produces underdispersed predictions with a low SSR. This behavior is visible in Figure 10: CorrDiff ensemble samples remain highly similar across members, as shown in Figure 10a, and their residual magnitudes are relatively small, as shown in Figure 10c, compared with ReMatch residuals in Figure 10d. This suggests that CorrDiff fails to generate sufficiently corrective residuals under residual misspecification. Given the error-prone mean prediction shown by the UNet result in Figure 2, the residual model should correct large-scale mean errors rather than merely add high-frequency details. However, CorrDiff largely fails to do so, whereas ReMatch produces more diverse and corrective residuals.

A.13 ACC Score analysis

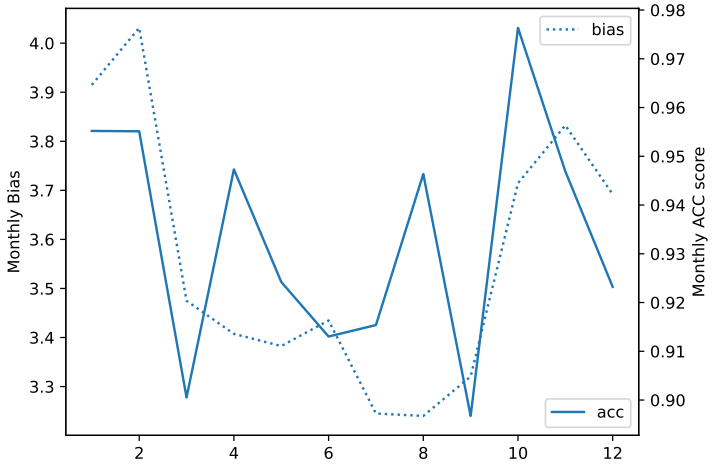


Figure 11: **Comparison of ReMatch$_{\text{U}}$ ACC and input/output bias per month.** ACC score per month calculated for 2021 and average bias (RMSE) per month calculated over 2018-2021.

With the ACC Score calculated across 2021, we can compare how spatially similar ReMatch$_{\text{U}}$ anomalies and HRRR anomalies are with the estimated climatological mean from 2018-2020 HRRR data. Anomalies are when the atmospheric variable deviates from the expected historical value at a certain time and location. We note that having a longer range of years to estimate the climatological mean would likely lead to more accurate estimates of anomalies. Fig. 11 shows a monthly breakdown of downscaling bias and captures how our model adapts to the bias gap between ERA5 and HRRR. In winter months the bias is higher, but our model maintains high ACC scores, so ReMatch$_{\text{U}}$ is able to capture anomaly structure.

A.14 Kinetic Energy Analysis

Figure 12 compares the 1D kinetic energy spectra of the ground-truth HRRR field, the ERA5 input, CorrDiff predictions, and ReMatch predictions. This diagnostic measures how kinetic energy is distributed across spatial scales: low wavenumbers correspond to large-scale structure, whereas high wavenumbers correspond to finer-scale variability. A successful downscaling model should therefore not only reduce pointwise error, but also recover the target spectrum over the relevant range of wavenumbers.

To compute this quantity, we form the horizontal kinetic energy from the wind components and evaluate its one-dimensional Fourier spectrum. For each sample, we compute the Fourier coefficients of the horizontal wind components u and v, and form

