Evaluating the Performance of AI in Crisis Detection: A Multi-Scenario Hindcast of Extreme Precipitation Forecasts

Feng Huang
School of Safety Science,
Tsinghua University, China
huangf23@outlook.com

Lida Huang
School of Safety Science,
Tsinghua University, China
hld999@yeah.net

Jing Zhang
School of Safety Science,
Tsinghua University, China
zhangjing0@mail.tsinghua.edu.cn

Tao Chen
School of Safety Science,
Tsinghua University, China
chentao.b@tsinghua.edu.cn

Guofeng Su
School of Safety Science,
Tsinghua University, China
Key Laboratory of Investigation on Disaster and Accident,
Ministry of Emergency Management, China
sugf@tsinghua.edu.cn

ABSTRACT

Artificial Intelligence Weather Prediction (AIWP) models excel in global mean-error metrics, yet their efficacy in detecting low-probability, high-impact extreme events--critical for emergency response--remains under-examined. This study evaluates three leading models (GraphCast, FuXi, and Artificial Intelligence Forecasting System (AIFS)) against satellite observations and a numerical baseline across four diverse historical crises. Using a crisis-centric evaluation framework comprising Peak Amplitude Ratio (PAR), Spatial Correlation (SC), Root Mean Square Error (RMSE), volumetric Bias, and the Symmetric Extremal Dependence Index (SEDI), preliminary results reveal a systemic intensity deficit in AIWP models. While GFS maintains a PAR above 0.65 across most scenarios, AI models underestimate peak rainfall by over 90% and exhibit significant spatial displacement. These findings suggest that inherent statistical smoothing transforms catastrophic signals into benign forecasts. Consequently, over-reliance on current AIWP models for crisis detection may yield a false sense of security, potentially exacerbating rather than mitigating emergency vulnerabilities.

Keywords

Crisis Detection, Extreme Precipitation, AI Weather Prediction, Hindcast Evaluation.

INTRODUCTION

The timely detection of emerging threats is a foundational requirement for effective crisis management (Bonev et al., 2025). Despite advances in forecasting technologies, organizations and governments frequently encounter "rude surprises," such as unexpected and devastating extreme weather events (Sun et al., 2025). This recurring failure to anticipate severe disruptions is often characterized as a "detection doom loop," highlighting systemic vulnerabilities in existing early warning infrastructures and organizational sense-making processes.

Recently, Artificial Intelligence Weather Prediction (AIWP) models, including GraphCast, FuXi, and the

Artificial Intelligence Forecasting System (AIFS), have emerged as prominent tools in meteorology (L. Chen et al., 2023; Lam et al., 2022; Lang et al., 2024). Evaluated primarily through global mean-error metrics, such as Root Mean Square Error (RMSE) for 500 hPa geopotential height, these data-driven models have demonstrated statistical superiority over traditional Numerical Weather Prediction (NWP) systems (Xia et al., 2026). This performance has generated significant advertised promises regarding their potential to fundamentally enhance crisis detection and societal resilience (Landsberg & Barnes, 2025).

However, the reliance on global average metrics presents a critical blind spot for crisis management (K. Chen et al. 2023; Bonev et al. 2025). Extreme meteorological events, such as catastrophic localized floods or supercharged cyclones, represent the low-probability, high-impact "long tail" of atmospheric distributions. AIWP models, typically optimized using loss functions like Mean Squared Error (MSE), inherently penalize large, localized variances. Consequently, they exhibit a systemic "smoothing effect," wherein extreme precipitation peaks are statistically di-minished to minimize overall error (K. Chen et al., 2023; Zhong et al., 2024). In the context of crisis detection, accurately predicting the average regional weather while completely missing the extreme disaster peak renders a forecast operationally ineffective.

To address this gap, this study empirically investigates the operational performances of leading AIWP models in detecting extreme precipitation crises. Using a multi-scenario hindcast design, we evaluate the performance of GraphCast, FuXi, and AIFS against satellite-derived ground truth (Global Satellite Mapping of Precipitation (GSMaP)) and a traditional physics-based baseline (Global Forecast System (GFS)). The analysis encompasses four high-impact events from 2024: arid-region flash floods (UAE-Oman), tropical convective rainfall (Tanzania), complex-terrain monsoons (Nepal), and mid-latitude winter storms (UK). To isolate model performance during crises, we employ crisis-centric indicators—specifically the Peak Amplitude Ratio (PAR) and Spatial Correlation (SC)—to quantify intensity deficits and spatial displacements.

By analyzing the specific failure mechanisms of AI models in these severe scenarios, this study critically assesses the gap between technological promises and operational realities. We argue that the statistical smoothing inherent in current AI forecasts may induce a false sense of security among decision-makers, thereby complicating organizational response and inadvertently exacerbating the detection doom loop in disaster governance.

DATA AND METHODS

Case Selection and Observational Data

To rigorously evaluate model performance against the most severe meteorological crises, this study selects four record-breaking extreme precipitation events from 2024 (Green et al., 2025). Following the climatological classification, these cases encompass the primary synoptic drivers of recent global extremes:

Low-pressure systems and blocking: The historic UAE-Oman floods (April 14–17), where an anomalous cut-off low delivered over 240 mm of rainfall in 24 hours to arid regions like Dubai.

Supercharged tropical cyclones: Tropical Cyclone Hidaya (May 4), which brought unprecedented rainfall (>89 mm) to Mtwara, Tanzania, triggering mandatory mass evacuations.

Increased monsoon rains: The Kathmandu, Nepal monsoon surge (September 26–28), which dumped 240 mm of rain in 24 hours over complex mountainous terrain, causing devastating landslides.

Winter storms: The successive January winter storms (Storm Ingunn) in England and Wales, UK, driven by an intensified jet stream bringing widespread 50–100 mm rainfall and severe flooding.

