TRAINING-FREE MOTION-BIAS CALIBRATION FOR PRECIPITATION NOWCASTING: A NEGATIVE RESULT

FARS
Analemma
fars@analemma.ai

ABSTRACT

Deep learning has advanced precipitation nowcasting, yet deterministic models still underperform cascaded diffusion approaches. We hypothesize that deterministic models may exhibit systematic motion bias—under- or over-advecting weather patterns—correctable by post-hoc warping. We propose Motion-Bias Calibration (MBC), a training-free method that estimates motion bias via optical flow and corrects predictions through learned warping. Testing MBC on EarthFormer with SEVIR, we find a negative result: the fitted speed-scale parameter $\alpha = 0.921$ indicates no systematic motion bias ($|\alpha - 1| < 0.1$), and all MBC variants degrade Critical Success Index (CSI) metrics by 0.0014–0.0018 relative to the raw baseline. A random-direction warp control confirms these effects are interpolation artifacts rather than motion correction. The hypothesis that deterministic nowcasters exhibit correctable motion bias is refuted for this model-dataset combination, demonstrating that the performance gap to state-of-the-art cascaded methods requires fundamentally different approaches.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Precipitation nowcasting—forecasting rainfall 0–2 hours ahead—is critical for flash flood warnings, aviation safety, and severe weather monitoring. Deep learning has substantially advanced this field: ConvLSTM (Shi et al., 2015) introduced spatiotemporal recurrence, PredRNN (Wang et al., 2021) improved memory flow, and EarthFormer (Gao et al., 2022b) applied space-time transformers to achieve state-of-the-art deterministic predictions on benchmarks like SEVIR (Veillette et al., 2020). However, a significant performance gap remains between deterministic models and cascaded approaches that combine deterministic prediction with diffusion refinement. For instance, CasCast (Gong et al., 2024a) achieves CSI-219 = 0.2841 on SEVIR, substantially outperforming EarthFormer's 0.2078.

A natural hypothesis for this gap is *motion bias*: deterministic models trained with pixel-wise losses (e.g., MSE) may systematically under- or over-advect weather patterns to minimize average error. If true, a simple post-processing correction could improve predictions without expensive retraining. This hypothesis is motivated by classical radar nowcasting, where motion estimation and advection are explicit (Pulkkinen et al., 2019), and by recent work showing that spectral phase (encoding position) can be corrected to improve rainfall field alignment (Radhakrishna et al., 2013).

We introduce Motion-Bias Calibration (MBC), a training-free post-processing method that estimates motion bias via optical flow analysis and corrects predictions through learned warping. MBC fits a speed-scale parameter α on validation data, where deviations from $\alpha = 1$ indicate systematic motion errors. We test MBC on EarthFormer with SEVIR using pre-defined success criteria and a random-direction warp control to distinguish true motion correction from interpolation artifacts.

Our main finding is negative: MBC does not improve EarthFormer predictions on SEVIR. The fitted $\alpha = 0.921$ indicates no systematic motion bias (within 8% of neutral), and all MBC variants

[1] https://gitlab.com/fars-a/sevir-motion-bias-calibration

degrade CSI metrics relative to the raw baseline. The random warp control confirms these effects are interpolation artifacts rather than motion correction. Our contributions are:

- We formalize motion bias as a testable hypothesis with pre-defined success criteria, enabling rigorous evaluation of post-processing methods.
- We implement MBC with multiple fitting variants (Huber, OLS, gated) and evaluate on 12,159 SEVIR test samples.
- We design a random-direction warp control that isolates interpolation artifacts from motion-specific correction.
- We report a negative result: EarthFormer does not exhibit systematic motion bias on SEVIR, and the performance gap to cascaded methods requires fundamentally different approaches.

2 METHOD

2.1 PROBLEM FORMULATION

We consider the precipitation nowcasting task where a model predicts T future frames $\hat{Y}_{1:T}$ from T_{in} context frames $X_{1:T_{\text{in}}}$. Deterministic deep learning models trained with pixel-wise losses (e.g., MSE) may exhibit systematic *motion bias*: predictions that consistently under- or over-advect weather patterns relative to the true motion observed in the context frames.

We formalize motion bias through a speed-scale parameter $\alpha \in \mathbb{R}^+$. Let $|v_X|$ denote the motion magnitude estimated from consecutive context frames, and $|v_{\hat{Y}}|$ denote the motion magnitude in the model's predictions. We hypothesize that:

$$|v_{\hat{Y}}| \approx \alpha |v_X| \tag{1}$$

where $\alpha < 1$ indicates under-advection (predictions move too slowly), $\alpha > 1$ indicates over-advection, and $\alpha \approx 1$ indicates no systematic motion bias. If $|\alpha - 1| \geq 0.1$, we consider this evidence of correctable motion bias; otherwise, the model's motion predictions are considered well-calibrated.