$$E(k) \propto \frac{1}{2}\Big(|\hat{u}(k)|^2 + |\hat{v}(k)|^2\Big).$$

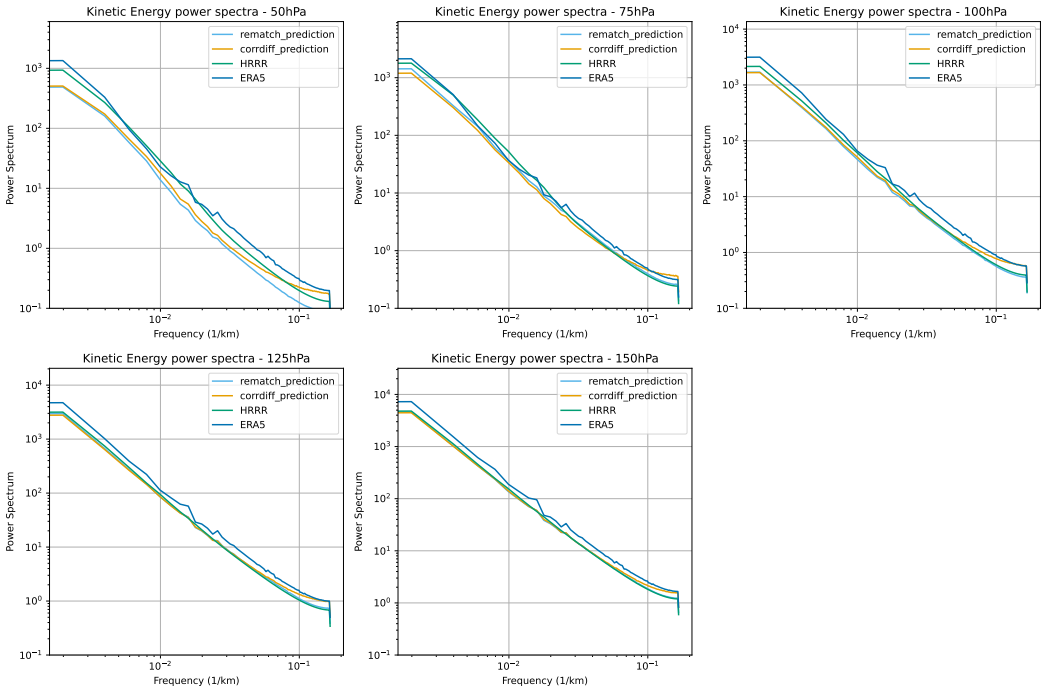


Figure 12: **Power spectra for kinetic energy in all pressure levels**

We average over the remaining spatial dimension and over samples. For stochastic methods, the spectrum is computed separately for each ensemble member and then averaged across ensemble members.

The figure shows that ERA5 overestimates kinetic energy at all wavenumbers relative to HRRR, indicating a scale-dependent bias. ReMatch removes a substantially larger fraction of this excess power and tracks the HRRR spectrum more closely over a broad range of wavenumbers. The gap between CorrDiff and ReMatch is most visible at higher wavenumbers, where CorrDiff preserves more of the excessive energy inherited from ERA5, whereas ReMatch remains closer to the HRRR target.

This pattern differs from surface-level wind fields shown in Mardani et al. (2025), where HRRR often contains much stronger high-frequency variability due to land–sea contrast, topography, and other surface effects. At pressure levels, the flow is smoother, so even the HRRR target has relatively weak high-frequency energy. In this regime, the main issue is therefore not only missing fine-scale structure, but a broader mismatch in scale-dependent energy.

A.15 Uncertainty Conditioned Experiments

Previous work on image regression Angelopoulos et al. (2022); Kendall & Gal (2017) often learns pixelwise epistemic or aleatoric uncertainty jointly with the regression predictor. In our setting, we adapt this idea to the mean–residual CorrDiff architecture by training an auxiliary model to predict uncertainty about the upstream mean and using that quantity as an additional conditioning signal for residual diffusion. Empirically, however, dense pixelwise uncertainty maps did not learn meaningful or predictive structure, whereas low-dimensional scalar summaries of the regression error—such as MAE, Q_5, Q_{90}, and Q_{95}—were much easier to predict and matched the ground-truth values more reliably on the test set (Fig. 13). We explored two other spatially focused methods called Moran's I (a local indicator of spatial autocorrelation) and VGGT Wang et al. (2025) inspired confidence maps. While the trained networks provided more meaningful feature maps 14, they ultimately did not perform as well as conditioning to diffusion.

We interpret this through the lens of residual target misspecification. Under downscaling bias, the regression error $Y - \mu_\phi(X)$ is itself a predictor-defined correction target whose distribution shifts across regimes. Learning a full pixelwise uncertainty field therefore amounts to inferring a high-dimensional summary of a misspecified residual target from limited observable information, which is substantially more difficult than predicting a coarse scalar descriptor of its magnitude. This explains why scalar uncertainty signals can still provide useful information to the diffusion model, as reflected in the UC baselines, while dense pixelwise uncertainty prediction remains unreliable. More broadly, these results suggest that uncertainty prediction alone is not sufficient to resolve residual misspecification: compressing the error into low-dimensional summaries may help, but recovering its full spatial structure requires a more direct treatment of the shifted residual target itself.