The observational reference for evaluating forecast accuracy is derived from the GSMaP_Gauge v8 reanalysis product, which combines satellite retrievals with rain gauge corrections to provide high-resolution (0.1 degrees) precipitation estimates at hourly temporal resolution. While GSMaP represents one of the most reliable satellite-derived precipitation products available, it is important to acknowledge that satellite retrievals have known limitations--particularly in complex terrain and mountainous regions where orographic effects can introduce systematic biases in precipitation estimation. The Nepal case (complex mountainous terrain) may therefore be subject to higher reference uncertainty. However, since all models are evaluated against the same observational reference, the relative performance comparisons remain valid. For the traditional NWP baseline, we utilize the operational forecast data from the Global Forecast System (denoted as GFS_FX). All datasets are uniformly regridded to a 0.25 degrees x 0.25 degrees spatial resolution to ensure consistency in the grid-to-grid comparison.

AIWP Models and Experimental Setup

All AIWP model experiments in this study were conducted using the NVIDIA Earth2Studio platform, which

provides standardized computational infrastructure and accessible interfaces for running state-of-the-art weather prediction models. The hindcast experiments evaluate three leading data-driven weather prediction models: FuXi, AIFS, and GraphCast. The selection of these three models was deliberate: they represent the most prominent open-source AIWP models capable of directly generating precipitation forecasts. Other available models either do not provide direct precipitation outputs or require additional diagnostic/prognostic models to derive precipitation from primary meteorological variables (such as geopotential height or temperature), introducing additional sources of uncertainty. By focusing on models with native precipitation prediction capabilities, this study ensures a fair and direct comparison of crisis detection performance.

A critical aspect of our experimental design is the standardization of initial conditions. To ensure a controlled and fair comparison across all models, the atmospheric states used to initialize FuXi, AIFS, and GraphCast are strictly sourced from the GFS operational analysis data. By forcing the AI models with the same GFS initial fields, any deviations in the subsequent forecasts can be directly attributed to the internal architectural and algorithmic differences of the models, rather than discrepancies in data assimilation.

The primary prognostic variable analyzed in this study is the 6-hour accumulated precipitation (tp06). For each crisis scenario, multi-lead-time hindcasts were generated to assess the temporal stability and convergence of the predictive signals leading up to the disaster peak.

Crisis-Centric Evaluation Metrics

To comprehensively assess the performance of AIWP models and isolate their specific failure mechanisms during extreme events, this study employs a crisis-centric evaluation framework augmented by conventional meteorological metrics, designed to quantify intensity deficits and spatial displacements. For all subsequent formulations, let N represent the total number of grid points within the target bounding box for a specific case. Let F_i and O_i denote the forecasted (model) and observed (satellite) 6-hour accumulated precipitation at the i-th grid point, respectively.

Conventional and Volumetric Metrics

Conventional metrics evaluate the general grid-to-grid accuracy and the overall water budget of the forecast within the target domain.

RMSE serves as the standard baseline metric and the primary optimization target for the loss functions of the evaluated AI models. It quantifies the average magnitude of the forecast error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}$$

While valuable for assessing average performance, RMSE heavily penalizes spatial displacement (the "double-penalty" problem) and implicitly rewards statistically smoothed, conservative forecasts.

To determine whether intensity errors are driven by a systemic failure to generate precipitation or by spatial smoothing, the total volumetric bias is calculated:

$$\mathrm{BIAS} = \frac{\sum_{i=1}^{N} F_i}{\sum_{i=1}^{N} O_i}$$

where F_i and O_i represent the forecasted and observed 6-hour accumulated precipitation at the i-th grid point, respectively. A BIAS value significantly less than 1.0 indicates a systemic underestimation of total precipitation, whereas a value greater than 1.0 indicates overestimation.

Spatial Reliability

Accurate geographic localization of extreme precipitation is critical for the deployment of emergency resources and evacuation protocols.

SC measures the linear relationship and spatial alignment between the forecasted and observed precipitation patterns. It is calculated using the Pearson correlation coefficient across the flattened grid arrays:

$$\mathrm{SC} = \frac{\sum_{i=1}^{N}(F_i - \bar{F})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N}(F_i - \bar{F})^2 \sum_{i=1}^{N}(O_i - \bar{O})^2}}$$

where $\bar{F}$ and $\bar{O}$ are the spatial means of the forecast and observation, respectively. Values approaching 1.0 indicate perfect spatial alignment, whereas values near or below 0 signify severe spatial displacement, representing a critical vulnerability in crisis warning.

Crisis-Centric Extremal Metrics

To evaluate the operational utility of the forecasts in anticipating "rude surprises," metrics must focus exclusively on the extreme tail of the precipitation distribution.

PAR directly quantifies the intensity deficit at the core of the disaster. It compares the absolute maximum precipitation predicted by the model to the observed maximum within the domain:

$$\mathrm{PAR} = \frac{\max(F_i)}{\max(O_i)}$$

A PAR significantly below 1.0 provides direct evidence of the statistical smoothing effect, indicating the model's inability to capture the destructive magnitude of the event.

Traditional categorical metrics (e.g., Critical Success Index) degenerate toward zero for rare events, making them unsuitable for extreme crises. SEDI is specifically designed to evaluate the detection rate of rare anomalies exceeding a critical threshold (T_{ext}, defined in this study as region-specific thresholds based on climatological characteristics: 10 mm/6h for UAE & Oman and England & Wales, 15 mm/6h for Mtwara, and 20 mm/6h for Kathmandu, derived from the GSMaP_Gauge v8 reanalysis product) (Ferro & Stephenson, 2011). Note that SEDI is here applied to deterministic forecast fields by binary thresholding, and the scores should be interpreted as spatial pattern detection skill rather than probabilistic calibration. Based on a contingency table, let H represent the Hit Rate (probability of detection) and F represent the False Alarm Rate. SEDI is formulated as:

$$\mathrm{SEDI} = \frac{\ln F - \ln H + \ln(1-H) - \ln(1-F)}{\ln F + \ln H + \ln(1-H) + \ln(1-F)}$$

SEDI scores range from -1 to 1. A score approaching 1 indicates perfect extremal dependence, whereas a score near 0 indicates no skill in detecting the extreme threshold, directly reflecting a failure in crisis early warning capability.

