DRIVEGUARD-WM: DYNAMIC-ONLY CROSS-REGION METEOROLOGICAL DOWNSCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised meteorological downscaling relies on paired low-resolution (LR) and high-resolution (HR) data, yet observationally supported HR data remain geographically limited. This motivates cross-region downscaling from paired source-region data to target regions with only LR inputs. However, regional differences in spatiotemporal dynamics limit the transferability of learned reconstruction relationships. Our central premise is that transfer should account for the predictive contributions of source-learned relationships and adapt their use to current LR conditions. To address this challenge, we propose DriveGuard-WM, a cross-region downscaling framework that reconstructs HR fields from dynamic LR inputs alone, without target HR supervision or fine-tuning. The framework translates contribution assessments from a world model into LR-conditioned guidance for adaptive spatiotemporal fusion. Across six directed transfers among Australia, the contiguous United States, and Europe, DriveGuard-WM reduces macro-averaged, source-standardized mean absolute error (MAE) by 38.6% for temperature and 46.0% for precipitation relative to the best evaluated baseline on this metric under the reported reference-field protocol. Our code is available at https://anonymous.4open.science/r/WeatherDownscaling-7732/.

1 INTRODUCTION

Global meteorological products provide seamless coverage, but their low-resolution (LR) grids cannot fully represent local variability. High-resolution (HR) data with adequate observational support remain geographically limited (Harder et al., 2026), although regions without such data also need fine-scale information for water management, agriculture, and climate risk assessment. Downscaling reconstructs HR meteorological fields from LR predictors, spanning classical statistical formulations, neural super-resolution, and stochastic generative modeling (Maraun et al., 2010; Vandal et al., 2017; Leinonen et al., 2021; Dai & Ushijima-Mwesigwa, 2025; Fotiadis et al., 2025; Hess et al., 2025; Sun et al., 2026). In target regions where HR reference fields are unavailable for training or adaptation, supervised downscalers cannot be fitted locally even though LR inputs remain available (Harder et al., 2026). We therefore study **Cross-Region Downscaling**: training on paired LR–HR data from a source region and reconstructing HR fields in an unseen target region using only its LR inputs. Differences in climatic backgrounds and meteorological processes make this transfer challenging.

Diffusion and flow-based models can reconstruct fine-scale atmospheric structure (Mardani et al., 2025; Fotiadis et al., 2025), yet geographic-transfer benchmarks report substantial degradation in unseen regions even when the training area is expanded (Harder et al., 2026). In our source-trained setting, this gap exposes three linked challenges. **First, it is unclear which source-learned cross-scale relationships actually improve HR reconstruction.** Coarse LR fields leave fine-scale outcomes underdetermined (Hess et al., 2025), and geography-dependent high-frequency detail is difficult to recover in complex terrain (Xu et al., 2026). Similar predictive scores can also mask different, sometimes physically implausible, learned predictor relationships (González-Abad et al., 2026). Thus, a large HR-induced change in a learned representation need not reduce reconstruction error (Section 5.3); transferring every such change may carry source-specific structure rather than useful information. **Second, relationships that help in the source may not remain useful in a new region.** Regional differences in climate, topography, and precipitation dynamics can alter

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.

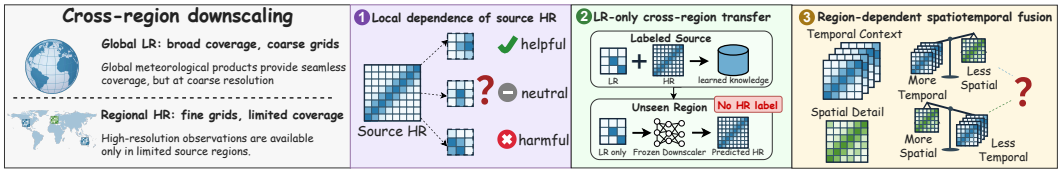


Figure 1: Cross-region downscaling and its three connected challenges: local dependence of HR detail, LR-conditioned use of learned reconstruction relationships, and regional variation in spatiotemporal contributions.

LR-to-HR relationships (Hernanz et al., 2024; Harder et al., 2026); one recent conditional diffusion downscaler required retraining when moved to a new geographic region (Aich et al., 2026). Without target HR, the model cannot directly assess or recalibrate these relationships and must decide which to use from target LR alone. **Third, the value of recent temporal evolution relative to current spatial structure is not constant across regions.** Recent downscalers use temporal conditioning to preserve evolving fine-scale fields (Srivastava et al., 2024; Schmidt et al., 2025), while learnable spatial context reduces region- and season-specific downscaling errors (Fuentes-Franco et al., 2025). Our held-out source comparisons show that different regions favor different temporal inputs (Section 5.5). A fixed fusion can thus overuse uncertain spatial detail or discard useful history, and its balance must be chosen without target HR feedback.

To address these challenges, we propose **DriveGuard-WM**, which reconstructs HR fields from dynamic LR inputs alone. **First**, a World Model for Contribution Assessment (WMCA) links global context, cross-scale messages, and local states. Controlled HR-informed message corrections quantify the reconstruction benefits of refining background–local relationships. **Second**, LR-Conditioned Guidance Learning (LGL) uses correction gains and temporal stability to supervise Drive/Guard scores predicted from LR inputs and prior states. Current content assignments map these scores to spatial guidance. **Third**, Adaptive Spatiotemporal Reconstruction (ASR) combines a guided spatial path with sequence predictions through a source-calibrated, sign-constrained fusion gate. Training and calibration use source data; all modules remain frozen in unseen regions.

Our main contributions are:

- **Novel LR-only cross-region framework:** DriveGuard-WM learns how source HR details depend on local conditions and reconstructs HR fields in unseen regions from dynamic LR inputs, without target HR supervision or fine-tuning.
- **Benefit-guided adaptive reconstruction:** We assess source relationships by reconstruction benefit, learn LR-conditioned guidance from correction gains and temporal stability, and adapt temporal and spatial fusion to regional conditions.
- **Consistent cross-region gains:** Across six directed transfers, DriveGuard-WM reduces macro-averaged, source-standardized MAE by **38.6%** for temperature and **46.0%** for precipitation relative to the strongest evaluated baseline under the reported reference-field protocol.

2 Related Work

Meteorological downscaling and geographic generalization. Downscaling spans statistical methods (Wilby & Wigley, 1997; Maraun et al., 2010), neural super-resolution, and generative models (Vandal et al., 2017; Leinonen et al., 2021; Ling et al., 2024; Hess et al., 2025). Studies of geographic transfer and zero-shot downscaling differ in their target supervision, conditioning, adaptation, and external priors (Harder et al., 2026; Zhu & Zhou, 2024; González-Abad et al., 2026; Tie et al., 2026). We isolate transfer from dynamic LR inputs after single-source training (Appendix D.1).

World models and latent dynamics. World models represent temporal dynamics in compact latent states (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2021; 2025; Micheli et al., 2023). We draw on the prior–posterior structure of recurrent state-space models (Hafner et al., 2019) to test whether HR-conditioned state replacements improve prediction. The temporal prior remains at deployment; the HR posterior is removed (Appendix D.2).

Privileged information and distillation. Generalized distillation transfers privileged training-time knowledge to a deployment-time student (Lopez-Paz et al., 2016). DriveGuard-WM follows this information-flow paradigm but differs in its supervision: controlled replacement benefits define cross-scale relation targets rather than directly matching teacher outputs or features (Appendix D.3).

Transfer learning and out-of-distribution generalization. Transfer learning and adaptation use related tasks or target samples (Pan & Yang, 2010; Yosinski et al., 2014; Ganin et al., 2016; Wang et al., 2021a), whereas domain generalization excludes target access during training (Arjovsky et al., 2019; Cha et al., 2021; Rame et al., 2022). We summarize relation benefits across temporal partitions of one labeled geographic source. Their stability is a source-side criterion; geographic transfer requires separate evaluation (Appendix D.4).

3 Preliminaries

Let s and u denote a source and an unseen target region. Each sample pairs a length-T dynamic LR sequence with the HR field at its final time:

$$X_t^s = (x_{t-T+1}^s, \ldots, x_t^s), \qquad \widehat{y}_t^s = f_\theta(X_t^s). \tag{1}$$

The LR input contains seven dynamic meteorological variables; temperature and precipitation are reconstructed separately. The current frame is included, so the task is contemporaneous spatial downscaling. Each training run uses one source region; rotating the three source regions gives six directed transfers. Training minimizes source reconstruction loss together with the auxiliary objectives in Section 4:

$$\theta_s^\star = \arg\min_\theta \frac{1}{|\mathcal{D}_s|} \sum_{(X,y)\in\mathcal{D}_s} \ell(f_\theta(X), y), \qquad \widehat{y}_{t;s\to u} = f_{\theta_s^\star}(X_t^u). \tag{2}$$

All parameters, running statistics, and calibration choices are fixed before target evaluation. Target HR is used only for offline scoring. The protocol excludes static covariates, coordinates, region identity, target-derived statistics, and test-time adaptation; all deployed quantities depend on dynamic LR inputs.

4 Method: DriveGuard-WM

4.1 Overall Framework

Given a dynamic LR sequence $X_t = (x_{t-T+1}, \ldots, x_t)$, DriveGuard-WM transfers source-region reconstruction knowledge through three modules (Fig. 2). **World Model for Contribution Assessment (WMCA)** models global context, cross-scale messages, and local evolution to produce a sequence prediction $\widehat{y}_A$ and diagnose relationship contributions. **LR-Conditioned Guidance Learning (LGL)** learns Drive/Guard guidance from source HR corrections and predicts it from LR content at deployment. **Adaptive Spatiotemporal Reconstruction (ASR)** uses this guidance to produce a spatial prediction $\widehat{y}_B$ and fuse the two paths. Temperature and precipitation are modeled separately; target inference uses dynamic LR inputs only.

4.2 World Model for Contribution Assessment

The first challenge is identifying which source-learned cross-scale relationships actually improve HR reconstruction. We address it with WMCA, an LR-conditioned model inspired by recurrent state-space models (Hafner et al., 2019). It factorizes regional context, background–local interaction, and local evolution into G, M, and L, producing an LR-only sequence prediction and exposing cross-scale messages for the source-HR contribution assessment in Section 4.3. Here, W denotes linear projections, σ sigmoid, and brackets concatenation.

Multiscale observation encoding. A two-stage convolutional encoder extracts multiscale features. Global cross-attention summarizes regional context, while learned local queries aggregate content-dependent features:

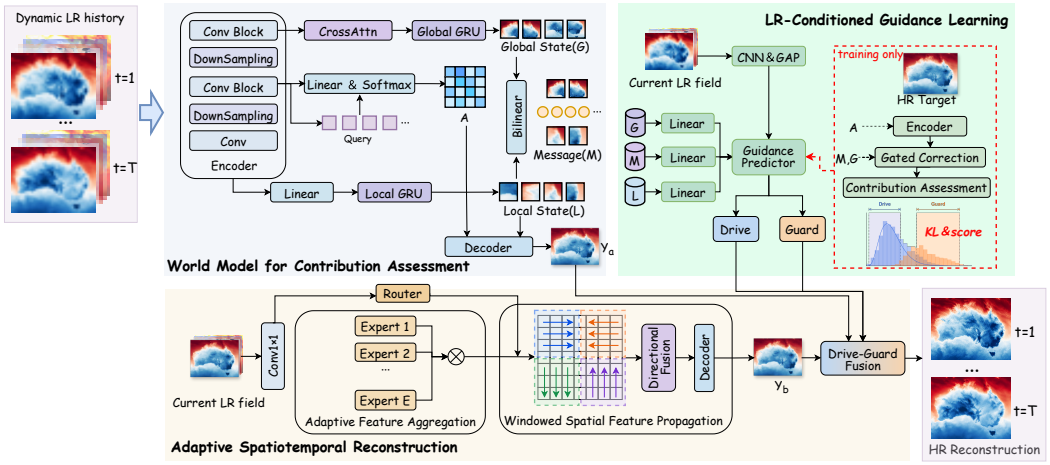


Figure 2: DriveGuard-WM architecture. WMCA maps dynamic LR history to global, message, and local states (G, M, and L) and produces the sequence prediction $\widehat{y}_A$. A source-only HR branch supervises LGL's LR-conditioned Drive/Guard scores. ASR performs guided spatial reconstruction to obtain $\widehat{y}_B$, followed by Drive/Guard fusion.

$$\begin{aligned} (F_t^G, F_t^L) &= \mathrm{CNN}_{\mathrm{enc}}(x_t), & o_t^G &= \mathrm{CrossAttn}(Q^G, F_t^G), \\ A_t &= \mathrm{Softmax}_n\big(W_Q(Q^L)W_F(F_t^L)^\top\big), & o_t^L &= W_O\big(\mathrm{Pool}_{A_t}(F_t^L)\big). \end{aligned} \tag{3}$$

CrossAttn produces K global observations. The assignment $A_t \in \mathbb{R}^{N\times P}$ maps P grid positions to N local queries with $\sum_n A_t(n,u) = 1$; Pool_A normalizes weights within each query.

Global states G: regional temporal context. Global cross-attention yields K observations o_t^G; a global GRU integrates them with recurrent history, and a distributional head produces G_t. The result carries evolving regional context into message construction.

Cross-scale messages M: background–local interaction. A bilinear interaction couples each $G_t(k)$ with local observation $o_t^L(n)$ and query q_n, producing a message $M_t(k,n)$ for every global–local pair. Keeping pairs explicit enables relation-wise correction and contribution assessment.

