Evaluating AI Weather Models for Crisis Detection

A multi-scenario hindcast of extreme precipitation forecasts

Presenter: Feng Huang

huangf23@outlook.com

School of Safety Science, Tsinghua University

2026.5.24

Significance

Why this matters

Extreme precipitation forecasts are not only scientific products. They trigger evacuation, resource deployment, emergency staffing, and public warnings.

- AI weather prediction models are increasingly presented as operationally transformative
  [Lam et al., 2022, Bi et al., 2023, Chen et al., 2023, Lang et al., 2024].
- Crisis management depends on detecting the dangerous tail, not only the average weather state.
- A calm-looking forecast can delay action and create a false sense of security, especially when extremes fall outside historical training ranges
  [Watson, 2022, Zhang et al., 2026].

Literature Review

AIWP progress

- GraphCast [Lam et al., 2022]
- Pangu-Weather [Bi et al., 2023]
- FuXi [Chen et al., 2023]
- AIFS [Lang et al., 2024]

Evaluation tradition

- Global mean-error metrics
- Medium-range forecast skill
- Large-scale atmospheric fields

Gap

Crisis management requires tail-aware, location-aware evaluation of low-probability and high-impact extremes
[Ferro and Stephenson, 2011, Gneiting, 2011, Lerch et al., 2017].

Evaluation Gap

Standard evaluation asks

- Is the mean field accurate?
- Is RMSE low?
- Does the forecast look smooth?

Crisis evaluation asks

- Is the peak intensity preserved?
- Is the peak geographically aligned?
- Is the warning signal stable?

A forecast can be statistically clean and operationally unsafe.

Recent record-breaking-extreme benchmarks show the same split: AI models can perform well on average metrics while underestimating the rare events that matter most [Zhang et al., 2026].

Why Tail Events Stress AIWP

Training distribution

Rare extremes contribute little to the loss function and may sit outside the historical range learned by the model.

Spatial scale

Disaster rainfall is often defined by sharp local peaks rather than broad synoptic patterns.

Decision asymmetry

Missing one dangerous peak can matter more than small errors over many harmless grid cells.

Operational implication

A crisis-oriented evaluation must ask whether the model preserves the actionable signal, not only whether it minimizes average error.

Research Question

Core question

Can current AI weather prediction models reliably detect localized extreme precipitation crises?

Forecasting skill

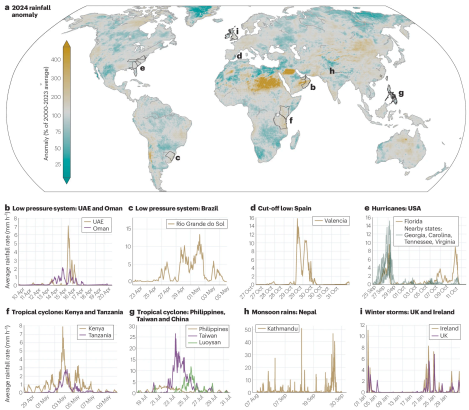
How accurate is the model on average across a large domain?

Crisis detection skill

Does the model preserve the dangerous peak in the right place, early enough?

Do global mean-error gains translate into emergency warning skill?

Data Collection: Four 2024 Extreme Precipitation Cases



Cases are selected from Green et al. (2025), *Nature Reviews Earth & Environment*: Precipitation extremes in 2024 [Green et al., 2025].

Study Cases

Event	Driver	Crisis signature
UAE and Oman	Cut-off low and blocking	More than 240 mm in 24 h in an arid region
Mtwara, Tanzania	Tropical Cyclone Hidaya	Extreme tropical rainfall and mass evacuations
Kathmandu, Nepal	Monsoon surge over complex terrain	240 mm in 24 h, landslides, terrain forcing
England and Wales	Winter storms and atmospheric rivers	Widespread 50–100 mm rainfall and flooding

Case Selection Logic

Why these four cases

- High-impact 2024 precipitation events
- Different climate regimes and storm drivers
- Clear crisis-management relevance
- Localized peak behavior that stresses AIWP

What variation helps test

- Arid-region flash flooding
- Tropical cyclone rainfall
- Monsoon and complex terrain
- Mid-latitude winter storm flooding

The goal is not a climatology of all rainfall. It is a stress test of crisis-relevant forecast behavior.