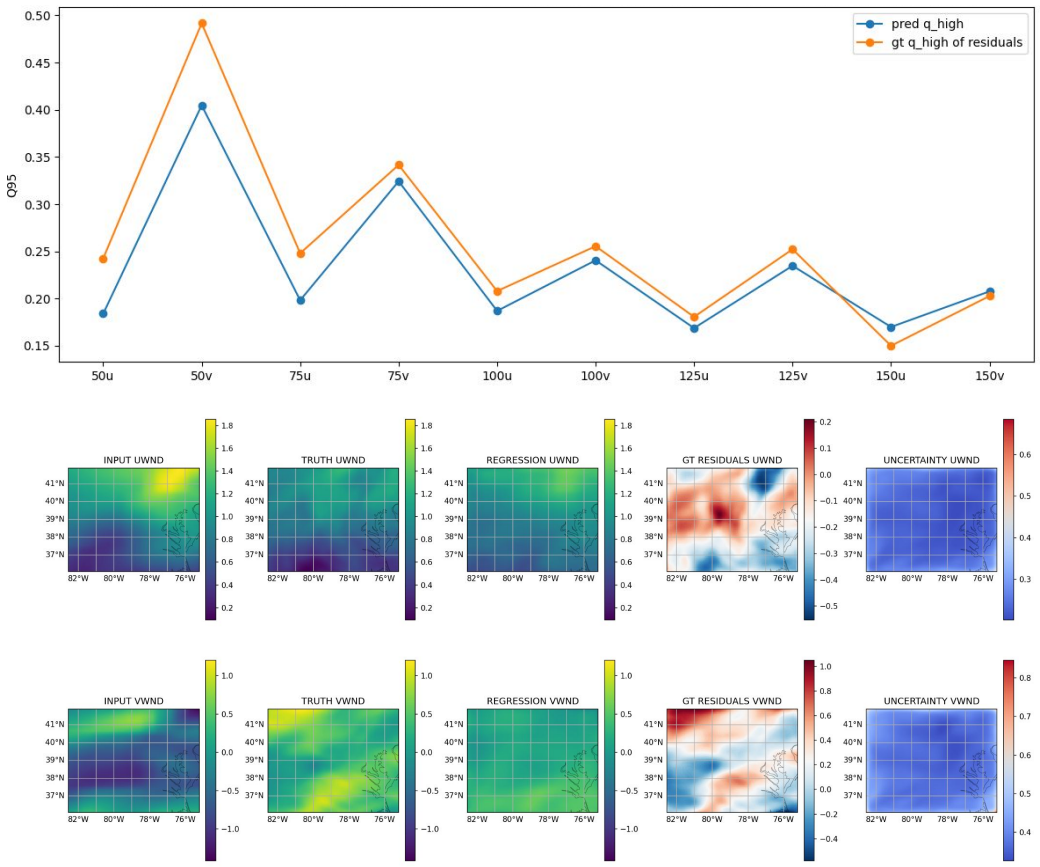


Figure 13: **Comparison of scalar and pixelwise uncertainty estimators.** Both networks predict q5 and q95 of the residuals per channel on the test set. The top plot is a scalar prediction while the bottom plot displays a pixelwise prediction at 2021-01-02 11:00:00 at 50 hPa u-wind field.

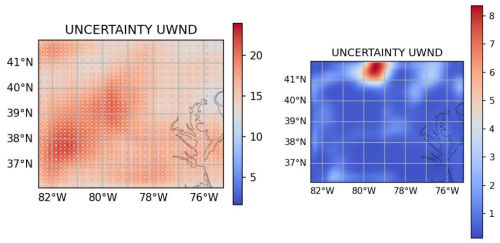


Figure 14: **Comparison of VGGT inspired (left) and Moran's I (right) pixelwise uncertainty networks.** Both samples are also from 2021-01-02 11:00:00 at 50 hPa u-wind field. The Moran's I uncertainty focuses more on "interest points" while the VGGT inspired confidence is more spatially aware. However, the VGGT inspired confidence has systematic artifacts.