Local states L: content-specific temporal evolution. A Local GRU combines o_t^L, preceding local states, and $\bar{M}_t = K^{-1}\sum_k M_t(k,\cdot)$. The assignment $A_t^\top$ returns local priors to the grid, and a convolutional decoder with PixelShuffle upsampling produces $\widehat{y}_{A,t}$:

$$\begin{aligned} h_t^G &= \mathrm{GRU}_G([G_{t-1}, o_t^G], h_{t-1}^G), & G_t &\sim p_G(\cdot \mid h_t^G), \\ \eta_{kn,t} &= \mathrm{Bilinear}\big(W_g(G_t(k)), W_q([o_t^L(n), q_n])\big), & M_t(k,n) &\sim p_M(\cdot \mid \eta_{kn,t}), \\ h_t^L &= \mathrm{GRU}_L([L_{t-1}, o_t^L, \bar{M}_t], h_{t-1}^L), & L_t &\sim p_L(\cdot \mid h_t^L), \\ \widehat{y}_{A,t} &= \mathrm{Dec}_A(A_t^\top L_t^p). \end{aligned} \tag{4}$$

Linear heads parameterize diagonal Gaussian priors; p denotes the LR-conditioned prior and q the HR-informed posterior introduced below. Shared updates produce input-dependent G^p, M^p, L^p, A and the sequence prediction $\widehat{y}_A$.

4.3 LR-Conditioned Guidance Learning

The second challenge is selecting useful source-learned relationships when target HR is unavailable. We address it with LGL. Using source HR, it measures message-correction benefits and combines them with temporal stability to construct Drive/Guard targets. At deployment, it predicts the guidance from LR inputs and prior states.

Contribution assessment and guidance learning. The source HR branch reuses A to align fine-scale evidence. With WMCA frozen, it replaces only message (k, n) while holding other messages, states, history, and random draws fixed. The resulting correction benefit is

$$b_{kn} = \ell_n(\widehat{y}_A, y) - \ell_n(\widehat{y}^{(kn)}, y). \tag{5}$$

The error ℓ_n weights valid pixels by normalized assignments; $b_{kn} > 0$ denotes improvement. The LR Guidance Predictor combines globally pooled current features with projected G^p, M^p, L^p, then uses two Softplus heads to predict nonnegative Drive and Guard per relationship.

Frozen HR reference scores D^*, R^* supervise the LR heads through KL distribution matching and weighted score regression. The former learns relative Drive priorities; the latter preserves Drive/Guard magnitudes for message weighting and fusion. Learning details appear in Appendix C.1; inference uses only LR features and prior states.

Source-scaled gains are aggregated over non-overlapping blocks: $b_{kn}^{(e)}$ is the block mean, v_{kn} the between-block variance, and $s_{kn} = 2\,\mathrm{Mean}_e\,\mathbf{1}[b_{kn}^{(e)} > 0] - 1$ the sign consistency.

Drive: prioritizing stable benefits. Drive weights positive gain by variance attenuation and sign consistency, prioritizing relationships that are both useful and stable. Predicted Drive controls expert routing and message weighting.

Guard: suppressing unreliable refinement. Guard combines adverse gain, temporal variability, message posterior–prior KL, and sign inconsistency. These complementary signals avoid treating representation discrepancy alone as harm; predicted Guard directly suppresses unsupported spatial refinement. The targets are

$$\begin{aligned} D_{kn} &= [b_{kn}]_+ \exp(-v_{kn})\sigma(\alpha_D s_{kn}), \\ R_{kn} &= [-b_{kn}]_+ + \lambda_v v_{kn} + \lambda_g\,\mathrm{KL}(q_{kn} \| p_{kn}) + \lambda_c(1 - |s_{kn}|). \end{aligned} \tag{6}$$

Here, $[\cdot]_+$ is the positive part; all coefficients are nonnegative and $\alpha_D > 0$. Drive and Guard may both be high because they encode usefulness and unreliability separately.

Content-aware spatial application. The current assignment A maps relation scores and messages onto the grid:

$$\begin{aligned} S(k, u) &= \sum_n A(n, u)\widehat{D}_{kn}, \\ U(u) &= \sum_n A(n, u) \sum_k \widehat{D}_{kn} M_{kn}^p, \\ (d, r) &= \mathrm{Up}\left(\frac{1}{K} \sum_{k,n} A(n, \cdot)[\widehat{D}_{kn}, \widehat{R}_{kn}] \right). \end{aligned} \tag{7}$$

S guides expert routing, U provides Drive-weighted messages, and d and r regulate fusion. None of these operations requires pixel correspondence between regions.

4.4 Adaptive Spatiotemporal Reconstruction

The third challenge is the varying predictive value of recent temporal evolution and current spatial structure across regions and conditions. We address it with ASR. It combines WMCA's sequence prediction with LGL-guided spatial reconstruction: adaptive aggregation selects local transformations, windowed propagation models spatial dependencies, and a source-calibrated Drive/Guard gate adjusts their fusion using LR-available information.

Adaptive feature aggregation. A convolutional embedding feeds a mixture of experts whose dense router uses current features and spatial guidance:

$$\begin{aligned} s &= \mathrm{Conv}_{\mathrm{emb}}(x), \\ \pi &= \mathrm{Softmax}_E(\mathrm{CNN}_r([s, S])), \\ h &= \mathrm{MoE}(s; \pi) = \sum_{e=1}^{E} \pi_e \odot \mathrm{CNN}_e(s). \end{aligned} \tag{8}$$

Each expert uses convolutional layers. All experts process the full representation, and $\sum_e \pi_e(u) = 1$; content s and guidance S therefore jointly determine the spatially varying transformation.

Windowed spatial feature propagation. A windowed state-space model captures bounded spatial dependencies. Independently initialized windows are scanned along rows and columns in both directions; their features are fused and conditioned on sequence states and Drive-weighted messages:

$$\begin{aligned} &\bar{h} = \sum_{j \in \mathcal{D}} w_j \ \mathrm{WSSM}_j(h), \qquad w = \mathrm{Softmax}(\omega), \ C = \mathrm{Conv}_c([\bar{G}, \bar{L}, U]), \\ &z = \mathrm{Conv}_o\big(\bar{h} \odot \sigma(\mathrm{Conv}_g(C))\big), \quad \widehat{y}_B = \mathrm{Dec}_B([z, \bar{L}]). \end{aligned} \tag{9}$$

Here, $\mathcal{D}$ contains four directions, $\bar{L} = A^\top L^p$, and $\bar{G}$ broadcasts the mean global state; bilinear resizing aligns grids. Convolutional layers implement the conditioning, gating, and output projections. The decoder also receives $\bar{L}$ and uses convolutional layers with PixelShuffle upsampling. Source HR trains this path with WMCA and LGL frozen.

Drive/Guard fusion. A sign-constrained gate combines Drive d, Guard r, and path disagreement $\delta = |\widehat{y}_A - \widehat{y}_B|$:

$$\begin{aligned} g &= \sigma(ad - br + w_\delta \delta + b_0), \qquad a, b > 0, \\ \widehat{y} &= (1 - g) \odot \widehat{y}_A + g \odot \widehat{y}_B. \end{aligned} \tag{10}$$

Softplus ensures $a, b > 0$: Drive increases spatial-path weight, whereas Guard favors the sequence path. Because message benefit does not directly equal path advantage, the gate is calibrated on frozen source predictions using path-error preferences and fused reconstruction loss.

5 Experiments

Four questions guide the experiments: **Q1**, does the frozen model transfer using only target LR under the reported reference-field protocol? **Q2**, do source HR-conditioned state replacements reveal reconstruction utility beyond KL? **Q3**, can LGL recover useful Drive priorities from LR and improve source reconstruction? **Q4**, do regional input preferences vary, and do separate fusion controls show gains and fallback as Guard increases? Sections 5.2–5.5 address these questions in order. Component controls use separate source records and do not decompose the cross-region gains in Table 1.

5.1 Experimental Setup

Protocol and data. We evaluate daily temperature and precipitation over Australia (AUS), the contiguous United States (CONUS), and Europe (EU). Inputs comprise five consecutive frames, including the current time, of seven dynamic ERA5 variables at $0.25°$. Each model is trained in one source region and frozen for evaluation in the other two, yielding six directed transfers. Source data from 2015–2017 support model fitting; 2018 source data support selection and calibration; 2019 data are used for evaluation. Target HR is used only for scoring. Data products and preprocessing appear in Appendix A.

Table 1: Cross-region results under the reference-field protocol. MAE and RMSE are source-standardized; SSIM is dimensionless. Parenthesized subscripts equal method minus DriveGuard-WM. Bold and underlined values rank first and second; C/A/E denote CONUS/Australia/Europe.