Methodology

Models

- GraphCast
- FuXi
- AIFS
- GFS forecast baseline

Experimental setup

- GSMaP_Gauge v8 reference
- 6-hour accumulated precipitation
- Common 0.25 degree grid
- Standardized GFS initial fields
- Multi-lead-time hindcasts

This precipitation-focused design complements recent record-extreme benchmarks, which generally exclude precipitation because reanalysis precipitation carries substantial regional and resolution-dependent uncertainty [Lavers et al., 2022, Zhang et al., 2026].

From Forecast Field To Warning Signal

1. Generate multi-lead-time hindcasts from the same initial-condition framework.
2. Convert model precipitation outputs to a common spatial and temporal verification grid.
3. Compare forecasted 6-hour accumulation against GSMaP_Gauge reference fields.
4. Evaluate whether intensity, placement, and threshold exceedance would support early action.

Key design choice

The unit of analysis is not only a forecast map. It is the emergency signal that the map would create for a response organization.

Evaluation Metric System

Decision question	Metric dimension	Indicator	Operational interpretation
Is the dangerous rainfall strong enough to trigger action?	Intensity preservation	**PAR**	Tests whether the forecast keeps the extreme peak magnitude visible.
Is the warning placed in the correct response area?	Spatial reliability	**SC**	Tests whether the forecast rainfall pattern aligns with the observed crisis field.
Is the overall precipitation field close to observations?	Mean-field accuracy	**RMSE**, Bias	Provides conventional error support, but may hide localized peak loss and reward smoothing.
Would a threshold-based warning be triggered?	Rare-event detection	**SEDI**	Evaluates hit and false-alarm behavior for rare high-impact exceedances.
Does the warning signal become clearer over lead time?	Lead-time stability	PAR and SC trajectories	Checks whether evidence consolidates or deteriorates as impact approaches.

Principle: the metric system links forecast verification to emergency decision needs: intensity, location, average error, threshold detection, and lead-time stability.

Crisis-Centric Metric Definitions

Metric	Formula	Interpretation
PAR	$PAR = \frac{\max_i(\hat{y}_i)}{\max_i(y_i)}$	Peak amplitude ratio. $PAR = 1$ is ideal; $PAR < 1$ indicates peak underestimation.
SC	$SC = \frac{\sum_i(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_i(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_i(y_i - \bar{y})^2}}$	Spatial correlation across the event domain; measures location and pattern agreement.
RMSE	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$	Mean grid-level error. Useful, but can favor smoothed fields and hide localized peak loss [Gneiting, 2011, Zhang et al., 2026].
Bias	$Bias = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$	Domain-mean signed error; negative values indicate overall underprediction.
SEDI	$SEDI = \frac{\log F - \log H - \log(1-F) + \log(1-H)}{\log F + \log H + \log(1-F) + \log(1-H)}$	Rare-event detection skill. $H = \frac{hits}{hits+misses}$ and $F = \frac{false\ alarms}{false\ alarms+correct\ negatives}$ [Ferro and Stephenson, 2011].
Stability	$\Delta PAR = PAR_{t_2} - PAR_{t_1}, \quad \Delta SC = SC_{t_2} - SC_{t_1}$	Lead-time change in intensity and spatial skill. Positive consolidation is expected as impact approaches.

How To Read The Metrics Together

A forecast can pass

- Low RMSE over the broad domain
- Small domain-mean bias
- Visually plausible rainfall pattern

And still fail

- Low PAR: peak too weak
- Low SC: peak in the wrong place
- Low SEDI: warning threshold missed

For crisis detection, the metrics must be interpreted as a bundle, not as isolated scores.

Finding Preview

Main finding

AIWP models show systematic intensity deficits and unstable warning signals in localized precipitation crises, even when conventional error metrics appear acceptable.

- Peak rainfall can be underestimated by more than 90%.
- Smooth forecasts can produce deceptively low RMSE.
- Warning signals may fail to converge as the crisis approaches.
- This pattern is consistent with broader evidence that AI models can underpredict record-breaking extremes [Zhang et al., 2026].

What Counts As Operational Failure

Weak signal

The forecast shows rainfall, but the peak is too weak to trigger escalation.

Wrong signal

The forecast places the hazard outside the actual response area.

Unstable signal

The forecast does not become more decisive as the event approaches.

Reading the following results

The question is therefore not simply which model has the lowest error, but which model preserves warning-worthy evidence under time pressure.