RESULT

The 72-Hour Benchmark: Intensity Deficit at the Golden Window

To evaluate the operational utility of the models for early warning systems, this analysis anchors on a 72-hour lead time (approximately 3 days prior to the event). In disaster governance, this temporal window constitutes the critical threshold for proactive resource deployment and mass evacuation. Table 1 presents the performance metrics of the four models across the selected extreme precipitation scenarios at this specific lead time. PAR and RMSE are computed over all forecast lead times (0-168 h); SC, BIAS, and SEDI are aggregated across all grid points and time steps. PAR, SC, and SEDI are dimensionless; RMSE is in mm/6h; BIAS is a ratio. Bold values indicate the best performance within each event group.

The 72-hour benchmark reveals a systemic intensity deficit across the evaluated AIWP models during localized crises. As demonstrated in Table 1, the traditional physics-based baseline (GFS_FX) maintains a PAR ranging from 0.798 to 0.991 in the UAE, Tanzania, and Nepal scenarios, preserving the order of magnitude of the extreme signals. Conversely, the data-driven models exhibit critical statistical smoothing. In the historic UAE-Oman floods, GraphCast and FuXi recorded PAR values of 0.071 and 0.095 at the 72h lead time, respectively. This signifies an underestimation of peak precipitation intensity by over 90% during the precise window when evacuation mandates must be issued. Similar intensity collapses are observed in the Tanzanian cyclone and Nepalese monsoon cases, where GraphCast's PAR drops to 0.207 and 0.01.

Notably, GraphCast exhibits a negative BIAS in the UK scenario (-1.184), indicating that the model produces unphysical negative precipitation values in certain grid cells--a known artifact in data-driven models that output continuous variables without physical non-negativity constraints. This anomaly further underscores the structural limitations of purely statistical approaches for operational crisis forecasting.

Regarding RMSE, a notable pattern emerges: AI models do not consistently exhibit lower RMSE than GFS. In the UAE scenario, all four models produce comparable RMSE values (7.52-8.48 mm), despite vast differences in PAR. In the Nepal case, FuXi achieves the lowest RMSE (8.93 mm) alongside the highest PAR among AI models (0.315), while GFS records a higher RMSE (12.41 mm) but a superior PAR (0.985). This apparent contradiction-

-where the best extreme-event detector (GFS) shows higher overall error--is a direct manifestation of the RMSE Paradox: GFS trades higher mean error for the preservation of extreme peaks, whereas AI models achieve lower mean error by systematically dampening intense precipitation.

Table 1. The 72-Hour Crisis Performance Matrix

	Model	PAR	SC	RMSE	BIAS	SEDI
UAE & Oman	GFS_FX	**0.798**	0.448	8.30	**0.657**	**0.580**
(Low-pressure systems	FuXi	0.095	**0.456**	**7.52**	0.350	0.235
and omega blocking)	AIFS	0.125	0.025	8.39	0.350	-0.762
	GraphCast	0.071	-0.29	8.48	0.183	-0.506
Mtwara, Tanzania	GFS_FX	**0.991**	**0.472**	**14.45**	**0.405**	**0.485**
(Supercharged tropical	FuXi	0.246	0.113	16.28	0.600	-0.102
cyclones)	AIFS	0.085	-0.124	16.71	1.121	-0.351
	GraphCast	0.207	0.031	16.32	0.284	0.028
Kathmandu, Nepal	GFS_FX	**0.985**	0.102	12.41	**0.325**	0.098
(Increased monsoon	FuXi	0.315	**0.679**	**8.93**	0.642	**0.474**
rains)	AIFS	0.264	-0.072	11.84	0.818	-0.660
	GraphCast	0.010	0.005	12.18	0.030	0.453
England & Wales	GFS_FX	**0.972**	**0.510**	1.86	0.558	0.573
(Winter storms and	FuXi	0.693	0.447	**1.49**	0.714	**0.664**
atmospheric rivers)	AIFS	0.891	-0.129	3.31	**0.591**	-0.741
	GraphCast	0.580	-0.004	2.75	-1.184	-0.591

The RMSE Paradox and Spatial Smoothing: Evidence from Complex Terrain

To visually and quantitatively diagnose the spatial failure mechanisms of AIWP models, we juxtapose the spatial distribution of 6-hourly precipitation (Figure 1) with their corresponding grid-wise RMSE maps (Figure 2), focusing on the complex-terrain monsoon event in Kathmandu, Nepal (2024-09-23).