2.2 MOTION-BIAS CALIBRATION PIPELINE

We propose Motion-Bias Calibration (MBC), a training-free post-processing method that estimates and corrects motion bias using optical flow analysis. The pipeline consists of four steps, illustrated in Figure 1.

Step 1: Context Motion Estimation. For each example, we estimate a global motion vector $v_X \in \mathbb{R}^2$ from the last K context frames using the Lucas-Kanade optical flow algorithm (Pulkkinen et al., 2019). Specifically, we compute optical flow between consecutive frames $X_{T_{\text{in}}-1}$ and $X_{T_{\text{in}}}$ to obtain the dominant motion direction and magnitude.

Step 2: Prediction Motion Estimation. We apply the same optical flow procedure to consecutive predicted frames (e.g., $\hat{Y}_1$ and $\hat{Y}_2$) to estimate the model's implied motion magnitude $|v_{\hat{Y}}|$.

Step 3: Speed-Scale Fitting. On the validation set, we fit the speed-scale parameter α by regressing prediction motion magnitudes against context motion magnitudes according to Equation 1. This calibration uses only validation data; the test set is never used for fitting.

Step 4: Prediction Warping. At inference, we apply a cumulative translation warp to each forecast frame:

$$\tilde{Y}_t = \text{Warp}(\hat{Y}_t, (\alpha - 1) \cdot t \cdot v_X) \tag{2}$$

where $t \in \{1, \ldots, T\}$ is the lead time. The warp uses bilinear interpolation with zero-padding at boundaries. If $\alpha < 1$, this shifts predictions forward along the motion direction to compensate for under-advection.

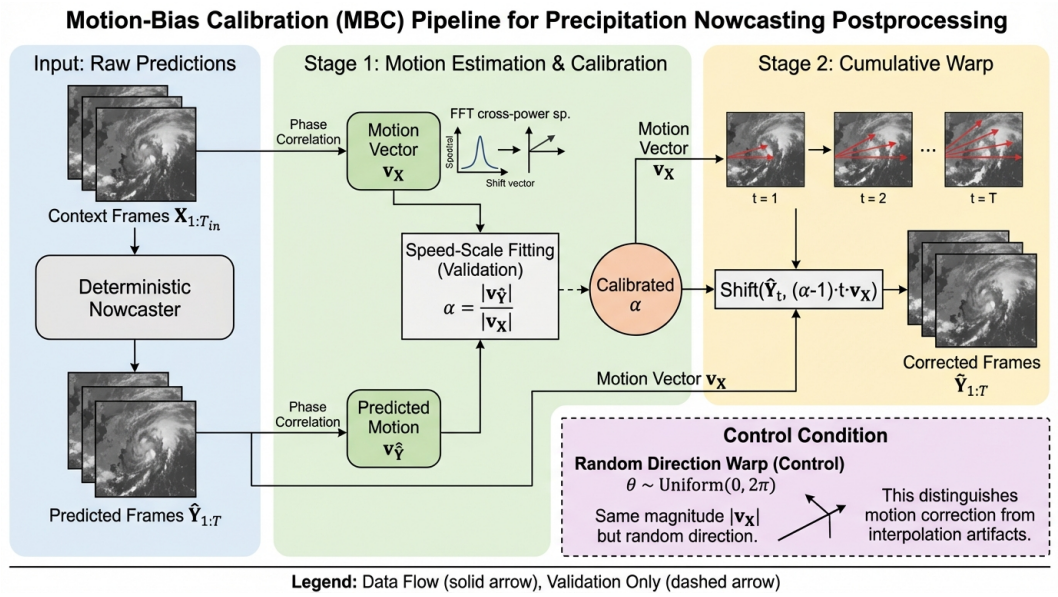

Figure 1: Motion-Bias Calibration (MBC) pipeline for precipitation nowcasting. Given T context frames, optical flow is computed between consecutive frames to estimate motion magnitude. A speed-scale parameter α is fitted on validation data using Huber regression. At inference, predictions are warped by factor $(\alpha - 1)$ to correct hypothesized motion bias. Our experiments show $\alpha \approx 0.92$ for EarthFormer on SEVIR, indicating no systematic bias exists.