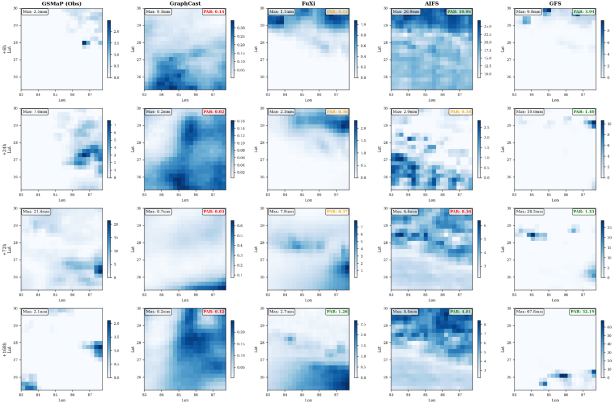
Analysis I: 72-Hour Warning Window

Event	GFS PAR	Best AI PAR	Worst AI PAR	Interpretation
UAE and Oman	0.798	0.125	0.071	Over 85–90% of peak missed
Tanzania	0.991	0.246	0.085	Tropical signal damped
Nepal	0.985	0.315	0.010	Near-total terrain failure
England and Wales	0.972	0.891	0.580	Better intensity, weak localization

Takeaway

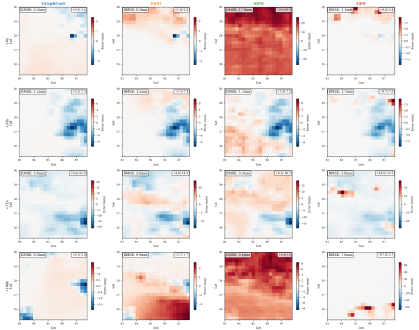
GFS preserves the magnitude of extreme rainfall more consistently than current AIWP models at the operationally critical 72-hour window.

Analysis II: Spatial Smoothing



Kathmandu monsoon event at the +72h forecast step. GraphCast produces a highly diffused field while GFS preserves the localized peak structure.

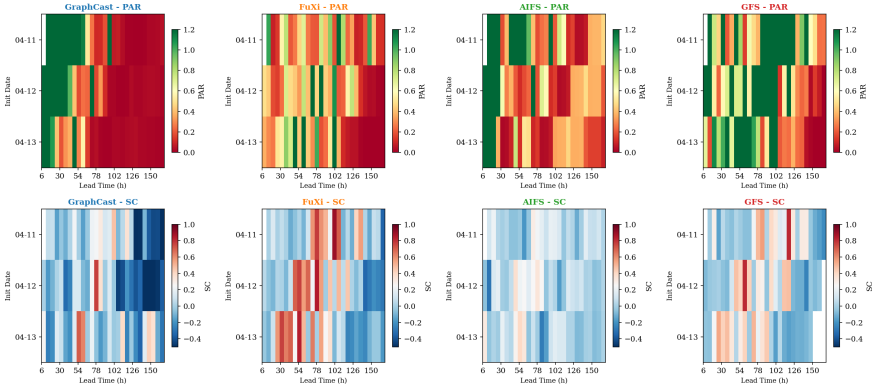
Analysis III: The RMSE Paradox



Interpretation

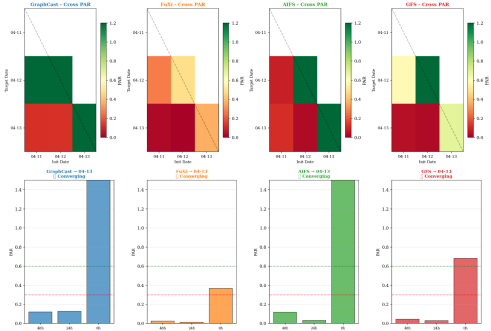
A model can obtain low domain-averaged RMSE by predicting a flattened field. But this same behavior removes the signal needed for crisis detection, echoing recent concerns about mean-error optimization and record-extreme underestimation [Zhang et al., 2026].

Analysis IV: Forecast Stability



UAE-Oman event: dark red PAR regions indicate persistent intensity deficits; blue SC bands indicate spatial displacement.

Analysis V: Negative Convergence



Crisis management implication

A stable warning system should consolidate as the event approaches. Here, AI forecast signals can oscillate or degrade exactly when emergency decisions must become more decisive.

Discussion: Scale-Dependent Failure

Topography

Complex terrain creates localized orographic lift that current AI architectures do not reliably preserve [Sun et al., 2025].