Metric	Method	$C \to A$	$C \to E$	$A \to C$	$A \to E$	$E \to A$	$E \to C$	Mean	Worst
Temperature									
MAE ↓	U-Net	$0.8561_{(+.400)}$	$\underline{0.0813}_{(+.025)}$	$0.1322_{(+.084)}$	$0.0931_{(+.037)}$	$0.9220_{(+.441)}$	$0.1455_{(+.063)}$	$0.3717_{(+.175)}$	$0.9220_{(+.441)}$
	GeoFAR	$0.7136_{(+.258)}$	$0.1819_{(+.125)}$	$0.2656_{(+.218)}$	$0.1732_{(+.117)}$	$1.0817_{(+.601)}$	$0.2954_{(+.213)}$	$0.4519_{(+.255)}$	$1.0817_{(+.601)}$
	SGD	$\underline{0.6204}_{(+.164)}$	$0.1036_{(+.047)}$	$0.5552_{(+.507)}$	$0.4712_{(+.415)}$	$\underline{0.8330}_{(+.352)}$	$0.1415_{(+.059)}$	$0.4542_{(+.257)}$	$\underline{0.8330}_{(+.352)}$
	STVD	$0.6674_{(+.211)}$	$0.0816_{(+.025)}$	$\underline{0.1072}_{(+.059)}$	$\underline{0.0773}_{(+.022)}$	$0.8584_{(+.377)}$	$0.1309_{(+.048)}$	$\underline{0.3205}_{(+.124)}$	$0.8584_{(+.377)}$
	WeatherGFM	$0.7398_{(+.284)}$	$0.0830_{(+.026)}$	$0.1118_{(+.064)}$	$0.0971_{(+.041)}$	$0.9108_{(+.430)}$	$\underline{0.1299}_{(+.047)}$	$0.3454_{(+.149)}$	$0.9108_{(+.430)}$
	DriveGuard-WM	$\mathbf{0.4561}_{(+.000)}$	$\mathbf{0.0567}_{(+.000)}$	$\mathbf{0.0481}_{(+.000)}$	$\mathbf{0.0558}_{(+.000)}$	$\mathbf{0.4811}_{(+.000)}$	$\mathbf{0.0828}_{(+.000)}$	$\mathbf{0.1968}_{(+.000)}$	$\mathbf{0.4811}_{(+.000)}$
RMSE ↓	U-Net	$1.3482_{(+.768)}$	$0.1788_{(+.057)}$	$0.2443_{(+.116)}$	$0.1974_{(+.073)}$	$1.0406_{(+.427)}$	$0.2580_{(+.149)}$	$0.5446_{(+.265)}$	$1.3482_{(+.735)}$
	GeoFAR	$0.8546_{(+.274)}$	$0.3704_{(+.249)}$	$0.5054_{(+.377)}$	$0.4013_{(+.277)}$	$1.2289_{(+.616)}$	$0.5034_{(+.395)}$	$0.6440_{(+.365)}$	$1.2289_{(+.616)}$
	SGD	$\underline{0.7668}_{(+.186)}$	$0.2196_{(+.098)}$	$0.9327_{(+.804)}$	$0.8440_{(+.720)}$	$0.9755_{(+.362)}$	$0.2535_{(+.145)}$	$0.6654_{(+.386)}$	$0.9755_{(+.362)}$
	STVD	$0.7797_{(+.199)}$	$\underline{0.1642}_{(+.043)}$	$0.1975_{(+.069)}$	$\underline{0.1723}_{(+.048)}$	$\underline{0.9324}_{(+.319)}$	$\underline{0.2196}_{(+.111)}$	$\underline{0.4110}_{(+.132)}$	$\underline{0.9324}_{(+.319)}$
	WeatherGFM	$0.8578_{(+.277)}$	$0.1753_{(+.054)}$	$\underline{0.1961}_{(+.067)}$	$0.2084_{(+.084)}$	$1.0055_{(+.392)}$	$0.2224_{(+.114)}$	$0.4443_{(+.165)}$	$1.0055_{(+.392)}$
	DriveGuard-WM	$\mathbf{0.5804}_{(+.000)}$	$\mathbf{0.1215}_{(+.000)}$	$\mathbf{0.1287}_{(+.000)}$	$\mathbf{0.1242}_{(+.000)}$	$\mathbf{0.6133}_{(+.000)}$	$\mathbf{0.1087}_{(+.000)}$	$\mathbf{0.2795}_{(+.000)}$	$\mathbf{0.6133}_{(+.000)}$
SSIM ↑	U-Net	$0.7914_{(-.143)}$	$\underline{0.9463}_{(-.026)}$	$0.9229_{(-.042)}$	$\underline{0.9475}_{(-.027)}$	$0.8834_{(-.057)}$	$0.9110_{(-.055)}$	$0.9004_{(-.058)}$	$0.7914_{(-.143)}$
	GeoFAR	$0.6443_{(-.290)}$	$0.8301_{(-.142)}$	$0.7983_{(-.167)}$	$0.8685_{(-.106)}$	$0.6608_{(-.280)}$	$0.8356_{(-.130)}$	$0.7729_{(-.186)}$	$0.6443_{(-.290)}$
	SGD	$0.8805_{(-.054)}$	$0.9106_{(-.062)}$	$0.8843_{(-.081)}$	$0.9044_{(-.070)}$	$0.8829_{(-.058)}$	$0.9154_{(-.050)}$	$0.8964_{(-.062)}$	$0.8805_{(-.054)}$
	STVD	$\underline{0.8992}_{(-.036)}$	$0.9426_{(-.030)}$	$\underline{0.9233}_{(-.042)}$	$0.9471_{(-.027)}$	$\underline{0.9092}_{(-.031)}$	$\underline{0.9193}_{(-.047)}$	$\underline{0.9235}_{(-.035)}$	$\underline{0.8992}_{(-.036)}$
	WeatherGFM	$0.8240_{(-.111)}$	$0.9184_{(-.054)}$	$0.9105_{(-.055)}$	$0.9211_{(-.053)}$	$0.8562_{(-.084)}$	$0.9114_{(-.054)}$	$0.8903_{(-.068)}$	$0.8240_{(-.111)}$
	DriveGuard-WM	$\mathbf{0.9347}_{(+.000)}$	$\mathbf{0.9721}_{(+.000)}$	$\mathbf{0.9652}_{(+.000)}$	$\mathbf{0.9740}_{(+.000)}$	$\mathbf{0.9405}_{(+.000)}$	$\mathbf{0.9658}_{(+.000)}$	$\mathbf{0.9587}_{(+.000)}$	$\mathbf{0.9347}_{(+.000)}$
Precipitation									
MAE ↓	U-Net	$0.1379_{(+.099)}$	$0.1781_{(+.150)}$	$0.7167_{(+.593)}$	$0.1864_{(+.132)}$	$0.0897_{(+.052)}$	$0.1741_{(+.111)}$	$0.2472_{(+.190)}$	$0.7167_{(+.593)}$
	GeoFAR	$0.0900_{(+.051)}$	$0.0933_{(+.065)}$	$0.2171_{(+.093)}$	$0.1125_{(+.058)}$	$0.0755_{(+.038)}$	$0.1670_{(+.104)}$	$0.1259_{(+.068)}$	$0.2171_{(+.093)}$
	SGD	$\underline{0.0650}_{(+.026)}$	$0.1372_{(+.109)}$	$0.3096_{(+.186)}$	$0.1423_{(+.088)}$	$0.0777_{(+.040)}$	$0.2203_{(+.157)}$	$0.1587_{(+.101)}$	$0.3096_{(+.186)}$
	STVD	$0.0666_{(+.028)}$	$\underline{0.0645}_{(+.037)}$	$\underline{0.2041}_{(+.080)}$	$\underline{0.0965}_{(+.042)}$	$0.0636_{(+.026)}$	$\underline{0.1463}_{(+.083)}$	$\underline{0.1069}_{(+.049)}$	$\underline{0.2041}_{(+.080)}$
	WeatherGFM	$0.0733_{(+.035)}$	$0.1078_{(+.080)}$	$0.2106_{(+.087)}$	$0.1492_{(+.095)}$	$\underline{0.0592}_{(+.022)}$	$0.1505_{(+.087)}$	$0.1251_{(+.067)}$	$0.2106_{(+.087)}$
	DriveGuard-WM	$\mathbf{0.0388}_{(+.000)}$	$\mathbf{0.0280}_{(+.000)}$	$\mathbf{0.1238}_{(+.000)}$	$\mathbf{0.0546}_{(+.000)}$	$\mathbf{0.0377}_{(+.000)}$	$\mathbf{0.0634}_{(+.000)}$	$\mathbf{0.0577}_{(+.000)}$	$\mathbf{0.1238}_{(+.000)}$
RMSE ↓	U-Net	$0.6926_{(+.532)}$	$1.0542_{(+.917)}$	$2.8320_{(+2.377)}$	$1.0921_{(+.891)}$	$0.3472_{(+.157)}$	$0.5075_{(+.179)}$	$1.0876_{(+.842)}$	$2.8320_{(+2.377)}$
	GeoFAR	$0.2744_{(+.114)}$	$0.2568_{(+.120)}$	$0.7605_{(+.305)}$	$0.3923_{(+.191)}$	$0.3052_{(+.115)}$	$0.5317_{(+.204)}$	$0.4202_{(+.175)}$	$0.7605_{(+.305)}$
	SGD	$0.2557_{(+.096)}$	$0.5445_{(+.408)}$	$0.7018_{(+.246)}$	$0.6156_{(+.414)}$	$0.3085_{(+.118)}$	$0.5044_{(+.176)}$	$0.4884_{(+.243)}$	$0.7018_{(+.246)}$
	STVD	$\underline{0.2398}_{(+.080)}$	$\underline{0.2068}_{(+.070)}$	$\underline{0.6443}_{(+.189)}$	$\underline{0.2957}_{(+.094)}$	$\underline{0.2879}_{(+.098)}$	$\underline{0.4920}_{(+.164)}$	$\underline{0.3611}_{(+.116)}$	$\underline{0.6443}_{(+.189)}$
	WeatherGFM	$0.2485_{(+.088)}$	$0.3261_{(+.189)}$	$0.6781_{(+.223)}$	$0.7517_{(+.550)}$	$0.2978_{(+.108)}$	$0.5214_{(+.193)}$	$0.4706_{(+.225)}$	$0.7517_{(+.296)}$
	DriveGuard-WM	$\mathbf{0.1602}_{(+.000)}$	$\mathbf{0.1368}_{(+.000)}$	$\mathbf{0.4555}_{(+.000)}$	$\mathbf{0.2016}_{(+.000)}$	$\mathbf{0.1902}_{(+.000)}$	$\mathbf{0.3282}_{(+.000)}$	$\mathbf{0.2454}_{(+.000)}$	$\mathbf{0.4555}_{(+.000)}$
SSIM ↑	U-Net	$0.9231_{(-.062)}$	$0.8957_{(-.088)}$	$0.8412_{(-.135)}$	$0.9104_{(-.072)}$	$0.9563_{(-.031)}$	$0.9287_{(-.051)}$	$0.9092_{(-.073)}$	$0.8412_{(-.135)}$
	GeoFAR	$0.9528_{(-.032)}$	$0.9486_{(-.035)}$	$0.9315_{(-.044)}$	$0.9502_{(-.033)}$	$0.9619_{(-.025)}$	$0.9431_{(-.036)}$	$0.9480_{(-.034)}$	$0.9315_{(-.044)}$
	SGD	$0.9642_{(-.020)}$	$0.9360_{(-.047)}$	$0.9473_{(-.029)}$	$0.9448_{(-.038)}$	$0.9655_{(-.022)}$	$0.9517_{(-.028)}$	$0.9516_{(-.031)}$	$0.9360_{(-.040)}$
	STVD	$\underline{0.9657}_{(-.019)}$	$\underline{0.9661}_{(-.017)}$	$\underline{0.9524}_{(-.023)}$	$\underline{0.9653}_{(-.017)}$	$\underline{0.9682}_{(-.019)}$	$\underline{0.9580}_{(-.021)}$	$\underline{0.9626}_{(-.020)}$	$\underline{0.9524}_{(-.023)}$
	WeatherGFM	$0.9613_{(-.023)}$	$0.9504_{(-.033)}$	$0.9488_{(-.027)}$	$0.9466_{(-.036)}$	$0.9651_{(-.022)}$	$0.9532_{(-.026)}$	$0.9542_{(-.028)}$	$0.9466_{(-.029)}$
	DriveGuard-WM	$\mathbf{0.9846}_{(+.000)}$	$\mathbf{0.9832}_{(+.000)}$	$\mathbf{0.9758}_{(+.000)}$	$\mathbf{0.9827}_{(+.000)}$	$\mathbf{0.9873}_{(+.000)}$	$\mathbf{0.9794}_{(+.000)}$	$\mathbf{0.9822}_{(+.000)}$	$\mathbf{0.9758}_{(+.000)}$

Baselines and metrics. Baselines are U-Net, SGD, STVD, WeatherGFM, and GeoFAR (Ronneberger et al., 2015; Tu et al., 2025; Srivastava et al., 2024; Zhao et al., 2025; Xu et al., 2026). Source-standardized MAE/RMSE and dimensionless SSIM use common scales and masks within each transfer. Regional reference products and unresolved ERA5-filled-cell scoring limit conclusions to the reported reference-field protocol. Component records are separate from transfer results; refined single-message correction, conditional guidance, and dense MoE still require implementation-matched validation (Appendix C.2).

5.2 Main Results: Cross-Region Downscaling

Table 1 evaluates frozen LR-only transfer without target HR supervision or fine-tuning. DriveGuard-WM leads all **36 comparisons**. Relative to STVD, the strongest macro-MAE baseline, it lowers macro MAE by **38.6%** for temperature and **46.0%** for precipitation. Macro RMSE falls by **32.0%** for both variables, while macro SSIM rises from 0.9235 to 0.9587 for temperature and from 0.9626 to 0.9822 for precipitation. Worst-direction MAE falls by **42.2%** and **39.3%**, respectively. Against the strongest baseline chosen separately for each transfer, MAE reductions range from **26.5%** to **55.1%** for temperature and from **36.3%** to **56.7%** for precipitation. These direction-specific margins support consistent LR-only transfer across the three regions under the reported reference-field protocol. Full directional results appear in Appendix B.1.

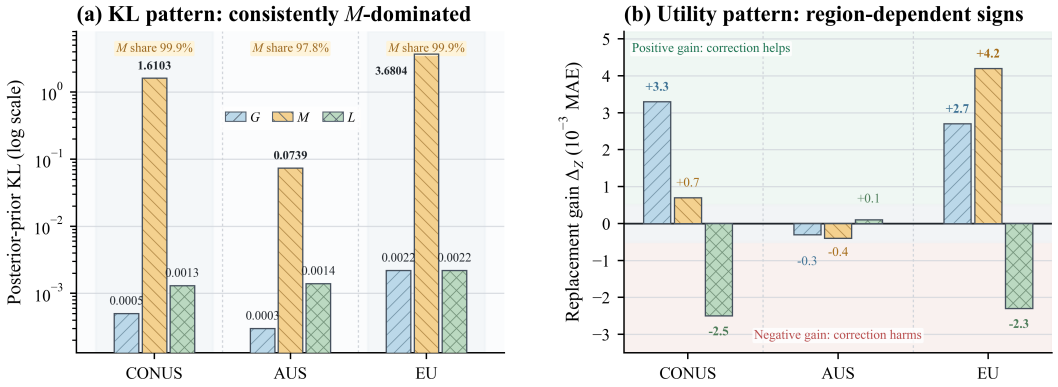


Figure 3: WMCA diagnostics. (a) State KL. (b) HR-replacement MAE gain Δ_Z (10^{-3}).

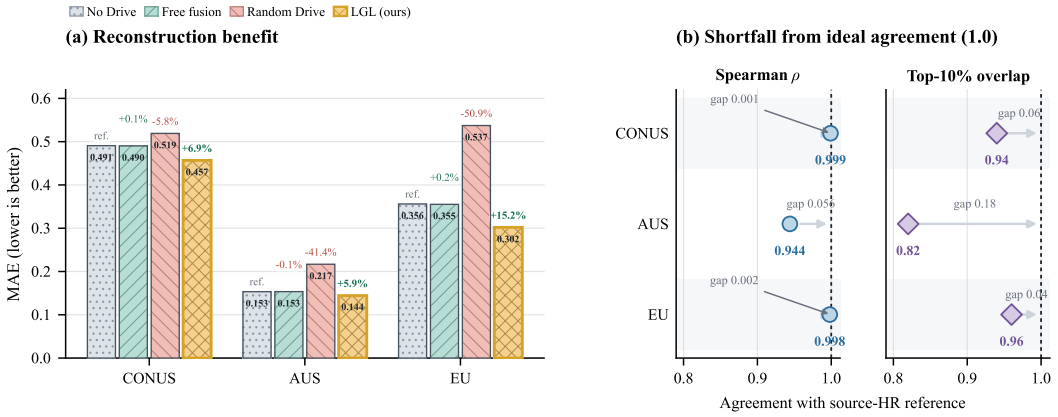


Figure 4: LGL guidance. (a) MAE by guidance control. (b) Source-HR Drive rank agreement.

5.3 WMCA State Change and Reconstruction Utility

Figure 3 compares posterior/prior KL with the MAE gain from controlled HR-conditioned state replacement, testing whether large state changes are also useful for reconstruction.

With $\Delta_Z = \mathrm{MAE}_{\mathrm{before}} - \mathrm{MAE}_{\mathrm{after}}$, M accounts for at least **97.8%** of KL in every region. Yet in CONUS, replacing G yields a $+3.3$ gain, compared with only $+0.7$ for M; in EU, M yields $+4.2$, exceeding G at $+2.7$. The ordering of useful corrections changes across regions despite the persistent concentration of KL in M. Replacing L is harmful in CONUS (-2.5) and EU (-2.3), while all AUS effects are near zero, giving a third regime with little correction benefit. These interventions use held-out source data and fixed model parameters, so each Δ_Z reflects the error change under the same predictor. The mismatch between KL and error reduction makes reconstruction utility the relevant source signal for relationship assessment: positive and stable effects motivate Drive, while negative or uncertain effects motivate Guard (Section 4.3). Cross-region transfer is evaluated separately in Section 5.2; further controls appear in Appendix B.2.