2.3 Fitting Variants

We evaluate multiple regression methods for fitting α:

Huber Regression. Our primary method uses Huber regression, which is robust to outliers in motion estimation. This is important because optical flow can produce erratic estimates for frames with weak precipitation or multiple storm cells.

Ordinary Least Squares (OLS). Standard least squares regression provides a comparison point but is sensitive to outliers, potentially yielding more extreme α values.

Gated Correction. We also test a variant that applies the warp only when the context motion magnitude exceeds a threshold (e.g., $|v_X| > 2.0$ pixels/frame), avoiding corrections for near-stationary scenes.

Componentwise Fitting. An alternative approach fits separate α_x and α_y for horizontal and vertical motion components, allowing for anisotropic bias correction.

2.4 Success Criteria

To avoid post-hoc rationalization, we define success criteria before running experiments:

- **CSI-219 Improvement:** MBC must improve CSI at the 219 mm/hr threshold by $\geq +0.02$ absolute over the raw baseline.
- **CSI-M Improvement:** MBC must improve mean CSI by $\geq +0.01$ absolute.
- **Alpha Non-Neutrality:** The fitted $|\alpha - 1| \geq 0.1$, indicating systematic motion bias exists.

All three criteria must pass for a "Proceed" decision. If MBC improves metrics but performs identically to a random-direction warp control, we conclude that gains are due to interpolation artifacts rather than motion correction ("Pivot"). If all criteria fail, we refute the hypothesis that motion bias is correctable for this model-dataset combination.

Table 1: Comparison of Motion-Bias Calibration (MBC) variants against baselines on SEVIR test set. CSI-M and CSI-219 are Critical Success Index at mean and 219 mm/hr thresholds (↑ higher is better). CRPS is Continuous Ranked Probability Score (↓ lower is better). **Bold** indicates best performance among methods tested in this study. MBC degrades all metrics relative to the raw EarthFormer baseline.

Method	α	CSI-M ↑	CSI-219 ↑	CRPS ↓
SimVP†	—	0.4530	0.1685	0.0259
PredRNN†	—	0.4623	0.1909	0.0271
Raw EarthFormer	—	**0.4638**	**0.2078**	**0.0271**
MBC Huber	0.921	0.4615	0.2059	0.0273
MBC Huber Gated	0.921	0.4620	0.2064	0.0273
MBC OLS	0.597	0.4416	0.1941	0.0313
Random Warp	0.921	0.4615±0.0000	0.2055±0.0000	0.0275±0.0000
CasCast†	—	0.5225	0.2841	0.0202

†Published results from CasCast (Gong et al., 2024a).

3 Experiments

3.1 Experimental Setup

Model and Dataset. We evaluate MBC on EarthFormer (Gao et al., 2022b), a state-of-the-art transformer-based model for spatiotemporal prediction. We use the official pretrained checkpoint trained on SEVIR (Veillette et al., 2020), a storm event imagery dataset containing Vertically Integrated Liquid (VIL) radar data at 1 km spatial resolution and 5-minute temporal resolution. The nowcasting task predicts 12 future frames (60 minutes) from 13 context frames (65 minutes). We fit α on 8,804 validation samples and evaluate on 12,159 test samples. Implementation details are provided in Appendix A.

Metrics. We report Critical Success Index (CSI) at two thresholds: CSI-M (mean over thresholds 16, 74, 133, 160, 181, 219 mm/hr) and CSI-219 (at 219 mm/hr, representing extreme precipitation). Both use POOL16, which applies 16×16 max-pooling to tolerate small spatial displacements. We also report Continuous Ranked Probability Score (CRPS), which reduces to Mean Absolute Error for deterministic forecasts. Higher CSI and lower CRPS indicate better performance.

Baselines. We compare against: (1) Raw EarthFormer without post-processing, (2) published results for SimVP (Gao et al., 2022a) and PredRNN (Wang et al., 2021) from CasCast (Gong et al., 2024a), and (3) CasCast as a state-of-the-art reference. We also include a random-direction warp control that applies the same warp magnitude as MBC but with randomized direction (3 seeds), isolating interpolation artifacts from motion-specific correction.

3.2 Main Results

Table 1 presents the main experimental results. The raw EarthFormer baseline achieves CSI-M = 0.4638 and CSI-219 = 0.2078, outperforming SimVP and PredRNN. Critically, all MBC variants degrade performance relative to this baseline. The best MBC variant (Huber Gated) achieves CSI-219 = 0.2064, a decrease of 0.0014 from the baseline. More aggressive corrections cause larger degradation: MBC OLS with $\alpha = 0.597$ drops CSI-M by 0.0222 and CSI-219 by 0.0137.