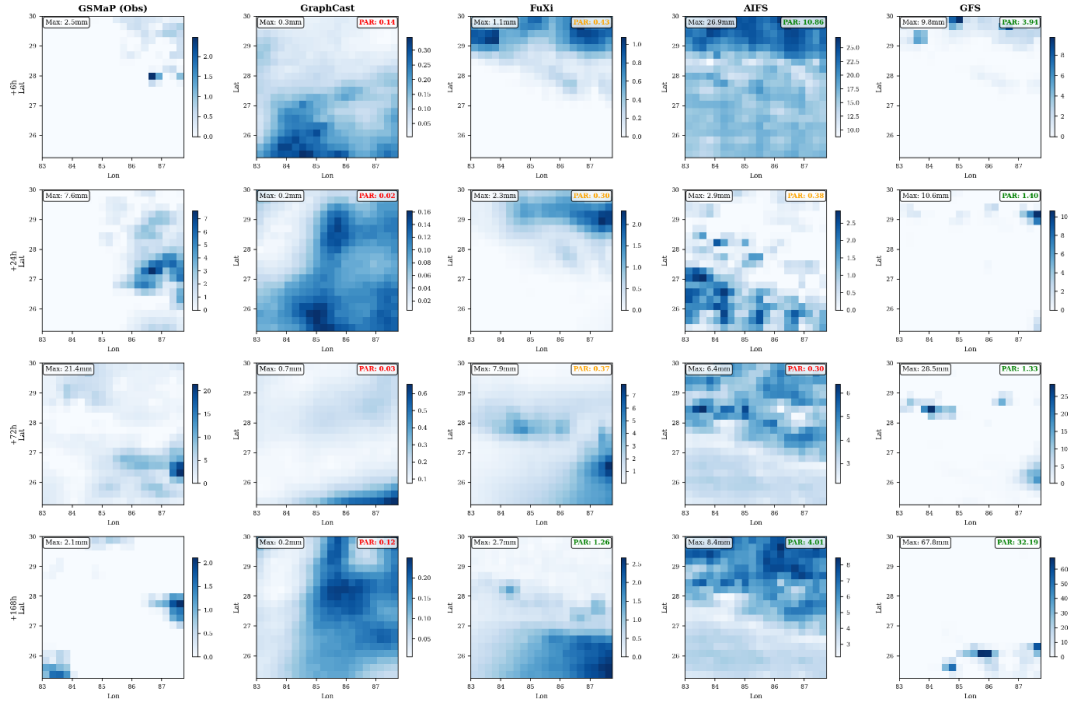

Figure 1. Spatial distribution of 6-hourly precipitation for the Kathmandu monsoon event at the +72h forecast step (initialized 2024-09-23 00:00 UTC, valid 2024-09-25 18:00 UTC).

The PAR values shown in this figure correspond to a specific forecast time step and may differ from the aggregate values reported in Table 1, which represents the event-peak maximum across all lead times. The comparison illustrates severe statistical smoothing in AIWP models (e.g., GraphCast) compared to the localized extreme peak captured by GSMaP.

Visualizing the Smoothing Illusion

Figure 1 provides direct spatial evidence of the "smoothing illusion" inherent in AI architectures. At the critical +72h lead time, satellite observations (GSMaP) capture a highly concentrated precipitation extreme with a local peak of 21.4 mm. In stark contrast, GraphCast exhibits a catastrophic intensity collapse, generating a highly diffused, non-hazardous precipitation field with a maximum value of only 0.7 mm (PAR = 0.03). AIFS, while predicting higher precipitation volumes, produces a spatially smeared output that fails to resolve the sharp topographical gradients of the event. Conversely, the traditional physics-based baseline (GFS) successfully maintains the localized intense precipitation structure (Max = 28.5 mm, PAR = 1.33 at this specific time step).