5.4 From Contribution Supervision to LR Guidance: LGL Effectiveness

Figure 4(a) compares LGL with no Drive, random Drive, and free fusion on held-out source data; Figure 5 evaluates Guard separately. Relative to no Drive, LGL reduces MAE by **6.9%**, **5.9%**, and **15.2%** in CONUS, AUS, and EU, respectively, and outperforms the other controls in every region. Random Drive increases MAE in all three regions; LGL reduces its error by **12.0%**, **33.4%**, and **43.8%**, respectively. Free fusion changes MAE by at most **0.2%** relative to no Drive. Within these controls, reconstruction benefit follows the learned relation priorities rather than generic fusion or arbitrary Drive scores. The LR predictor also agrees with the source-HR Drive reference, with Spearman correlations of 0.944–0.999 and top-10% overlaps of 0.82–0.96 (Figure 4(b)). AUS has both the lowest rank agreement (0.944 Spearman and 0.82 overlap) and the smallest MAE reduction,

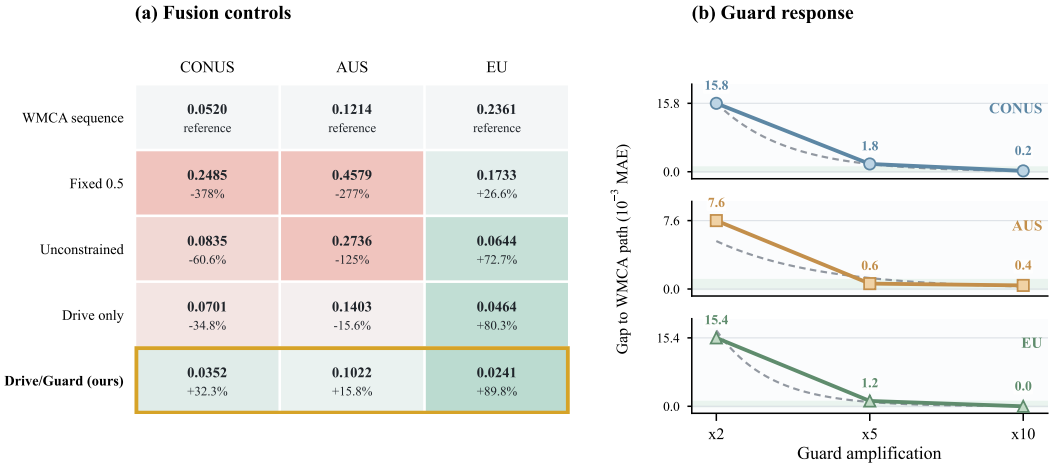


Figure 5: Drive/Guard fusion and fallback. (a) MAE across fusion controls. (b) Gap to the WMCA path under Guard amplification (10^{-3} MAE). Full values appear in Appendix Tables A8 and A10.

whereas EU combines the highest overlap (0.96) with the largest reduction. This pattern links LR-recovered source priorities to reconstruction utility, supporting the source-side guidance step needed when target HR is unavailable.

5.5 Guard-Controlled Adaptive Fusion

At full LR resolution, held-out source tests favor the current frame in CONUS and EU but temporal history in AUS (Appendix Figure A3 and Table A7). Adding history lowers AUS MAE by **4.3%**, whereas it raises CONUS and EU MAE by **2.0%** and **6.5%**. This reversal motivates adaptive information use; separate fusion controls evaluate the gate itself.

Figure 5(a) shows that the full gate reduces sequence-path MAE by **32.3%**, **15.8%**, and **89.8%** in CONUS, AUS, and EU, respectively. Relative to Drive-only fusion, it reduces MAE by another **49.8%**, **27.2%**, and **48.1%**. The spatial path alone is substantially weaker than the sequence path in CONUS (0.4843 versus 0.0520) and AUS (0.9098 versus 0.1214), yet their joint gate improves on the stronger path in both. Neither single-term gate does so in these regions; in EU, both single-term gates help, but the joint gate performs best (Appendix Table A8). It also outperforms fixed and unconstrained fusion throughout. Thus, useful spatial detail needs controlled contribution rather than a fixed preference for either path.

With path predictions and other gate inputs fixed, amplifying Guard progressively reduces the gap to the WMCA sequence path in all three regions (Figure 5(b)). The gap is already at most 1.8×10^{-3} MAE at $5\times$ Guard and reaches at most 0.4×10^{-3} at $10\times$. This graded response shows smooth fallback toward $\widehat{y}_A$ under the controlled intervention.

6 Conclusion

DriveGuard-WM transfers source-region reconstruction knowledge from dynamic LR inputs through contribution assessment, LR-conditioned Drive/Guard guidance, and adaptive fusion. Across six directed transfers, it lowers macro MAE by **38.6%** for temperature and **46.0%** for precipitation relative to the strongest macro-MAE baseline, without target HR training or adaptation. Separate held-out source analyses show that the largest state change is not always the most useful correction, learned Drive priorities outperform random guidance, and joint Drive/Guard fusion improves on either single-term gate. Together, these results support assessing source relationships by reconstruction benefit and regulating the contribution of spatial detail. Future work will extend evaluation to more climates and extremes and study calibration, efficiency, and matched structural ablations.

AI USE STATEMENT

Generative AI assisted with language editing, LaTeX formatting, and consistency checks of author-provided materials. It did not generate experimental data or results. The authors are responsible for verifying the manuscript and its claims.

REPRODUCIBILITY STATEMENT

Appendix A records data and evaluation choices; Appendix B preserves complete numerical evidence; Appendix C specifies training interfaces and the distinction between recorded and refined formulations. Missing controls and configuration details are explicitly identified.

ETHICS STATEMENT

This work does not involve human subjects, private data, or other significant ethical concerns. All datasets used are publicly available and used for research purposes.

REFERENCES

Michael Aich, Philipp Hess, Baoxiang Pan, Sebastian Bathiany, Yu Huang, and Niklas Boers. Conditional diffusion models for downscaling and bias correction of earth system model precipitation. *Geoscientific Model Development*, 19(4):1791–1808, 2026.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Jorge Baño-Medina, Rodrigo Manzanas, and José Manuel Gutiérrez. Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4):2109–2124, 2020.

Jorge Baño-Medina, Rodrigo Manzanas, and José Manuel Gutiérrez. On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections. *Climate Dynamics*, 57(11):2941–2951, 2021.

Jorge Baño-Medina, Rodrigo Manzanas, Ezequiel Cimadevilla, Jesús Fernández, Jose González-Abad, Antonio S. Cofiño, and José Manuel Gutiérrez. Downscaling multi-model climate projection ensembles with deep learning (DeepESD): contribution to CORDEX EUR-44. *Geoscientific Model Development*, 15(17):6747–6758, 2022. doi: 10.5194/gmd-15-6747-2022.

Jorge Baño-Medina, Maialen Iturbide, Jesús Fernández, and José Manuel Gutiérrez. Transferability and explainability of deep learning emulators for regional climate model projections: Perspectives for future applications. *Artificial Intelligence for the Earth Systems*, 3(4):e230099, 2024.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of machine learning research*, 22(2):1–55, 2021.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22405–22418, 2021.

Richard C Cornes, Gerard Van Der Schrier, Else JM Van Den Besselaar, and Philip D Jones. An ensemble version of the e-obs temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, 123(17):9391–9409, 2018.

Ting-Yu Dai and Hayato Ushijima-Mwesigwa. PrecipDiff: Leveraging image diffusion models to enhance satellite-based precipitation observations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):27932–27939, 2025. doi: 10.1609/aaai.v39i27.35010.

Christopher Daly, Michael Halbleib, Joseph I Smith, Wayne P Gibson, Matthew K Doggett, George H Taylor, Jan Curtis, and Phillip P Pasteris. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states. *International Journal of Climatology: a Journal of the Royal Meteorological Society*, 28(15):2031–2064, 2008.

Imke Durre, Michael F. Squires, Russell S. Vose, Anthony Arguez, Wendy S. Gross, Jared R. Rennie, and Carl J. Schreck. NOAA nClimGrid-Daily Version 1 – Daily Gridded Temperature and Precipitation for the Contiguous United States since 1951. NOAA National Centers for Environmental Information, https://doi.org/10.25921/c4gt-r169, 2022. Dataset.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17489–17521. PMLR, 2025. URL https://proceedings.mlr.press/v267/fotiadis25a.html.

Ramón Fuentes-Franco, Kristofer Krus, Mikhail Ivanov, Torben Koenigk, Fuxing Wang, and Aitor Aldama-Campino. Pan-european high-resolution downscaling using deep learning. *Journal of Geophysical Research: Machine Learning and Computation*, 2(4):e2025JH000630, 2025.

Chris Funk, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell, et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific data*, 2(1):150066, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Luca Glawion, Julius Polz, Harald Kunstmann, Benjamin Fersch, and Christian Chwala. Global spatio-temporal era5 precipitation downscaling to km and sub-hourly scale using generative ai. *npj Climate and Atmospheric Science*, 8(1):219, 2025.

Jose González-Abad, Maialen Iturbide, Alfonso Hernanz, and José Manuel Gutiérrez. Pre-training for deep statistical climate downscaling: enhancing consistency and robustness across regional datasets. *Geoscientific Model Development*, 19(12):5781–5804, 2026.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Paula Harder, Luca Schmidt, Francis Pelletier, Nicole Ludwig, Matthew Chantry, Christian Lessig, Alex Hernandez-Garcia, and David Rolnick. Benchmarking the geographic generalization of deep learning models for precipitation downscaling. *Scientific Reports*, 16(1):3733, 2026.

Alfonso Hernanz, Carlos Correa, Juan-Carlos Sánchez-Perrino, Ignacio Prieto-Rico, Esteban Rodríguez-Guisado, Marta Domínguez, and Ernesto Rodríguez-Camino. On the limitations of deep learning for statistical downscaling of climate change projections: The transferability and the extrapolation issues. *Atmospheric Science Letters*, 25(2):e1195, 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.

Philipp Hess, Michael Aich, Baoxiang Pan, and Niklas Boers. Fast, scale-adaptive and uncertainty-aware downscaling of earth system model fields with generative machine learning. *Nature Machine Intelligence*, 7(3):363–373, 2025.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

David A Jones, William Wang, and Robert Fawcett. High-quality spatial climate data-sets for australia. *Australian Meteorological and Oceanographic Journal*, 58(4):233–248, 2009.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Wouter M. Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):766–785, 2021. doi: 10.1109/TPAMI.2019.2945942.

Jussi Leinonen, Daniele Nerini, and Alexis Berne. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223, 2021. doi: 10.1109/TGRS.2020.3032790.

Fenghua Ling, Zeyu Lu, Jing-Jia Luo, Lei Bai, Swadhin K Behera, Dachao Jin, Baoxiang Pan, Huidong Jiang, and Toshio Yamagata. Diffusion model-based probabilistic downscaling for 180-year east asian climate reconstruction. *npj Climate and Atmospheric Science*, 7(1):131, 2024.

Yumin Liu, Auroop R Ganguly, and Jennifer Dy. Climate downscaling using ynet: A deep convolutional network with skip connections and fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3145–3153, 2020.

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.

David Lopez-Paz, Leon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. URL https://bottou.org/papers/lopez-paz-2016.

Douglas Maraun, Frederick Wetterhall, Anderson M Ireson, Richard E Chandler, Elizabeth J Kendon, Martin Widmann, Stephan Brienen, Henning W Rust, Tobias Sauter, Matthias Themeßl, et al. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of geophysics*, 48(3), 2010.

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.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023.

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.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International conference on machine learning*, pp. 18347–18377. PMLR, 2022.

Neelesh Rampal, Peter B Gibson, Steven Sherwood, and Gab Abramowitz. On the extrapolation of generative adversarial networks for downscaling precipitation extremes in warmer climates. *Geophysical Research Letters*, 51(23):e2024GL112492, 2024.

Neelesh Rampal, Peter B Gibson, Steven Sherwood, Gab Abramowitz, and Sanaa Hobeichi. A reliable generative adversarial network approach for climate downscaling and weather generation. *Journal of Advances in Modeling Earth Systems*, 17(1):e2024MS004668, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Jonathan Schmidt, Luca Schmidt, Felix M Strnad, Nicole Ludwig, and Philipp Hennig. A generative framework for probabilistic, spatiotemporally coherent downscaling of climate simulation. *npj Climate and Atmospheric Science*, 8(1):270, 2025.

Yingkai Sha, David John Gagne II, Gregory West, and Roland Stull. Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: Daily maximum and minimum 2-m temperature. *Journal of Applied Meteorology and Climatology*, 59(12):2057–2073, 2020a. doi: 10.1175/JAMC-D-20-0057.1.

Yingkai Sha, David John Gagne II, Gregory West, and Roland Stull. Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation. *Journal of Applied Meteorology and Climatology*, 59(12):2075–2092, 2020b. doi: 10.1175/JAMC-D-20-0058.1.

Pedro MM Soares, Frederico Johannsen, Daniela CA Lima, Gil Lemos, Virgílio A Bento, and Angelina Bushenkova. High-resolution downscaling of cmip6 earth system and global climate models using deep learning for iberia. *Geoscientific Model Development*, 17(1):229–259, 2024.