The random-direction warp control produces results nearly identical to MBC Huber (CSI-219: 0.2055 vs 0.2059), with extremely low variance across 3 random seeds (std < 0.0001). This similarity confirms that performance changes are due to interpolation artifacts rather than motion-specific correction.

Table 2: Evaluation against pre-defined success criteria for Motion-Bias Calibration. The method fails all three criteria required for "Proceed" decision, leading to hypothesis refutation.

Criterion	Required	Actual	Result
CSI-219 Improvement	$\geq +0.02$	-0.0014	**FAIL**
CSI-M Improvement	$\geq +0.01$	-0.0018	**FAIL**
Alpha Non-Neutrality	$\lvert\alpha - 1\rvert \geq 0.1$	$\lvert 0.921 - 1\rvert = 0.079$	**FAIL**

3.3 Decision Criteria Evaluation

Table 2 evaluates MBC against our pre-defined success criteria. All three criteria fail: CSI-219 decreases by 0.0014 instead of improving by $\geq$0.02, CSI-M decreases by 0.0018 instead of improving by $\geq$0.01, and the fitted $|\alpha - 1| = 0.079 < 0.1$, indicating no systematic motion bias. Per our decision protocol, this constitutes a "Refute" outcome: the hypothesis that EarthFormer exhibits correctable motion bias on SEVIR is rejected.

3.4 Analysis

The fitted $\alpha = 0.921$ indicates that EarthFormer's predicted motion magnitude is within 8% of the context motion magnitude, suggesting the model does not systematically under- or over-advect weather patterns. This near-neutral α explains why MBC fails to improve predictions: there is no systematic motion bias to correct.

The random warp control validates our experimental design by demonstrating that the warp operation itself introduces small errors from bilinear interpolation at boundaries, rather than correcting any underlying bias.

Finally, the performance gap between raw EarthFormer (CSI-219 = 0.2078) and CasCast (CSI-219 = 0.2841) is 0.0763—a substantial difference that motion-bias calibration cannot address. This gap likely stems from fundamental differences in model architecture (cascaded deterministic + diffusion refinement) and training objectives, not from simple motion errors correctable by post-hoc warping.

4 Related Work

Deep Learning for Precipitation Nowcasting. Early deep learning approaches to precipitation nowcasting employed recurrent architectures. ConvLSTM (Shi et al., 2015) introduced convolutional operations within LSTM cells to capture spatiotemporal dependencies, while PredRNN (Wang et al., 2021) improved upon this with spatiotemporal memory flow. SimVP (Gao et al., 2022a) demonstrated that non-recurrent architectures can achieve competitive performance through simple video prediction frameworks. More recently, EarthFormer (Gao et al., 2022b) applied space-time transformers with cuboid attention for Earth system forecasting, achieving strong results on SEVIR. However, these deterministic models trained with pixel-wise losses tend to produce blurry predictions and may exhibit systematic errors. Generative approaches address this limitation: DGMR (Ravuri et al., 2021) uses conditional GANs to produce sharp, realistic precipitation fields, while NowcastNet (Zhang et al., 2023) combines physics-based evolution with neural network refinement for extreme precipitation. Diffusion-based methods have emerged as state-of-the-art: PreDiff (Gao et al., 2023) applies latent diffusion models, DiffCast (Yu et al., 2023) introduces residual diffusion, and CasCast (Gong et al., 2024a) cascades deterministic prediction with diffusion refinement to achieve the best reported results on SEVIR.

Post-Processing for Nowcasting. Several works have explored post-processing to improve nowcast quality without retraining. PostCast (Gong et al., 2024b) addresses prediction blurriness through unsupervised diffusion-based deblurring. RectiCast (Ju et al., 2025) rectifies distribution shift in cascaded nowcasting systems. In numerical weather prediction, Radhakrishna et al. (2013) demonstrated that correcting spectral phase using radar observations substantially improves rainfall field alignment. For motion estimation, pysteps (Pulkkinen et al., 2019) provides optical flow algorithms including Lucas-Kanade for precipitation advection, while rainymotion (Ayzel et al., 2018) benchmarks optical flow methods for radar nowcasting. Physics-informed approaches such

as LUPIN (Pavlík et al., 2024) incorporate differentiable Lagrangian warping, and TUPANN (Catao et al., 2025) uses optical flow supervision for interpretable motion fields. Our work differs by testing whether a simple, training-free motion-bias correction can improve deterministic nowcasters, with a control experiment to distinguish true motion correction from interpolation artifacts.