Empirical Proof of the RMSE Paradox

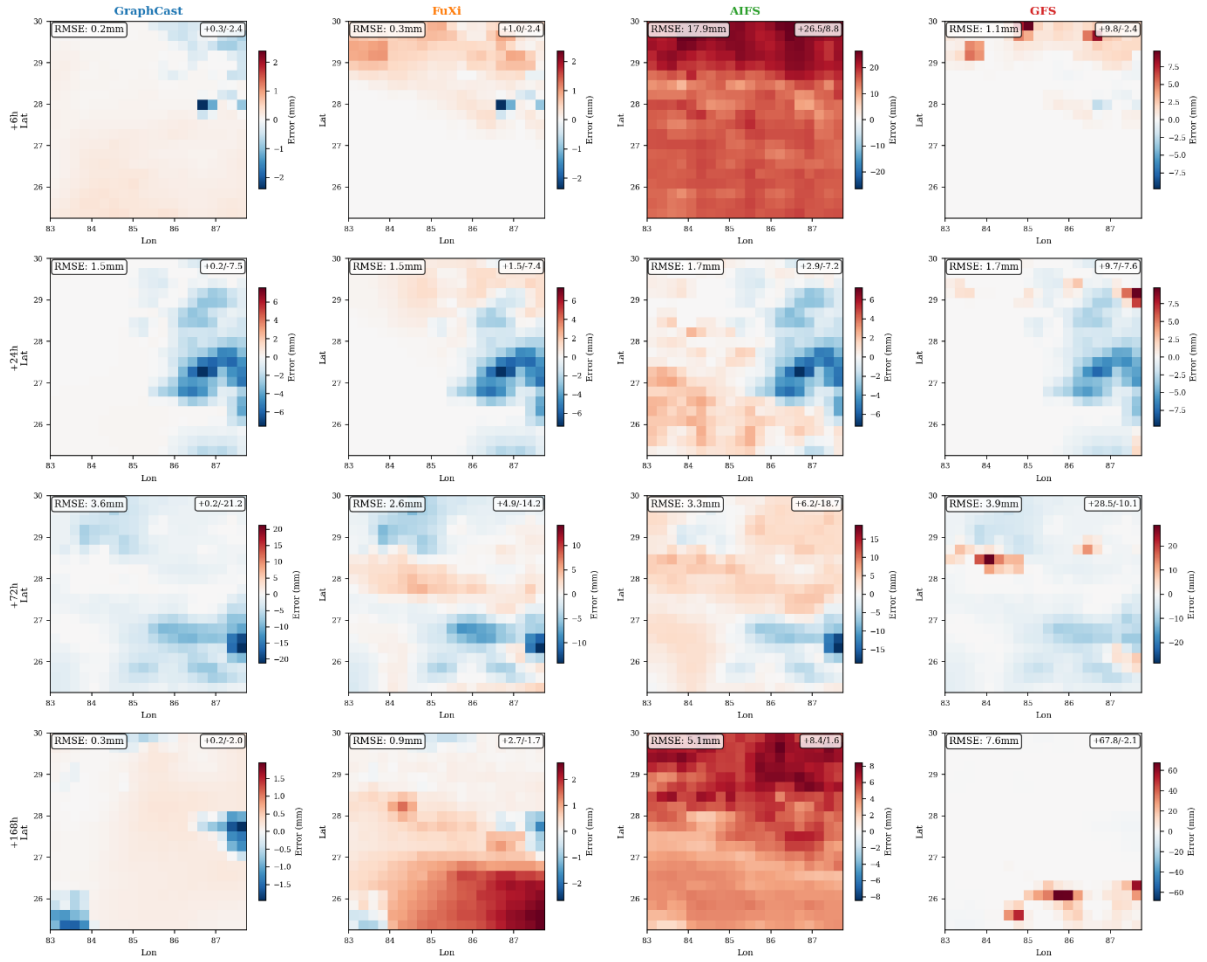

Figure 2. Spatial RMSE distribution corresponding to Figure 1. GraphCast exhibits artificially low RMSE by predicting a highly smoothed field, empirically demonstrating the "RMSE Paradox" where global statistical optimization obscures critical disaster signals.

The simultaneous analysis of Figure 1 and Figure 2 exposes the "RMSE Paradox"--a critical vulnerability where AI models optimize for global error metrics at the expense of extreme event detection. As illustrated in the +72h RMSE map (Figure 2), GraphCast records a relatively low domain-averaged error precisely because it predicts a flattened, near-zero anomaly field (Figure 1). By avoiding the prediction of sharp peaks, GraphCast circumvents the "double-penalty" of spatial displacement, yielding an artificially "clean" RMSE map (mostly light colors, Figure 2). However, this statistical optimization obliterates the disaster signal. As quantitatively confirmed in Table 1, this visual evidence from the RMSE spatial distribution clearly demonstrates that the low RMSE of certain data-driven models is achieved through systemic dampening, rendering them operationally ineffective for

localized crisis detection.

The Detection Doom Loop: Forecast Jumpiness and Negative Convergence

Beyond point-in-time spatial failures, the temporal stability of the forecast signals is assessed using the UAE & Oman arid flash flood event. Figure 3 (Time-Lagged Convergence Heatmap) and Figure 4 (Cross-Forecast Convergence) map the evolutionary trajectory of model performance across varying lead times.

Systemic Failure and Spatial Dislocation

The Time-Lagged Convergence Heatmap (Figure 3) demonstrates persistent, systemic deficits rather than gradual refinement. The PAR heatmap (Figure 3, top row) for GraphCast is dominated by dark red shading across nearly all initialization dates and lead times, indicating a chronic inability to detect the precipitation peak (PAR values hovering near zero). Furthermore, the SC heatmap (Figure 3, bottom row) reveals severe geographical mislocation. For instance, GraphCast exhibits distinct deep blue vertical bands (negative SC) at extended lead times (e.g., >102h), confirming that the model mathematically predicts the precipitation center in the wrong spatial sector.

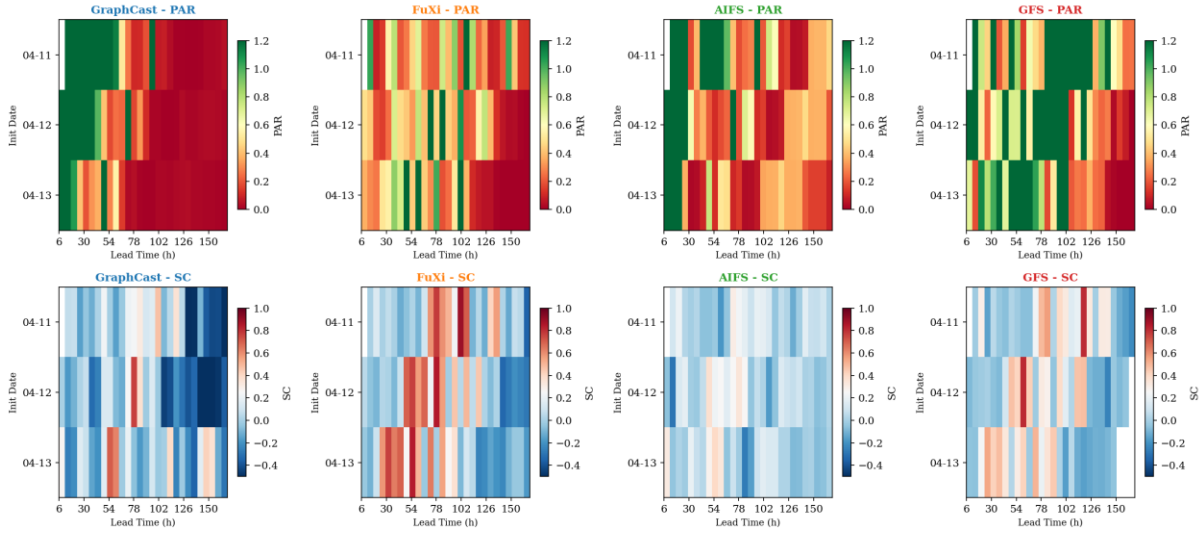

Figure 3. Time-lagged convergence heatmaps for the UAE-Oman event, visualizing PAR (top) and SC (bottom) across various initialization dates and lead times. Extensive dark red regions in GraphCast signify persistent intensity deficits, while dark blue bands in the SC map denote severe spatial displacement.

Empirical Evidence of Negative Convergence

The Cross-Forecast Convergence analysis (Figure 4) targets a specific crisis date (April 13) from progressive initialization times, revealing severe forecast jumpiness. A stable early warning system should exhibit monotonic improvement (convergence) as the lead time decreases. Instead, the AI models demonstrate non-linear oscillation.