Prakhar Srivastava, Ruihan Yang, Gavin Kerrigan, Gideon Dresdner, Jeremy McGibbon, Christopher Bretherton, and Stephan Mandt. Precipitation downscaling with spatiotemporal video diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 56374–56400, 2024. doi: 10.52202/079017-1795.

Sencan Sun, Congyi Nai, Baoxiang Pan, Wentao Li, Lu Li, Xin Li, Efi Foufoula-Georgiou, and Yanluan Lin. Fusion of multi-source precipitation records via coordinate-based generative models. *Nature Communications*, 17:1227, 2026. doi: 10.1038/s41467-025-67987-9.

Ruian Tie, Wenbo Xiong, Zhengyu Shi, Xinyu Su, Libo Wu, Hao Li, et al. Zero-shot statistical downscaling via diffusion posterior sampling. *arXiv preprint arXiv:2601.21760*, 2026.

Siwei Tu, Ben Fei, Weidong Yang, Fenghua Ling, Hao Chen, Zili Liu, Kun Chen, Hang Fan, Wanli Ouyang, and Lei Bai. Satellite observations guided diffusion model for accurate meteorological states at arbitrary resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28071–28080, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tu_Satellite_Observations_Guided_Diffusion_Model_for_Accurate_Meteorological_States_at_CVPR_2025_paper.html.

Thomas Vandal, Evan Kodra, Sangram Ganguly, Andrew Michaelis, Ramakrishna Nemani, and Auroop R Ganguly. Deepsd: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 1663–1672, 2017.

Thomas Vandal, Evan Kodra, and Auroop R Ganguly. Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation: T. vandal et al. *Theoretical and Applied Climatology*, 137(1):557–570, 2019.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021a.

Fang Wang, Di Tian, Lisa Lowe, Latif Kalin, and John Lehrter. Deep learning for daily precipitation and temperature downscaling. *Water Resources Research*, 57(4):e2020WR029308, 2021b.

Robert L Wilby and Thomas ML Wigley. Downscaling general circulation model output: a review of methods and limitations. *Progress in physical geography*, 21(4):530–548, 1997.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pp. 2226–2240. PMLR, 2023.

Chang Xu, Gencer Sumbul, Li Mi, Robin Zbinden, and Devis Tuia. Geofar: Geography-informed frequency-aware super-resolution for climate data. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, pp. 3320–3328, 2014.

Xiangyu Zhao, Zhiwang Zhou, Wenlong Zhang, Yihao Liu, Xiangyu Chen, Junchao Gong, Hao Chen, Ben Fei, Shiqi Chen, Wanli Ouyang, Xiao-Ming Wu, and Lei Bai. WeatherGFM: Learning a weather generalist foundation model via in-context learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=izjNI5bcOV.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023. doi: 10.1109/TPAMI.2022.3195549.

Honglin Zhu and Qiming Zhou. Advancing satellite-derived precipitation downscaling in data-sparse area through deep transfer learning. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

APPENDIX ROADMAP

Appendix A: data, protocol, and recorded settings. Appendix B: complete transfer metrics, diagnostics for the world model, guidance learner, and adaptive reconstruction, and source-fit records. Appendix C: training interfaces, formulation scope, and outstanding validation. Appendix D: extended literature context.

A DATA, PROTOCOL, AND RECORDED SETTINGS

A.1 REGIONAL DATA AND TEMPORAL ROLES

The benchmark covers daily temperature and precipitation in CONUS, AUS, and EU. ERA5 inputs (Hersbach et al., 2020) have a 0.25° grid and seven dynamic channels: near-surface temperature, two horizontal wind components, surface pressure, downward surface solar radiation, top-of-atmosphere incident solar radiation, and total precipitation. The five-frame sequence includes the field being downscaled. Variables are modeled separately. Each run trains on paired LR and HR data from one source region and evaluates the frozen model in the other two, yielding six directed transfers. Source fitting and normalization use 2015–2017; source selection and calibration use 2018; held-out and cross-region scoring use 2019. Target labels never determine model choice, score direction, thresholds, normalization, or adaptation. Figure A1 summarizes the regional grids and source-only training, calibration, and target-evaluation protocol.

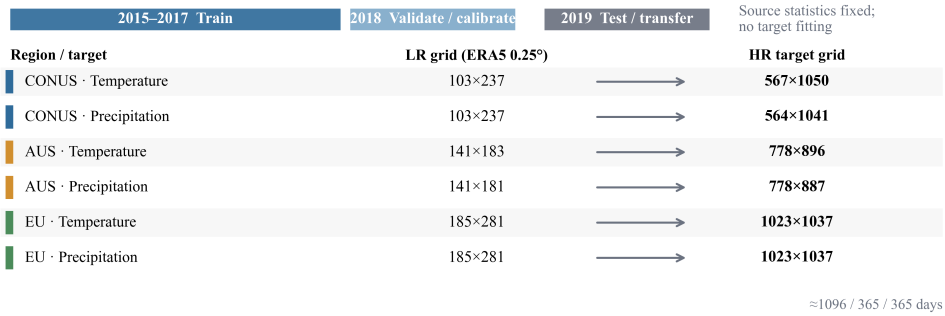
Freezing stages does not itself establish temporal independence. Reproduction requires the block boundaries used for relation statistics and out-of-fold predictions, whether upstream models are retrained per block, handling of five-frame windows that cross boundaries, and the separation of validation from gate calibration.

Reference fields and units. CONUS uses nClimGrid-Daily temperature and PRISM precipitation (Durre et al., 2022; Daly et al., 2008), AUS uses AGCD/AWAP temperature and CHIRPS precipitation (Jones et al., 2009; Funk et al., 2015), and EU uses E-OBS (Cornes et al., 2018). Temperature is converted from Celsius to Kelvin. Precipitation retains the product's millimeter or millimeter-per-day convention, is clipped to nonnegative values, and is resampled onto the regional HR grid. Reported preprocessing fills some missing ocean pixels with interpolated ERA5. Whether these filled cells enter scoring and HR normalization is not fully documented. Consequently, the reported comparison measures agreement with the prepared reference fields. Independent HR recovery requires separate original-observation and filled-cell masks and evaluation on original-HR valid pixels.

Processed array dimensions. In CONUS, the LR grid is 103×237 for both variables. Its temperature HR grid is 567×1050, and its precipitation HR grid is 564×1041. In AUS, the temperature LR and HR grids are 141×183 and 778×896, respectively; the precipitation grids are 141×181 and 778×887. In EU, both variables use an LR grid of 185×281 and an HR grid of 1023×1037. These are processed array dimensions, not native observational resolutions or uniform physical scale factors.

Table A1: Regional HR reference products used in the reported evaluation.

Region	Temperature reference	Precipitation reference
CONUS	nClimGrid-Daily Tavg, v1.0.0	PRISM, daily D2
AUS	AGCD/AWAP Tavg, AGCD v1.0.0	CHIRPS v2.0
EU	E-OBS TG ensemble mean, v31.0e	E-OBS RR ensemble mean, v29.0e



(a) Regional grids and common temporal split

Figure A1: Regional grids and the common temporal split. Models are trained and calibrated only in the source region, then frozen for target-region scoring in 2019. Array dimensions do not imply uniform physical resolution.

Table A2: Temporal roles and remaining boundary specifications.

Data	Reported role	Boundary to document
2015–2017 source	Fitting and normalization	Freeze source statistics thereafter
Non-overlapping training blocks	Relation statistics and out-of-fold path predictions	Block boundaries, upstream retraining, window crossings
2018 source	Model selection and fusion calibration	Document division between these roles
2019 source/target	Held-out and cross-region evaluation	No fitting, selection, or adaptation

A.2 Metrics and Aggregation

For source s and target u, let m_i identify valid scoring values, and let σ_s^{HR} be the HR standard deviation estimated on source training data. The reported normalized errors are

$$E_{s\to u} = \frac{\sum_i m_i |\widehat{y}_i - y_i|}{\sigma_s^{\mathrm{HR}} \sum_i m_i}, \tag{11}$$

$$\mathrm{RMSE}_{s\to u} = \frac{1}{\sigma_s^{\mathrm{HR}}} \sqrt{\frac{\sum_i m_i (\widehat{y}_i - y_i)^2}{\sum_i m_i}}, \tag{12}$$

$$E_{\mathrm{macro}} = \frac{1}{6} \sum_{s\neq u} E_{s\to u}, \qquad E_{\mathrm{worst}} = \max_{s\neq u} E_{s\to u}. \tag{13}$$

All methods share the same source scale and scoring mask within each transfer direction. Errors in physical units are obtained by multiplying normalized MAE or RMSE by the corresponding source HR standard deviation. Pearson correlation uses the same valid values across space and time, and SSIM is computed separately for each day before averaging over time. The mean of the six directional correlations is an arithmetic summary rather than a correlation pooled across samples. For SSIM and Pearson correlation, the worst-direction summary is the minimum; for MAE and RMSE, it is the maximum. All summary values are recomputed from the reported four-decimal directional entries. The headline macro-MAE reductions of 38.6% for temperature and 46.0% for precipitation use STVD as the fixed comparator. Figure A2 instead compares with the strongest baseline in each direction.

Table A3: Complete temperature transfer metrics. MAE and RMSE are source-standardized; Pearson correlation and SSIM are dimensionless. Boldface indicates the best entry for each metric and column.

Metric	Method	$C \to A$	$C \to E$	$A \to C$	$A \to E$	$E \to A$	$E \to C$	Mean	Worst
MAE ↓	U-Net	0.8561	0.0813	0.1322	0.0931	0.9220	0.1455	0.3717	0.9220
	GeoFAR	0.7136	0.1819	0.2656	0.1732	1.0817	0.2954	0.4519	1.0817
	SGD	0.6204	0.1036	0.5552	0.4712	0.8330	0.1415	0.4542	0.8330
	STVD	0.6674	0.0816	0.1072	0.0773	0.8584	0.1309	0.3205	0.8584
	WeatherGFM	0.7398	0.0830	0.1118	0.0971	0.9108	0.1299	0.3454	0.9108
	DriveGuard-WM	**0.4561**	**0.0567**	**0.0481**	**0.0558**	**0.4811**	**0.0828**	**0.1968**	**0.4811**
RMSE ↓	U-Net	1.3482	0.1788	0.2443	0.1974	1.0406	0.2580	0.5446	1.3482
	GeoFAR	0.8546	0.3704	0.5054	0.4013	1.2289	0.5034	0.6440	1.2289
	SGD	0.7668	0.2196	0.9327	0.8440	0.9755	0.2535	0.6654	0.9755
	STVD	0.7797	0.1642	0.1975	0.1723	0.9324	0.2196	0.4110	0.9324
	WeatherGFM	0.8578	0.1753	0.1961	0.2084	1.0055	0.2224	0.4443	1.0055
	DriveGuard-WM	**0.5804**	**0.1215**	**0.1287**	**0.1242**	**0.6133**	**0.1087**	**0.2795**	**0.6133**
Pearson r ↑	U-Net	0.7650	0.9671	0.9559	0.9595	0.9123	0.9585	0.9197	0.7650
	GeoFAR	0.5512	0.8313	0.7990	0.8012	0.6138	0.8265	0.7372	0.5512
	SGD	0.9251	0.9507	0.8282	0.8272	0.9473	0.9687	0.9079	0.8272
	STVD	0.9598	0.9723	0.9697	0.9664	0.9660	0.9736	0.9680	0.9598
	WeatherGFM	0.9391	0.9626	0.9720	0.9556	0.9595	0.9713	0.9600	0.9391
	DriveGuard-WM	**0.9788**	**0.9886**	**0.9890**	**0.9882**	**0.9821**	**0.9914**	**0.9864**	**0.9788**
SSIM ↑	U-Net	0.7914	0.9463	0.9229	0.9475	0.8834	0.9110	0.9004	0.7914
	GeoFAR	0.6443	0.8301	0.7983	0.8685	0.6608	0.8356	0.7729	0.6443
	SGD	0.8805	0.9106	0.8843	0.9044	0.8829	0.9154	0.8964	0.8805
	STVD	0.8992	0.9426	0.9233	0.9471	0.9092	0.9193	0.9235	0.8992
	WeatherGFM	0.8240	0.9184	0.9105	0.9211	0.8562	0.9114	0.8903	0.8240
	DriveGuard-WM	**0.9347**	**0.9721**	**0.9652**	**0.9740**	**0.9405**	**0.9658**	**0.9587**	**0.9347**

A.3 Recorded Architecture and Optimization

The original configuration uses $K = 8$ global states and $N = 64$ local states. Its latent dimension is 64, the GRU hidden dimension is 128, and each input contains five frames with seven dynamic channels. The spatial path has four experts and routes each position to the two experts with the highest scores. The windowed state-space model uses windows of size 8 with an overlap of 2, scans in four directions, and has a state dimension of 64. Training uses Adam with an initial learning rate of 10^{-4}, ReduceLROnPlateau, a batch size of 4, and a 20-epoch warm-up for the KL loss. The original gate has a reported temperature of 0.5 and an auxiliary monotonicity penalty. The model has a reported 15.20 million parameters and 21.2 billion floating-point operations (FLOPs). These settings document the historical records rather than the refined dense MoE and conditional supervision. The final configuration must specify training budgets for each stage, loss weights and reduction rules, handling at window boundaries, input-grid conventions for FLOP counts, and adaptations to baseline inputs. Baseline pretraining sources, sampling budgets, point-prediction aggregation, and model-selection choices also require documentation. The reported parameter and FLOP totals alone do not establish a capacity or efficiency advantage.