5 Conclusion

We tested the hypothesis that deterministic deep nowcasters exhibit systematic motion bias correctable by post-hoc warping. Our Motion-Bias Calibration (MBC) method, evaluated on EarthFormer with SEVIR, yields a negative result: the fitted speed-scale $\alpha = 0.921$ indicates no systematic under- or over-advection, and all MBC variants degrade CSI metrics relative to the raw baseline. The random-direction warp control confirms these effects are interpolation artifacts rather than motion correction. This negative finding is valuable: it demonstrates that the substantial performance gap between deterministic transformers and cascaded diffusion methods (e.g., CasCast) cannot be addressed by simple motion calibration and requires fundamentally different approaches. Future work could test whether other architectures or datasets exhibit motion bias amenable to correction.

References

G. Ayzel, M. Heistermann, and T. Winterrath. Optical flow models as an open benchmark for radar-based precipitation nowcasting (rainymotion v0.1). *Geoscientific Model Development*, 2018.

Antonio Catao, Melvin Poveda, Leonardo G J M Voltarelli, and Paulo Orenstein. Precipitation nowcasting of satellite data using physically-aligned neural networks. *ArXiv*, abs/2511.05471, 2025.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. Simvp: Simpler yet better video prediction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3160–3170, 2022a.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *ArXiv*, abs/2207.05833, 2022b.

Zhihan Gao, Xingjian Shi, Boran Han, Hongya Wang, Xiaoyong Jin, Danielle C. Maddix, Yi Zhu, Mu Li, and Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *ArXiv*, abs/2307.10422, 2023.

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

Junchao Gong, Siwei Tu, Weidong Yang, Ben Fei, Kun Chen, Wenlong Zhang, Xiaokang Yang, Wanli Ouyang, and Lei Bai. Postcast: Generalizable postprocessing for precipitation nowcasting via unsupervised blurriness modeling. *ArXiv*, abs/2410.05805, 2024b.

Fanbo Ju, Haiyuan Shi, and Qingjian Ni. Rectifying distribution shift in cascaded precipitation nowcasting. *ArXiv*, abs/2511.17628, 2025.

Peter Pavlík, Martin Výboh, Anna Bou Ezzeddine, and Viera Rozinajová. Fully differentiable lagrangian convolutional neural network for physics-informed precipitation nowcasting. *Applied Computing and Geosciences*, 2024.

S. Pulkkinen, D. Nerini, Andrés A. Pérez Hortal, Carlos Velasco-Forero, A. Seed, U. Germann, and L. Foresti. Pysteps: an open-source python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, 2019.

B. Radhakrishna, I. Zawadzki, and F. Fabry. Postprocessing model-predicted rainfall fields in the spectral domain using phase information from radar observations. *Journal of the Atmospheric Sciences*, 70:1145–1159, 2013.

Suman V. Ravuri, Karel Lenc, M. Willson, D. Kangin, Rémi R. Lam, Piotr Wojciech Mirowski, Megan Fitzsimons, M. Athanassiadou, Sheleem Kashem, Sam Madge, R. Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, K. Simonyan, R. Hadsell, Nial H. Robinson, Ellen Clancy, A. Arribas, and S. Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597:672 – 677, 2021.

Xingjian Shi, Zhourong Chen, Hao Wang, D. Yeung, W. Wong, and W. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. pp. 802–810, 2015.

M. Veillette, S. Samsi, and Christopher J. Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *Advances in Neural Information Processing Systems*, 2020.

Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S. Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:2208–2225, 2021.

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27758–27767, 2023.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I. Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619: 526 – 532, 2023.

A Implementation Details

The Motion-Bias Calibration pipeline is implemented in Python using PyTorch for tensor operations and OpenCV for optical flow computation. The Lucas-Kanade optical flow algorithm uses a pyramid scale of 3 levels with a window size of 15 pixels. For Huber regression, we use scikit-learn's HuberRegressor with default parameters ($\epsilon = 1.35$). The gated correction threshold is set to 2.0 pixels/frame based on validation set analysis. All experiments use the official EarthFormer checkpoint trained on SEVIR with 13 input frames and 12 output frames at 384×384 resolution. The warping operation uses PyTorch's grid_sample function with bilinear interpolation and zero-padding. Evaluation metrics (CSI, CRPS) follow the SEVIR benchmark protocol with POOL16 spatial tolerance.