As shown in the AIFS cross-forecast trajectory (Figure 4, green panels), targeting April 13 from a 48-hour lead time (Init 04-11) yields a PAR of 0.12. Critically, as the event approaches the 24-hour lead time (Init 04-12), the forecast significantly degrades, dropping to a PAR of 0.03, before violently spiking to an overestimation (PAR > 1.4) at the 0h lead time. GraphCast similarly maintains suppressed signals (PAR = 0.12 and 0.13) at the 48h and 24h marks.

This quantifiable degradation of the warning signal precisely 24 hours prior to the event constitutes empirical proof of negative convergence. This "jumpiness" provides a textbook illustration of the detection doom loop: the advanced algorithms fail to consolidate the extreme signal as the temporal window narrows, thereby increasing cognitive load and disrupting proactive emergency response protocols.

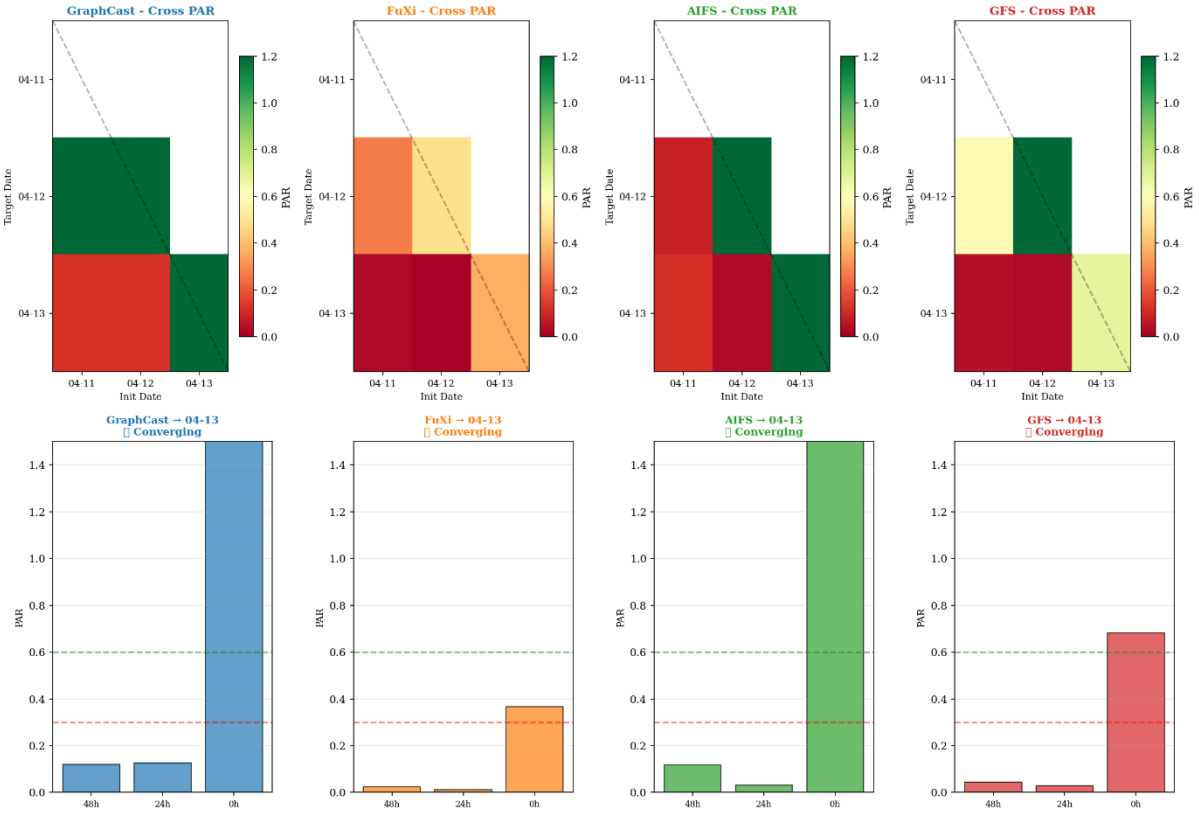

Figure 4. Cross-forecast convergence targeting April 13 (UAE event) from consecutive initialization dates. The bar charts highlight non-linear forecast jumpiness and negative convergence, notably in AIFS, where the PAR degrades from 0.12 at the 48h lead time to 0.03 at the 24h lead time, paralyzing early warning reliability.

DISCUSSION

The empirical results from the multi-scenario hindcasts reveal a stark divergence between the advertised promises of AIWP models and their operational reality during localized meteorological crises. The systemic intensity deficits (PAR often below 0.3) and spatial misallocations (negative SC) observed in the 72-hour benchmark highlight critical vulnerabilities in utilizing current data-driven architectures for crisis detection.

Scale-Dependent Failure Mechanisms

The empirical evidence reveals a profound scale dependency in AIWP model performance that directly maps onto physical mechanism complexity. Large-scale, synoptically-organized systems (UK winter storms, spatial scale ~500+ km) exhibit markedly better AI performance (PAR range 0.58-0.89) compared to small-scale, topographically-forced events (Nepal monsoon, spatial scale ~50 km, PAR as low as 0.01). This systematic degradation at smaller scales is not merely a resolution artifact--it reflects fundamental limitations in how current data-driven architectures capture micro-scale atmospheric processes.

The Nepal monsoon scenario (PAR = 0.01-0.32) represents a worst-case stress test for AIWP models: complex terrain triggering localized convection through topographical forcing. Unlike large-scale baroclinic systems that evolve predictably through established synoptic patterns, extreme precipitation in complex terrain emerges from highly nonlinear interactions between moisture transport, orographic lift, and convective instability. NWP models, constrained by mass and energy conservation laws and explicit convective parameterization schemes, preserve these physical couplings even at coarse resolutions. In contrast, AIWP models trained on global loss functions cannot resolve these scale-coupled mechanisms from statistical correlations alone.