B Additional Experimental Results

B.1 Complete Cross-Region Metrics

Tables A3 and A4 retain all three main-table metrics and add Pearson correlation under the identical six-direction protocol. The reported complete system is best on all four measures in each direction for both variables. The records do not include multi-seed uncertainty estimates, so these are numerical comparisons without statistical-significance claims.

Figure A2 visualizes the six directions. Temperature worst-direction MAE falls to 0.4811, 42.2% below the best baseline worst-direction value (0.8330, SGD); precipitation falls from 0.2041 to 0.1238 (39.3%). Figure-level improvements use each direction's strongest baseline rather than the fixed macro-MAE comparator.

Table A4: Complete precipitation transfer metrics under the same reference-field protocol.

Metric	Method	$C \to A$	$C \to E$	$A \to C$	$A \to E$	$E \to A$	$E \to C$	Mean	Worst
MAE ↓	U-Net	0.1379	0.1781	0.7167	0.1864	0.0897	0.1741	0.2472	0.7167
	GeoFAR	0.0900	0.0933	0.2171	0.1125	0.0755	0.1670	0.1259	0.2171
	SGD	0.0650	0.1372	0.3096	0.1423	0.0777	0.2203	0.1587	0.3096
	STVD	0.0666	0.0645	0.2041	0.0965	0.0636	0.1463	0.1069	0.2041
	WeatherGFM	0.0733	0.1078	0.2106	0.1492	0.0592	0.1505	0.1251	0.2106
	DriveGuard-WM	**0.0388**	**0.0280**	**0.1238**	**0.0546**	**0.0377**	**0.0634**	**0.0577**	**0.1238**
RMSE ↓	U-Net	0.6926	1.0542	2.8320	1.0921	0.3472	0.5075	1.0876	2.8320
	GeoFAR	0.2744	0.2568	0.7605	0.3923	0.3052	0.5317	0.4202	0.7605
	SGD	0.2557	0.5445	0.7018	0.6156	0.3085	0.5044	0.4884	0.7018
	STVD	0.2398	0.2068	0.6443	0.2957	0.2879	0.4920	0.3611	0.6443
	WeatherGFM	0.2485	0.3261	0.6781	0.7517	0.2978	0.5214	0.4706	0.7517
	DriveGuard-WM	**0.1602**	**0.1368**	**0.4555**	**0.2016**	**0.1902**	**0.3282**	**0.2454**	**0.4555**
Pearson r ↑	U-Net	0.9768	0.9543	0.9126	0.9684	0.9821	0.9615	0.9593	0.9126
	GeoFAR	0.9891	0.9865	0.9720	0.9871	0.9894	0.9782	0.9837	0.9720
	SGD	0.9912	0.9703	0.9784	0.9768	0.9915	0.9829	0.9819	0.9703
	STVD	0.9918	0.9924	0.9806	0.9916	0.9923	0.9861	0.9891	0.9806
	WeatherGFM	0.9905	0.9828	0.9792	0.9789	0.9917	0.9835	0.9844	0.9789
	DriveGuard-WM	**0.9971**	**0.9962**	**0.9925**	**0.9958**	**0.9974**	**0.9938**	**0.9955**	**0.9925**
SSIM ↑	U-Net	0.9231	0.8957	0.8412	0.9104	0.9563	0.9287	0.9092	0.8412
	GeoFAR	0.9528	0.9486	0.9315	0.9502	0.9619	0.9431	0.9480	0.9315
	SGD	0.9642	0.9360	0.9473	0.9448	0.9655	0.9517	0.9516	0.9360
	STVD	0.9657	0.9661	0.9524	0.9653	0.9682	0.9580	0.9626	0.9524
	WeatherGFM	0.9613	0.9504	0.9488	0.9466	0.9651	0.9532	0.9542	0.9466
	DriveGuard-WM	**0.9846**	**0.9832**	**0.9758**	**0.9827**	**0.9873**	**0.9794**	**0.9822**	**0.9758**

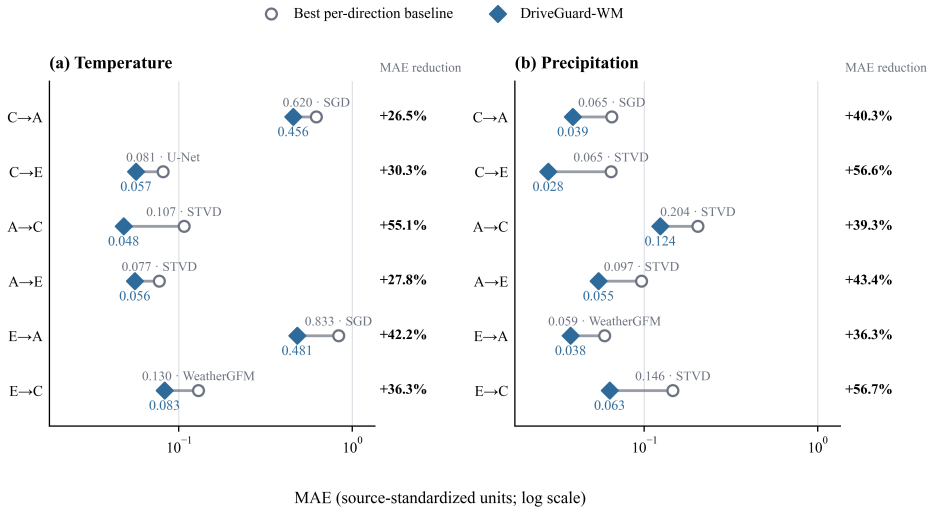


Figure A2: Six directed transfers. Each marker compares DriveGuard-WM with the strongest baseline for that direction. These directional improvements use different comparators from the fixed-baseline macro improvements in the text.

B.2 World Model for Contribution Assessment (WMCA)

Figure 3 reports the complete posterior–prior discrepancies and reconstruction changes after state replacement; positive Δ_Z indicates lower MAE after replacing state Z with its HR-conditioned counterpart. M has the largest KL in all regions, while G gives the largest benefit in CONUS and M in EU; AUS changes are close to zero. These records motivate task-error-based assessment, but do not establish M as the dominant bottleneck or validate the refined single-message intervention.

Content-aligned HR correction (pending). Freeze WMCA and running statistics. Under identical prior histories, global states, and random draws, replace a single message mean and recompute the downstream local states and decoder. Compare identity replacement, aligned HR correction,

shuffled HR content, and equal-magnitude random perturbation. Report the distribution of b_{kn}, positive-benefit fraction, and mean held-out reconstruction change. Identity replacement should yield zero benefit to numerical precision. The controls separate implementation consistency, content alignment, and effective fine-scale information.

Contribution assessment and hierarchical states (pending). Construct benefit-based, KL-based, and random relationship rankings on diagnostic time blocks, freeze them, and compare equal numbers of message corrections on independent held-out blocks. Do not rank and validate on the same known benefits. This is an offline HR-assisted diagnostic, separate from LR-only deployment. To test the state factorization, retrain variants that replace the Global GRU with a projection of the current context, replace pairwise bilinear messages with a shared broadcast projection, or replace the Local GRU with a feed-forward network that combines observations and messages. Preserve output dimensions and training budgets, report parameter counts and floating-point operation counts, and evaluate transfer of the sequence prediction $\widehat{y}_A$ first. These tests concern model-internal reconstruction relations, not identified physical causal processes.

B.3 LR-Conditioned Guidance Learning (LGL)

Table A5 preserves all guidance controls using the functional terminology of Method. LGL Drive guidance improves over spatial reconstruction without Drive by 6.9%, 5.9%, and 15.2% in CONUS, AUS, and EU, and also outperforms random Drive and free fusion.

Names of recorded controls. The WMCA sequence path produces $\widehat{y}_A$, and the ASR spatial path produces $\widehat{y}_B$. The recorded "Current local encoder" uses only the local encoding from the current time step. In the original records, "Path A only" refers to the WMCA sequence path, "Path B only" to the ASR spatial path, and "Unguided Path B" to spatial reconstruction without Drive. "Full Driver Coupling" corresponds to LGL Drive guidance, "Random driving" to random Drive guidance, and "Free fusion" to a learned fusion gate without structural constraints on Drive and Guard. These names describe how historical controls relate to the Method interfaces. Removing Drive removes learned relationship priorities; random Drive replaces those priorities with random relation scores; and LGL Drive uses predicted $\widehat{D}$ to form routing guidance S and Drive-weighted messages U. Exact implementations remain tied to the corresponding run configurations.

The Guidance Predictor matches the HR-informed reference ranking from Contribution Assessment with Spearman correlations of 0.999 in CONUS, 0.944 in AUS, and 0.998 in EU. The overlap among the top 10% of ranked relationships is 0.94, 0.82, and 0.96 in those regions, respectively. These values measure fitting to the reference targets. Because the recorded relation targets are fixed source matrices, they do not validate guidance that varies with each sample. Drive/Guard correlation describes joint variation, not benefit or Guard calibration. Positive Drive and high Guard can coexist because they preserve distinct evidence about possible improvement and unreliability.

LR-conditioned guidance prediction and content-aware spatial mapping (pending). Hold the upstream WMCA fixed and train each spatial path $\widehat{y}_B$ with identical budgets. Compare uniform Drive, fixed source-average scores for Drive and Guard, shuffled correspondence between scores and content, Drive scores derived from KL divergence, direct target supervision without the reference head, and full LGL. First report transfer of the spatial path alone, then transfer of the complete system, to distinguish guidance utility from its interaction with fusion. Uniform Drive retains the Guard head. The fixed-score control retains the current assignment A and message states M so that it isolates whether the scores themselves vary with the input. The KL control replaces only the positive-benefit term with a source-scaled discrepancy score, keeping the other target terms and training unchanged. A control without stability factors removes Drive's variance attenuation and sign-consistency terms. Report Drive and Guard fitting errors separately from their associations with actual reconstruction gains and excess error in the spatial path. A varying spatial guidance map does not by itself establish input-conditioned scores: a varying assignment A or message state M can produce such maps even when the scores are fixed.

Table A5: Complete recorded guidance controls on held-out source data. Functional labels follow Method; original record names are mapped in the text. These controls are not refined-module ablations.

Method-aligned variant	CONUS	AUS	EU
WMCA sequence path $\widehat{y}_A$	0.5641	0.1216	0.3129
Current local encoder	2.1810	0.8602	0.9524
ASR spatial path $\widehat{y}_B$	0.5189	0.2176	0.5368
Spatial reconstruction (without Drive)	0.4907	0.1533	0.3560
Free fusion (without Drive/Guard constraints)	0.4902	0.1535	0.3552
Random Drive guidance	0.5191	0.2168	0.5373
LGL Drive guidance (ours)	**0.4567**	**0.1443**	**0.3019**

Table A6: Guidance Predictor fidelity and score statistics. Rank agreement concerns Drive relative to the HR-informed reference in Contribution Assessment, not reconstruction benefit or Guard calibration.

Region	Spearman ρ	Top-10% overlap	Mean Drive	Mean Guard	$r(\text{Drive}, \text{Guard})$
CONUS	0.999	0.94	0.415	0.496	0.801
AUS	0.944	0.82	0.432	0.441	-0.998
EU	0.998	0.96	0.547	0.584	0.935

B.4 Adaptive Spatiotemporal Reconstruction (ASR)

Table A7 compares the current frame with frame history at coarsened and full LR resolution; Figure A3 visualizes how the two choices interact. Using the current frame at full LR resolution gives the lowest MAE in CONUS and EU, while using frame history at full LR resolution is best in AUS. Compared with the current frame at coarsened LR resolution, using it at full LR resolution reduces MAE by 8.5% in CONUS, 2.2% in AUS, and 9.9% in EU. This regional variation motivates adaptive information use; the input controls alone do not establish the performance of a particular adaptive architecture.

Adaptive feature aggregation and windowed spatial feature propagation (pending). With WMCA and LGL fixed, retrain variants that average the experts uniformly, continue spatial scans across windows instead of resetting them, or remove modulation by $\bar{G}$, $\bar{L}$, and U while retaining the local-state decoder skip. Hold data, seeds, and budgets constant, and evaluate transfer of the spatial prediction $\widehat{y}_B$ first. These tests isolate content-dependent aggregation, propagation extent, and sequence conditioning.

Drive/Guard fusion. Table A8 retains the independent fusion records. Relative to the WMCA sequence path $\widehat{y}_A$, fusion with both Drive and Guard reduces MAE by **32.3%** in CONUS, **15.8%** in AUS, and **89.8%** in EU. It achieves the lowest MAE among the recorded controls in every region. Equal-weight fusion, gates using only Drive or only Guard, and an unconstrained gate have different results across regions. The combined gate improves on the sequence path throughout, supporting the joint use of estimated contribution and unreliability.

The control labels follow the Method notation. The WMCA row reports only the sequence prediction $\widehat{y}_A$, whereas the ASR row reports only the spatial prediction $\widehat{y}_B$. Fixed fusion sets the spatial-path weight to $g = 0.5$. The Drive-only gate removes the Guard term from Eq. (10), and the Guard-only gate removes the Drive term. The unconstrained gate learns g directly without requiring $a, b > 0$. The full gate combines a positive Drive contribution with a negative Guard contribution.