Convection

Rapid convective initiation violates smooth statistical expectations learned from mean weather states [Zhong et al., 2024].

Scale interaction

Large synoptic systems are easier than micro-scale extremes embedded in complex terrain.

- Best AI performance appears in large-scale UK winter storms.
- Worst AI performance appears in Nepal and UAE localized extremes.
- The likely failure mode is not random noise alone, but difficulty extrapolating sharp, rare amplitudes beyond familiar training patterns [Bonavita, 2024, Zhang et al., 2026].

Discussion: Extrapolation Risk

What recent evidence shows

AI models can underestimate both the intensity and occurrence of record-breaking heat, cold, and wind events, with larger bias for larger record exceedance [Zhang et al., 2026].

Why it matters here

Localized rainfall crises also depend on rare peak amplitudes, spatial concentration, and short lead-time warning stability.

Interpretation

Our case results should be read as part of a broader operational risk: high average skill does not guarantee reliable extrapolation to the most decision-relevant extremes.

Discussion: The Detection Risk Chain

1. AI forecast suppresses or misplaces the extreme signal.
2. Decision-makers receive a forecast that appears calm or ambiguous.
3. Warnings, evacuation, and resource mobilization are delayed.
4. Missed impacts reduce trust in early warning systems.

Governance risk

The danger is not only model error. It is model error that looks authoritative enough to slow down organizational response.

Discussion: Operational Use Of AIWP

Use AIWP for

- Fast scenario scanning
- Additional early signals
- Cross-model comparison
- Situational awareness support

Do not use AIWP alone for

- Local evacuation triggers
- High-confidence all-clear messages
- Replacing physics-based guidance
- Ignoring field reports or radar updates

The safest operational role is complementary: fast, useful, but verified against tail-aware warning evidence.

Limitations

- The evaluation is based on four representative extreme precipitation cases.
- The current analysis uses deterministic single-run hindcasts rather than calibrated AIWP ensembles.
- GSMaP reference uncertainty may be higher in complex terrain, especially for the Nepal case.
- Future work should test longer continuous periods, record-threshold precipitation benchmarks, and probabilistic extreme-event detection [Price et al., 2025, Zhang et al., 2026].

Conclusion

- Current AIWP models can under-detect localized extreme precipitation crises despite strong average forecast skill.
- RMSE-dominated evaluation can reward smoothing that removes disaster peaks.
- Reliable AI crisis detection requires tail-aware metrics, physical constraints, and operational stress testing.

AI weather prediction should be judged by whether it preserves the signals that trigger action.

References I

[Bi et al., 2023] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619, 533–538.

[Bonavita, 2024] Bonavita, M. (2024). On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51, e2023GL107377.

[Chen et al., 2023] 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. arXiv:2306.12873.

[Ferro and Stephenson, 2011] 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.

[Gneiting, 2011] Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.

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

[Lam et al., 2022] Lam, R., Sanchez-Gonzalez, A., Willson, M., et al. (2022). GraphCast: Learning skillful medium-range global weather forecasting. arXiv:2212.12794.

[Lang et al., 2024] Lang, S., Alexe, M., Chantry, M., et al. (2024). AIFS: ECMWF's data-driven forecasting system. arXiv:2406.01465.

[Lavers et al., 2022] Lavers, D. A., Simmons, A., Vamborg, F., & Rodwell, M. J. (2022). An evaluation of ERA5 precipitation for climate monitoring. *Quarterly Journal of the Royal Meteorological Society*, 148, 3152–3165.

[Lerch et al., 2017] Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., & Gneiting, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, 32(1), 106–127.

References II

[Price et al., 2025] Price, I., Sanchez-Gonzalez, A., Alet, F., et al. (2025). Probabilistic weather forecasting with machine learning. *Nature*, 637, 84–90.

[Sun et al., 2025] 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. arXiv:2505.10241.

[Watson, 2022] Watson, P. A. G. (2022). Machine learning applications for weather and climate need greater focus on extremes. *Environmental Research Letters*, 17, 111004.

[Zhang et al., 2026] Zhang, Z., Fischer, E., Zscheischler, J., & Engelke, S. (2026). Physics-based models outperform AI weather forecasts of record-breaking extremes. *Science Advances*, 12, eaec1433. https://doi.org/10.1126/sciadv.aec1433

[Zhong et al., 2024] 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.

Q&A

Questions?

Feng Huang

huangf23@outlook.com