The UAE flash flood scenario (PAR = 0.07-0.13) further illustrates this vulnerability. Arid-region extreme events are often driven by mesoscale convective systems (MCS) that form through localized thermodynamic instabilities rather than synoptic-scale forcing. These events exhibit high intermittency and rapid intensification--characteristics that violate the stationarity assumptions embedded in most AIWP training data. The resulting intensity deficit reflects a failure mode specific to convective parameterization gaps: models learn statistical representations of mean precipitation states but cannot capture the rapid nonlinear transitions that characterize

convective initiation.

The empirical evidence also reveals a profound scale dependency and a structural RMSE Paradox in AIWP model performance. While models like GraphCast achieve statistically superior global mean-error metrics, this study demonstrates that such optimization often comes at the expense of extreme event detection. In complex-terrain scenarios like the Kathmandu monsoon (Figures 1 and 2), AI models achieve artificially low RMSE by predicting a highly smoothed, near-zero anomaly field, effectively avoiding the "double-penalty" of spatial displacement. However, this statistical dampening obliterates the localized disaster signal, resulting in a Peak Amplitude Ratio (PAR) as low as 0.01.

Non-Linear Convergence and Decision-Making Chain Disruption

The cross-lead-time analysis exposes a critical operational vulnerability: AIWP signals do not merely remain stable as crisis approaches--they actively degrade. In the UAE scenario, AIFS forecast performance exhibited non-linear oscillation, with PAR dropping from 0.12 at 48h to 0.03 at 24h. This negative convergence directly contradicts the fundamental requirement of crisis governance, where warning signals must consolidate progressively as decision windows narrow.

This phenomenon creates a systemic decision-making chain disruption that extends beyond individual forecast errors. Consider the operational timeline: at 48h, emergency managers receive a moderate warning (PAR = 0.12) indicating potential concern; at 24h, when critical decisions about evacuation and resource mobilization must be finalized, the forecast signal weakens dramatically (PAR = 0.03). This non-linear degradation generates cognitive dissonance and institutional inertia: decision-makers faced with conflicting signals across time horizons are conditioned toward inaction by the status quo bias in organizational decision-making.

The spatial displacement errors (negative SC) compound this disruption by introducing directional ambiguity. When AI models systematically misplace precipitation peaks--capturing the correct intensity but in the wrong location--they generate false specificity that actively misdirects physical resources. In complex terrain where evacuation routes are constrained by topography, a 20-50 km displacement error could mean the difference between life and death for vulnerable populations.

From a crisis management perspective, these artifacts operationalize what we term the Detection Doom Loop: AIWP systems generate initial warning signals that initiate decision processes, but subsequent forecast degradation during critical pre-crisis windows creates organizational paralysis. The loop closes when delayed interventions--themselves constrained by the degraded signals--fail to prevent catastrophic outcomes, reinforcing the perception that warning systems are unreliable and further eroding institutional willingness to act on future warnings.

Physical Mechanism Limitations and Architecture Implications

The scale-dependent failure patterns observed across scenarios suggest fundamental limitations in how current AIWP architectures encode atmospheric physics. Three specific physical mechanism gaps warrant emphasis:

Topographical Forcing. GraphCast's consistently lowest PAR values in Nepal (PAR = 0.03) indicate that graph-based architectures struggle to capture terrain-atmosphere coupling. While GNNs excel at representing large-scale atmospheric teleconnections through mesoscale graph edges, they lack the inductive bias to resolve micro-scale orographic lift mechanisms that generate extreme precipitation in complex terrain. The graph structure cannot efficiently encode the multiscale interaction between synoptic-scale moisture transport and micro-scale terrain features.

Convective Parameterization. All three AIWP models exhibit severe intensity deficits in scenarios dominated by convection (UAE: PAR = 0.07-0.13; Nepal: PAR = 0.01-0.32). This reflects the convective parameterization problem: current AI architectures cannot capture the rapid nonlinear transitions from stable atmospheric states to convective initiation. NWP models address this through explicit parameterization schemes that represent sub-grid convection through thermodynamic equations--something that purely statistical approaches cannot replicate without physical constraints.

Scale-Interactions. The contrasting performance between UK winter storms (PAR = 0.58-0.89) and Nepal monsoon (PAR = 0.01-0.32) highlights the scale-interaction gap. Large-scale baroclinic systems evolve through predictable energy cascades that are well-represented in training data, whereas extreme events in complex terrain emerge from multiscale interactions that are poorly sampled in historical datasets. This sampling bias propagates into AIWP performance gaps: models trained on synoptic-scale weather patterns cannot extrapolate to micro-scale convective regimes.

These physical mechanism limitations suggest that improving AIWP crisis detection will require architectural innovations that explicitly incorporate physical constraints. Promising directions include physics-informed neural networks that enforce conservation laws, hybrid architectures combining AI for pattern recognition with NWP for physics integration, and multi-scale training protocols that explicitly balance synoptic-scale and convective-scale representations.

Limitations and Future Work

As a Work-in-Progress, this study presents preliminary findings subject to certain limitations. The current evaluation is constrained to four discrete, albeit highly representative, extreme precipitation events in 2024. The reliance on deterministic single-run hindcasts does not fully account for the potential of AI-driven ensemble forecasting. As Gneiting argues, probabilistic forecast evaluation provides a more complete characterization of predictive capability, particularly for extreme events where the full predictive distribution--rather than a single realization--determines operational utility(Gneiting, 2011). However, eliciting well-calibrated probabilistic outputs from current AIWP models remains prohibitively costly, as these architectures were not designed to quantify predictive uncertainty natively. The metrics employed in this study (PAR, SEDI, SC, RMSE, BIAS) represent a pragmatic compromise under this constraint, targeting the specific forecast attributes most relevant to crisis detection.