Guard prediction and amplification response. In the separate calibration record of Table A9, the area under the receiver operating characteristic curve (AUROC) for Guard is 0.177 in CONUS, 0.313 in AUS, and 0.137 in EU. Under the convention that a higher Guard score indicates a more harmful spatial correction, these results do not support reliable ranking of harm. Interpreting the area under the precision-recall curve (AUPRC) also requires the prevalence of harmful corrections. Expected calibration error (ECE) and the Brier score compare the gate with a binary label indicating

Table A7: Recorded temporal/spatial input controls on held-out source data. Full remains LR input.

Region	Current–Coarse	Current–Full	History–Coarse	History–Full
CONUS	0.6142	0.5623	0.6151	0.5734
AUS	0.6032	0.5897	0.5819	0.5642
EU	0.5887	0.5307	0.5737	0.5650

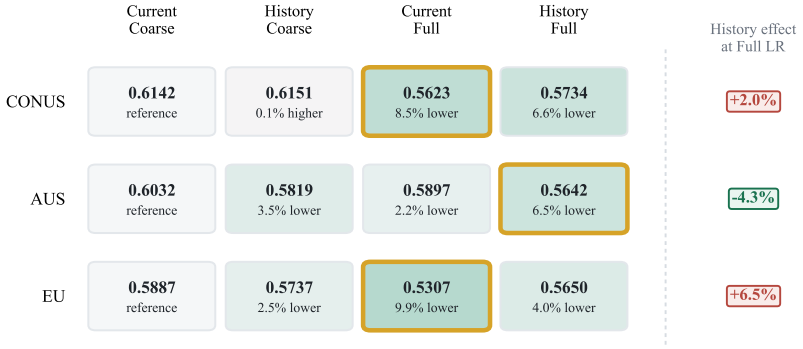


Figure A3: Temporal context × LR-resolution controls on held-out source data. Cells report MAE changes from Current–Coarse; gold borders mark regional minima, and right-hand badges isolate the History effect at Full LR. Full values appear in Table A7.

Table A8: Complete Drive/Guard fusion controls on held-out source data. This record differs from the guidance-control and primary-transfer records.

Fusion strategy	CONUS	AUS	EU
WMCA sequence path $\widehat{y}_A$	0.0520	0.1214	0.2361
ASR spatial path $\widehat{y}_B$	0.4843	0.9098	0.3418
Fixed equal-weight fusion ($g = 0.5$)	0.2485	0.4579	0.1733
Drive-only fusion gate	0.0701	0.1403	0.0464
Guard-only fusion gate	0.1317	0.5177	0.1096
Unconstrained fusion gate (without Drive/Guard)	0.0835	0.2736	0.0644
Drive/Guard gated fusion (ours)	**0.0352**	**0.1022**	**0.0241**

Table A9: Separate Guard and fusion-calibration records. Columns report sequence, spatial, and fused reconstruction errors. Expected calibration error and the Brier score measure agreement with a binary label for the preferred path, rather than calibrated HR uncertainty.

Region	$\widehat{y}_A$ MAE	$\widehat{y}_B$ MAE	$\widehat{y}$ MAE	$r(R, \text{excess})$	AUROC	AUPRC	ECE/Brier	Mean g
CONUS	0.0972	0.4011	0.1228	-0.25	0.177	0.684	0.483 / 0.439	0.320
AUS	0.1214	0.2170	0.1375	0.81	0.313	0.606	0.277 / 0.310	0.430
EU	0.0919	0.5899	0.1220	-0.583	0.137	0.798	0.652 / 0.560	0.256

which path is preferred. They do not establish optimal continuous mixing or calibrated uncertainty in the HR predictions. With other gate inputs and path predictions fixed, amplifying Guard moves fused errors toward the corresponding sequence-path $\widehat{y}_A$ values in all three regions, confirming the gate's conditional fallback response in Table A10.

Matched Drive/Guard fusion evaluation (pending). Reuse the same frozen sequence prediction $\widehat{y}_A$ and spatial prediction $\widehat{y}_B$ for every gate, fitting each gate only on source calibration data. In addition to the recorded variants, compare the best constant gate selected on source data and a gate based only on path disagreement. Match inputs and capacity for constrained and unconstrained gates. Freeze all thresholds and score directions before the six target transfers. Report directional MAE, changes relative to the sequence prediction $\widehat{y}_A$ and the source-selected constant, and the fraction of days improved. Evaluate harm labels $h_i = \mathbf{1}[\ell_B(i) > \ell_A(i)]$ on those same predictions, reporting granularity, score direction, AUROC, AUPRC, and harm prevalence. With other gate

Table A10: Response to amplified Guard with fixed path predictions and other gate inputs. Entries are fused reconstruction MAE.

Guard scaling	CONUS	AUS	EU
WMCA sequence path $\widehat{y}_A$ (reference)	0.0972	0.1214	0.0919
$\times 2$	0.113	0.129	0.1073
$\times 5$	0.099	0.122	0.0931
$\times 10$	0.097	0.121	0.0919
Absolute gap: $\times 10$ vs. $\widehat{y}_A$	**0.0002**	**0.0004**	**0.0000**

inputs fixed, verify separately that increasing Drive raises the spatial-path weight and increasing Guard lowers it. Retraining structure controls and fixed-output fusion controls answer different questions.

B.5 Source-Fit References and Training Records

Source-fit results are auxiliary to the transfer benchmark. Table A11 distinguishes matched temperature comparison from the unmatched precipitation WMCA sequence-path $\widehat{y}_A$ (S1) reference; the latter is not the complete system. Table A12 reports the full matched temperature metrics. The recorded model has lower MAE, while SGD has lower RMSE and higher correlation in all three regions. Source-domain advantages therefore depend on the metric and cannot independently explain transfer gains. Table A13 records selected WMCA training runs. Region-specific validation losses support selection within a region, not ranking regions by task difficulty.

Table A11: Source-fit MAE. The precipitation row for DriveGuard-WM is an unmatched WMCA sequence-path reference (S1 record), not the complete system.

Method	Temp. C	Temp. A	Temp. E	Prec. C	Prec. A	Prec. E
U-Net	0.0550	0.1059	0.0711	0.1141	0.0965	0.0649
SGD	0.0509	0.1025	0.0321	0.1163	0.0655	0.0534
STVD	0.0542	0.1071	0.0350	0.1155	0.0688	0.0511
WeatherGFM	0.0542	0.1155	0.0332	0.1264	0.0674	0.0536
GeoFAR	0.1610	0.2219	0.1440	0.1599	0.0824	0.0745
DriveGuard-WM	0.0375	0.0962	0.0301	0.0950	0.0537	0.0371

Table A12: Matched source-temperature metrics. Better MAE does not imply better RMSE, correlation, or SSIM.

Method	Region	MAE	RMSE	Pearson r	SSIM
U-Net	CONUS	0.0550	0.0992	0.9920	0.9366
	AUS	0.1059	0.1418	0.9907	0.9627
	EU	0.0711	0.1542	0.9732	0.9278
SGD	CONUS	0.0509	0.0916	0.9932	0.9400
	AUS	0.1025	0.1373	0.9913	0.9623
	EU	0.0321	0.0743	0.9938	0.9709
STVD	CONUS	0.0542	0.0976	0.9923	0.9400
	AUS	0.1071	0.1436	0.9905	0.9646
	EU	0.0350	0.0809	0.9926	0.9661
WeatherGFM	CONUS	0.0542	0.0973	0.9924	0.9464
	AUS	0.1155	0.1546	0.9890	0.9512
	EU	0.0332	0.0768	0.9934	0.9721
GeoFAR	CONUS	0.1610	0.2912	0.9329	0.8245
	AUS	0.2219	0.2929	0.9600	0.8375
	EU	0.1440	0.2973	0.8953	0.8867
DriveGuard-WM	CONUS	0.0375	0.1162	0.9919	0.9413
	AUS	0.0962	0.1773	0.9886	0.9463
	EU	0.0301	0.0977	0.9909	0.9653

Table A13: Selected WMCA training records. Validation losses are used for within-region selection only.

Region	Best epoch	Validation loss	Epochs	Minutes
CONUS	80	0.016466	91	481
AUS	35	0.025623	46	374
EU	37	0.017108	48	362

B.6 Qualitative Target-Region Transfer Examples

Figures A4 and A5 show selected target-region predictions in CONUS, AUS, and EU alongside HR references and pixelwise absolute errors. The temperature fields retain broad regional gradients, with residual errors around sharper local transitions. The precipitation fields capture broad event locations, while concentrated rainfall produces localized errors. The displayed map MAE values are case-specific errors in physical units; Tables A3 and A4 provide the six-direction, source-standardized comparisons. *Ours-Widened* and *Ours-S1* identify the plotted variants.

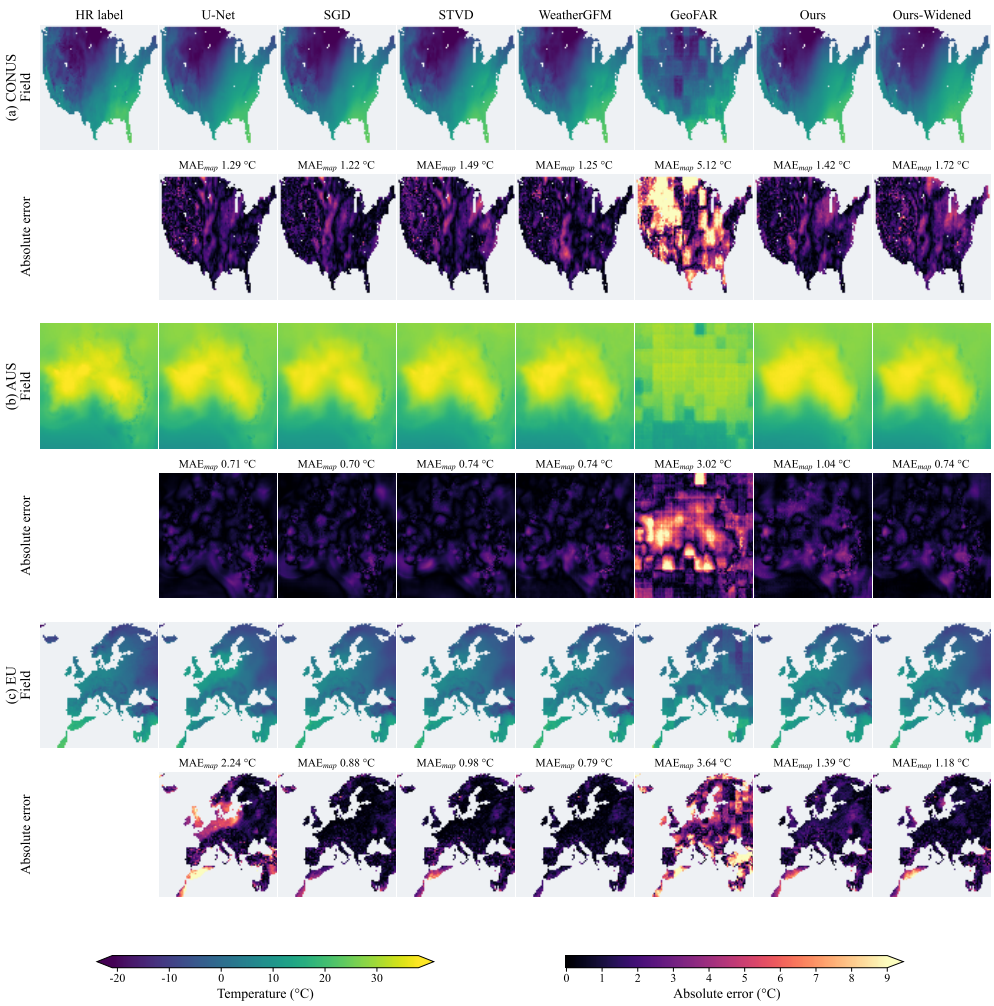


Figure A4: Selected target-region temperature transfer examples in CONUS, AUS, and EU. Each region shows the HR reference and model predictions above pixelwise absolute-error maps. $\mathrm{MAE}_{\mathrm{map}}$ is case-specific and measured in °C; Ours-Widened denotes the plotted widened variant.

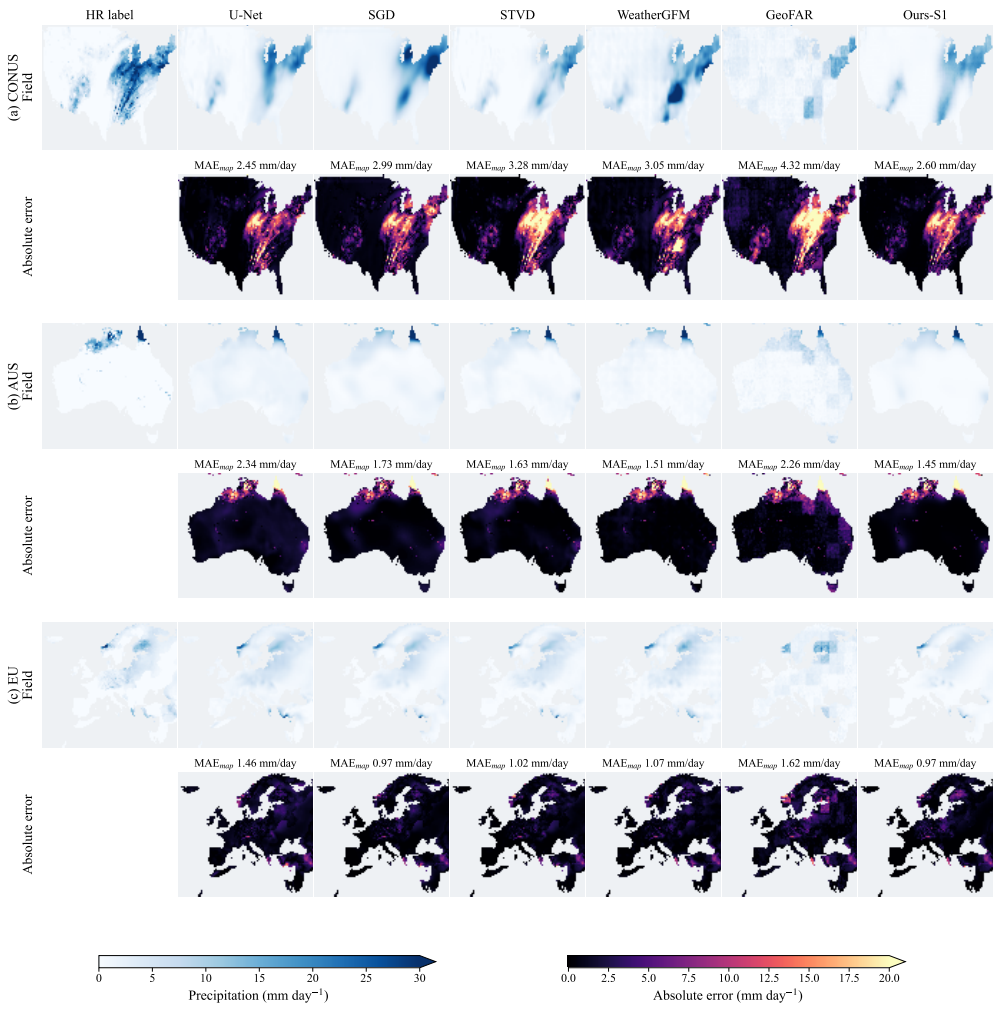


Figure A5: Selected target-region precipitation transfer examples in CONUS, AUS, and EU, with HR references, model predictions, and pixelwise absolute-error maps. $\mathrm{MAE}_{\mathrm{map}}$ is case-specific and measured in mm/day; Ours-S1 denotes the plotted stage-1 sequence-path checkpoint.

C Training Interfaces and Validation Scope

C.1 Learning Stages and Diagnostic Meaning

All learning stages use source data. First fit the LR-prior and HR-posterior reconstruction trajectories with shared parameters and separate histories. Freeze the model for controlled relationship diagnosis; then fit the HR reference head and LR Guidance Predictor, train the spatial path with upstream modules frozen, and finally calibrate the fusion gate on fixed path predictions. Deployment removes the HR Encoder, Gated Correction, and Contribution Assessment with its reference head; it preserves the WMCA prior, Guidance Predictor, spatial path, and fusion with frozen parameters and running statistics.

Content-weighted reconstruction error. Let $\widetilde{A}_n(i)$ be the assignment for query n resized to the HR grid, and let m_i be the HR-valid mask. The local diagnostic uses

$$w_n(i) = \frac{m_i \widetilde{A}_n(i)}{\sum_j m_j \widetilde{A}_n(j) + \epsilon}, \qquad \ell_n(\widehat{y}, y) = \sum_i w_n(i) |\widehat{y}_i - y_i|. \tag{14}$$

Assignments and masks are fixed between the prior and corrected predictions. Global states, prior history, other messages, and random draws remain identical. Thus b_{kn} measures the change attributable to a specific model-internal correction under those conditions. A positive value does not mean that the meteorological variable is physically causal or that the spatial path must outperform the sequence path.

Guidance and calibration interfaces. Drive and Guard targets share source-derived scaling with their predicted counterparts. Normalized Drive distributions support relative-priority matching, while score regression preserves absolute strength. Source error preferences provide a distinct calibration target for the fusion gate: message-correction gain and spatial-path advantage are not interchangeable. All score normalization and path-preference choices must be fixed before target use.

C.2 Recorded Implementation versus Refined Formulation

The supplied revision distinguishes proposed refinements from the available implementation records. We retain this distinction explicitly:

1. The diagnostic records replace global, message, and local states (G, M, and L) separately. The revised method corrects one message at a time while holding the prior history, global state, and other random variables fixed.
2. The recorded Drive/Guard supervision uses fixed source relation matrices. The revised targets combine current-sample correction gains with source temporal stability.
3. The original spatial path uses four experts and routes each position to the two experts with the highest scores. The revised mixture of experts assigns a soft weight to every expert at each position.

Consequently, the benchmark gains of 38.6% for temperature and 46.0% for precipitation remain complete-system results from the recorded pipeline. They must not be attributed to these refinements without matching the run version and re-evaluating the affected components. For every component record, reproduction requires the predicted variable, dates, units, aggregation granularity, random seed, checkpoint identifier, and retraining status. Different record families must not be combined to estimate a component contribution.

C.3 Remaining Evaluation Plan

Limited source HR supervision. Keep source LR dates and target test sets fixed while varying source training HR availability at 25%, 50%, and 100%, sampled in temporal blocks. All HR-dependent stages, including diagnostics and the reference head, must obey the same budget. Fix and separately disclose the labels used for validation and calibration. Report macro transfer MAE against HR availability.

Transfer field visualizations. Figures A4 and A5 provide selected transfer examples. A fuller visualization study should preselect target dates, identify each source-to-target direction and checkpoint, and compare interpolated LR, original-valid HR reference, the strongest baseline, both paths, the complete model, and error maps with shared scales. Show Drive d, Guard r, and fusion weights alongside whole fields and local crops, including cases with weak fusion.

Uncertainty, efficiency, and scoring masks. Repeat the main benchmark and key controls with at least three seeds and report the mean and standard deviation. Use paired date-block resampling to preserve temporal dependence. Report training cost by stage and inference cost for retained modules only, specifying grids, hardware, precision, and batch size. Recompute costs for the dense MoE. Add original-HR-valid-only scoring and physical-unit errors, and document the contribution of ERA5-filled reference cells.

Interpretation. These are specified validation procedures without new numerical results. The present evidence covers three regions and one evaluation year. Further geographic and temporal coverage, precipitation extremes, and implementation-matched controls are needed for stronger robustness or component-attribution claims.

D Extended Related Work and Positioning

This appendix positions DriveGuard-WM relative to meteorological downscaling, latent dynamics, domain generalization, and privileged-information distillation. Section 3 defines the single-source, dynamic-LR-only protocol; here we distinguish these established building blocks from the proposed use of replacement benefits as relation supervision.

D.1 Meteorological Downscaling and Geographic Generalization

Classical statistical downscaling relates coarse atmospheric predictors to local climate variables through regression, weather typing, analog methods, and bias correction (Wilby & Wigley, 1997; Maraun et al., 2010). Neural approaches recast gridded downscaling as super-resolution: DeepSD introduced stacked super-resolution networks for climate projections (Vandal et al., 2017), followed by architecture and intercomparison studies for temperature and precipitation (Vandal et al., 2019; Baño-Medina et al., 2020; Sha et al., 2020a;b). Multi-scale and residual designs further improved regional reconstruction and, in some cases, enabled direct cross-region application (Liu et al., 2020; Wang et al., 2021b). Continental and climate-model applications subsequently established deep statistical downscaling at larger spatial scales (Baño-Medina et al., 2021; 2022; Soares et al., 2024).

Because the LR-to-HR mapping is generally non-unique, stochastic and generative models have been developed to represent uncertainty and fine-scale variability. Representative directions include recurrent GANs, diffusion-based reconstruction, scale-adaptive generation, residual corrective diffusion, and spatiotemporally coherent climate emulation (Leinonen et al., 2021; Ling et al., 2024; Rampal et al., 2025; Hess et al., 2025; Mardani et al., 2025; Glawion et al., 2025; Schmidt et al., 2025; Lopez-Gomez et al., 2025). These models improve realism and ensemble generation, but a flexible decoder can still learn source-specific conditional structure; generative fidelity therefore does not by itself guarantee geographic robustness.

Recent studies make this limitation explicit. Downscalers can deteriorate under changes in climate state, driving GCM, or geographic region, even when their in-domain accuracy is strong (Hernanz et al., 2024; Baño-Medina et al., 2024; Rampal et al., 2024). RainShift directly benchmarks precipitation downscaling under geographic distribution shift and finds substantial region-dependent degradation (Harder et al., 2026). Target-side alignment, few-shot fine-tuning, and regional pretraining can mitigate such errors under their respective data-access assumptions (Harder et al., 2026; Zhu & Zhou, 2024; González-Abad et al., 2026); retraining-free generative approaches instead seek broader priors or inference-time guidance (Hess et al., 2025; Tie et al., 2026). Our setting is complementary but more restrictive at deployment: only dynamic target-region LR fields are available. We therefore emphasize which source-supervised information can be distilled into an LR-only predictor, rather than proposing another reconstruction or generative backbone.

D.2 World Models and Latent Dynamics

World models learn compact latent states whose dynamics can be predicted forward. Early formulations separated representation learning, dynamics, and decision making (Ha & Schmidhuber, 2018); PlaNet introduced recurrent latent state-space planning (Hafner et al., 2019), and Dreamer used latent imagination for behavior learning (Hafner et al., 2020; 2021; 2025). Later work extended latent world models through transformer dynamics, physical-robot learning, scalable continuous control, autonomous driving, and interactive video generation (Micheli et al., 2023; Wu et al., 2023; Hansen et al., 2024; Hu et al., 2023; Bruce et al., 2024). Across these settings, two ideas are especially relevant here: temporal history can be summarized in a predictive state, and a predictive prior can be contrasted with an observation-informed posterior.

We use the prior–posterior structure to compare information access within a learned predictor. The hierarchy separates global process state G_t, cross-scale message M_t, and local state L_t. During source training, the posterior additionally conditions on HR; replacing a prior state or relation and recomputing downstream states measures the predictive effect of that particular change. The result depends on posterior quality and the recomputation operator. It is a model-conditional sensitivity, with geographic transfer evaluated separately; the HR posterior is removed at inference.

The hierarchy provides a relation space in which mismatch and predictive effect can be measured separately. Aggregating replacement effects across source temporal blocks yields relation supervision for an LR head. This makes the world model an interface for constructing training targets as well as a sequence predictor. The quality of those targets depends on the learned posterior and replacement operator; geographic transfer remains a separate empirical property.

D.3 Positioning Relative to Existing Paradigms

Downscalers supply coarse-to-fine predictors, world models supply temporal states and prior–posterior structure, and domain generalization frames evaluation under unseen distributions. Privileged-information distillation is the closest paradigm for the training-to-deployment information flow. Generalized distillation trains a teacher with an additional representation and transfers its predictions to a student using the deployment representation (Lopez-Paz et al., 2016). Thus, an HR-informed teacher paired with an LR-only student follows an established principle.

Our design uses source HR to construct supervision from the predictive effect of controlled latent replacements. These effects are summarized over source temporal blocks and attached to cross-scale relations before being distilled into the LR head. The relevant distinction is therefore the construction and use of relation targets, rather than access to privileged training information itself. Testing its value requires comparison with conventional output or feature distillation, direct LR learning of the same targets, and fixed relation matrices.

The reported transfer benchmark evaluates the complete DriveGuard-WM system. Separate source ablations favor relation guidance over selected unguided alternatives, while capacity controls and routing failures limit stronger attribution. Guard is a fusion-control signal whose sign-constrained term provides conditional fallback; it does not constitute calibrated uncertainty or guarantee protection against negative transfer. Common-protocol cross-region controls are needed to determine whether replacement-based supervision explains the overall gains and whether adaptive fusion improves on simpler alternatives.

D.4 Transfer Learning and Out-of-Distribution Generalization

Transfer learning studies how knowledge from one task or domain supports another (Pan & Yang, 2010; Yosinski et al., 2014). Domain adaptation assumes access to target samples, which may be used for adversarial alignment or other distribution-matching objectives (Ganin et al., 2016; Kouw & Loog, 2021); test-time adaptation further updates a pretrained model during deployment (Wang et al., 2021a). These assumptions do not hold in our protocol: target batches cannot update parameters, normalization statistics, alignment transforms, or routing calibration. Such methods remain appropriate when historical target observations are available, but they address a different deployment regime.

Domain generalization (DG) removes target access during training and seeks predictors that remain effective on unseen distributions (Zhou et al., 2023). Representative approaches pursue invariance, marginal transfer, flat minima, or gradient-variance alignment (Arjovsky et al., 2019; Blanchard et al., 2021; Cha et al., 2021; Rame et al., 2022). Controlled benchmarks nevertheless show that model selection and protocol design can dominate apparent gains, and that naturally occurring shifts remain difficult (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021). Our setting is stricter than standard multi-source DG because each model has only one labeled geographic source. Temporal source partitions do not substitute for geographic diversity; they only test whether relation-level intervention effects remain stable under source-time variation.

Climate downscaling instantiates distribution shift through geographic, topographic, climatic, and data-source changes. Prior work spans direct cross-region transfer, analyses of extrapolation failure, target-data fine-tuning, and zero-shot alignment or generative guidance (Wang et al., 2021b; Hernanz et al., 2024; Baño-Medina et al., 2024; Rampal et al., 2024; Harder et al., 2026; Hess et al., 2025; Tie et al., 2026). Our protocol is specifically *inductive*, *single-source*, and *dynamic-LR-only*. Accordingly, the model cannot learn geographic invariance from multiple labeled regions or rely on target adaptation. It instead uses source HR during training to identify useful latent and cross-scale information, distills that evidence into an LR-only student, and retains a sequence-based reference path for fallback.