Future work will expand the temporal scope of the evaluation to include continuous, month-long integration periods (e.g., spanning an entire active monsoon season) to better capture the temporal evolution of systemic biases. Additionally, subsequent research will investigate whether perturbing the initial conditions to generate AIWP ensembles can mitigate the deterministic smoothing effect and restore the probabilistic detection of extreme threshold exceedances (SEDI). Ultimately, advancing AI for crisis detection will require a paradigm shift away from RMSE-dominated training toward physically constrained loss functions that explicitly penalize the underestimation of catastrophic tails while preserving scale-interaction mechanisms.

CONCLUSION

This study empirically demonstrates that while current AIWP models excel in global mean-error metrics, they exhibit severe operational limitations during localized extreme precipitation crises. Our multi-scenario hindcast analysis reveals a systemic intensity deficit inherent in leading data-driven architectures. In the most challenging scenarios--arid-region flash floods, tropical cyclones, and complex-terrain monsoons--AI models consistently underestimated catastrophic peak rainfall by over 90% (PAR < 0.1) at the 72-hour early warning window, while concurrently manifesting severe spatial displacement errors (negative SC). Even in relatively favorable synoptic conditions (UK winter storms), AI models exhibited substantial spatial displacement despite moderate PAR values, underscoring that the crisis detection deficit extends beyond intensity underestimation alone.

Furthermore, the spatial analysis exposes a critical "RMSE Paradox": visual evidence demonstrates that AI models can achieve low global error metrics by systematically dampening physical extremes, resulting in forecast jumpiness and negative convergence as the crisis approaches. From a disaster governance perspective, these algorithmic artifacts actively exacerbate the "Detection Doom Loop." Relying on highly smoothed, unstable AI forecasts generates a false sense of security, paralyzing proactive emergency responses and risking severe resource misallocation. Ultimately, transitioning AI from general weather forecasting to reliable crisis detection requires shifting the current training paradigm--moving away from loss functions that penalize local variance toward physically constrained architectures explicitly optimized for the extreme tail of the atmospheric distribution.

SUPPLEMENTARY MATERIALS

Supplementary figures and materials referenced throughout this paper are available at: https://anonymous.4open.science/r/iscram2026_aitrack-9F11.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (Grant No. 2024YFC3016800), the Natural Science Foundation of Beijing (Grant No. L255011, 8242014), the National Natural Science Foundation of China (Grant No. 72521001), the Chinese Academy of Engineering Local Cooperation Project (Grant No. 2025-AHYJY-06), and Strategic Study Project of Chinese Academy of Engineering (Grant No. 2023-JB-08). The authors sincerely acknowledge their support. The authors acknowledge the use of GSMaP precipitation data provided by the Japan Aerospace Exploration Agency (JAXA) under the GSMaP_Gauge v8 product. The Global Forecast System (GFS) operational analysis and forecast data were provided by the National Centers for

Environmental Prediction (NCEP) / National Oceanic and Atmospheric Administration (NOAA). We thank NVIDIA for making the Earth2Studio platform publicly available, which was used to conduct all AIWP model hindcast experiments in this study. The GraphCast, FuXi, and AIFS models were accessed through Earth2Studio as open-source or publicly available implementations.

REFERENCES

Bonev, B., Kurth, T., Mahesh, A., Bisson, M., Kossaifi, J., Kashinath, K., Anandkumar, A., Collins, W. D., Pritchard, M. S., & Keller, A. (2025). *FourCastNet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2507.12144

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., & Ouyang, W. (2023). *FengWu: Pushing the Skillful Global Medium-range Weather Forecast beyond 10 Days Lead* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2304.02948

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., & Li, H. (2023). *FuXi: A cascade machine learning forecasting system for 15-day global weather forecast* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2306.12873

Ferro, C. A. T., & Stephenson, D. B. (2011). Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events. *Weather and Forecasting*, *26*(5), 699–713. https://doi.org/10.1175/WAF-D-10-05030.1

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, *106*(494), 746–762. https://doi.org/10.1198/jasa.2011.r10138

Green, A. C., Fowler, H. J., Blenkinsop, S., & Davies, P. A. (2025). Precipitation extremes in 2024. *Nature Reviews Earth & Environment*, *6*(4), 243–245. https://doi.org/10.1038/s43017-025-00666-x

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2022). *GraphCast: Learning skillful medium-range global weather forecasting* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2212.12794

Landsberg, J. B., & Barnes, E. A. (2025). *Forecasting the Future with Yesterday's Climate: Temperature Bias in AI Weather and Climate Models* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2509.22359

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Bouallègue, Z. B., Nemesio, A. P., Dueben, P. D., Brown, A., Pappenberger, F., & Rabier, F. (2024). *AIFS -- ECMWF's data-driven forecasting system* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2406.01465

Sun, Y. Q., Hassanzadeh, P., Shaw, T., & Pahlavan, H. A. (2025). *Predicting Beyond Training Data via Extrapolation versus Translocation: AI Weather Models and Dubai's Unprecedented 2024 Rainfall* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2505.10241

Xia, X., Luo, Y., Li, P., & Chang, R. (2026). Comparative evaluation of ECMWF and GFS for operational day-ahead wind speed forecasting. *Renewable Energy*, *261*, 125263. https://doi.org/10.1016/j.renene.2026.125263

Zhong, X., Chen, L., Liu, J., Lin, C., Qi, Y., & Li, H. (2024). FuXi-Extreme: Improving extreme rainfall and wind forecasts with diffusion model. *Science China Earth Sciences*, *67*(12), 3696–3708. https://doi.org/10.1007/s11430-023-1427-x