



Quantifying the responses of AI precipitation forecast errors to reanalysis errors using analysis of covariance

Qiang Li¹ and Tongtiegang Zhao¹

¹Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Civil Engineering, Sun Yat-Sen University, Guangzhou, China

Correspondence to: Tongtiegang Zhao (zhaottg@mail.sysu.edu.cn)

Abstract. AI weather models are commonly trained by climate reanalysis datasets that inevitably influence AI weather forecasts. Given the importance of precipitation, it is essential to quantify how AI forecast errors respond to reanalysis errors. This paper proposes the use of the analysis of covariance (ANCOVA) method to diagnose the responses. Specifically, extreme precipitation events are identified from historical observations; both forecast and reanalysis errors are calculated against observations during the evaluation period; and the ANCOVA method is fitted to separate the responses to reanalysis errors from changes in response coefficients and additional offsets associated with extreme precipitation. A case study is devised for GraphCast forecasts and European Centre for Medium-range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5), using Daymet observations across 1426 catchments in the Catchment Attributes and MEteorology for Large-Sample SPATially distributed analysis (CAMELS-SPAT) dataset. The results show that the responses of GraphCast forecast errors to ERA5 reanalysis errors weaken progressively with increasing lead time. Under non-extreme conditions, the median response coefficient declines from 0.70 at the 1d lead time to 0.27 at 9d, indicating that each 1 mm/d increase in ERA5 precipitation errors leads to a median increase of 0.70 and 0.27 mm/d in GraphCast precipitation errors, respectively. Extreme precipitation amplifies the underestimation by GraphCast forecasts, with the median additional offset across 1426 catchments decreasing from -1.75 to -10.43 mm/d when the lead time increases from 1d to 9d. Overall, the ANCOVA method contributes to understandings of factors governing precipitation forecast errors and highlights the need to consider inherent uncertainties in climate reanalysis when training and evaluating AI weather models.



25 1 Introduction

Artificial intelligence (AI) weather models have advanced rapidly in recent years (Alet et al., 2026; Bodnar et al., 2025; Lam et al., 2023). Rather than explicitly solving the equations of atmospheric dynamics and physics, these models learn spatiotemporal relationships among atmospheric variables directly from historical climate reanalysis data (Rasp et al., 2024; Zhao et al., 2025). AI weather models offer substantial computational advantages while achieving competitive forecast skill (Chen et al., 2025; Bodnar et al., 2025). For example, Pangu-Weather (Bi et al., 2023) and GraphCast (Lam et al., 2023) can generate 10-day deterministic weather forecasts within a few minutes, with the predictive accuracy comparable to that of the Integrated Forecasting System (IFS) operated by the European Centre for Medium-range Weather Forecasts (ECMWF). Despite these advances, precipitation forecasting remains particularly challenging because of the intermittency, nonlinearity and spatial heterogeneity of precipitation (Huang and Zhao, 2025; Baño-Medina et al., 2025a; Tan et al., 2024). Although AI weather forecasts can reproduce large-scale precipitation patterns (Pan et al., 2025; Radford et al., 2025), substantial errors persist for heavy and extreme precipitation, including underestimation of intensity, spatial displacement and increasing errors with lead time (Li and Zhao, 2026a; Oh et al., 2026). These limitations underscore the importance of identifying the sources and characteristics of precipitation forecast errors for the further improvements of AI weather models.

Precipitation forecast errors in AI weather models may arise from the combined effects of biases inherited from training datasets, uncertainties accumulating with lead time and variations in performance across different climatic and topographic regimes (Liu et al., 2024; Oh et al., 2026; Pan et al., 2025). For AI models, forecast skill is inherently constrained by the information and errors embedded in the training data (Mu et al., 2026; Abel et al., 2026; Gupta et al., 2026). When climate reanalysis products, such as the ECMWF Reanalysis v5 (ERA5) (Hersbach et al., 2020), are used as training targets, reanalysis errors can be learned by AI weather models and subsequently propagated into their forecasts (Abel et al., 2026; Yuval et al., 2026). Such errors are spatially and temporally heterogeneous because the precipitation reanalysis is affected by observational constraints, physical parameterizations and data-assimilation procedures (Jiang et al., 2021; Hersbach et al., 2020). Consequently, the extent to which reanalysis errors are inherited by AI precipitation forecasts may vary with lead time as well as across climatic and topographic conditions (Baño-Medina et al., 2025a; Nie et al., 2026; Selz and Craig, 2023). It remains unclear whether reanalysis errors are amplified or attenuated in AI precipitation forecasts (Gentile et al., 2026; Mu et al., 2025). This issue is particularly important for extreme precipitation, whose intensity is often systematically underestimated in reanalysis products (Tang et al., 2020; Wang and Tian, 2025).

Quantitative methods are therefore in demand to assess how precipitation errors in AI weather forecasts respond to errors in climate reanalysis (Mu et al., 2026; Abel et al., 2026; Sun et al., 2025). The analysis of covariance (ANCOVA) method is well suited for this task because it can incorporate continuous covariates and categorical factors within a unified regression model (Gupta et al., 2015; Gao et al., 2023), thereby allowing response coefficients to be estimated and compared across different conditions (Fauchaux et al., 2022; McEachran et al., 2021). In this paper, we propose the use of the ANCOVA method to quantify the responses of AI precipitation forecast errors to reanalysis errors and to assess how these



60 responses vary under extreme precipitation. For GraphCast forecasts and ERA5 reanalysis, a case study is devised using
Daymet dataset across 1426 catchments in the Catchment Attributes and MEteorology for Large-Sample SPATially
distributed analysis (CAMELS-SPAT) dataset (Knoben et al., 2025). This analysis focuses on the catchment scale because it
provides the fundamental spatial unit for linking precipitation forecast errors to their impacts for potential hydrological
applications (Abbas et al., 2026; Meng and Zhao, 2025). As will be shown in the Methods and Results, the ANCOVA
method contributes to understandings of factors governing precipitation forecast errors, thereby providing useful information
for the further development and evaluation of AI weather models.

65

2 Methods

2.1 Overview of the analysis of covariance

70 There are three steps to quantify the responses of AI precipitation forecast errors to reanalysis errors, as shown in
Fig. 1. Prior to the analysis, precipitation forecasts, reanalysis data and observations are spatially and temporally aligned to
ensure consistency across these datasets. First, extreme precipitation events are identified separately for each catchment
using observations and the catchment-specific percentile threshold. Second, forecast and reanalysis errors are calculated
relative to the same observational reference and matched according to catchment, lead time and valid time. Third, the
ANCOVA method is fitted to quantify the responses of forecast errors to reanalysis errors while explicitly accounting for the
influence of extreme precipitation. The resulting model parameters characterize the responses coefficients under non-
75 extreme and extreme precipitation conditions.

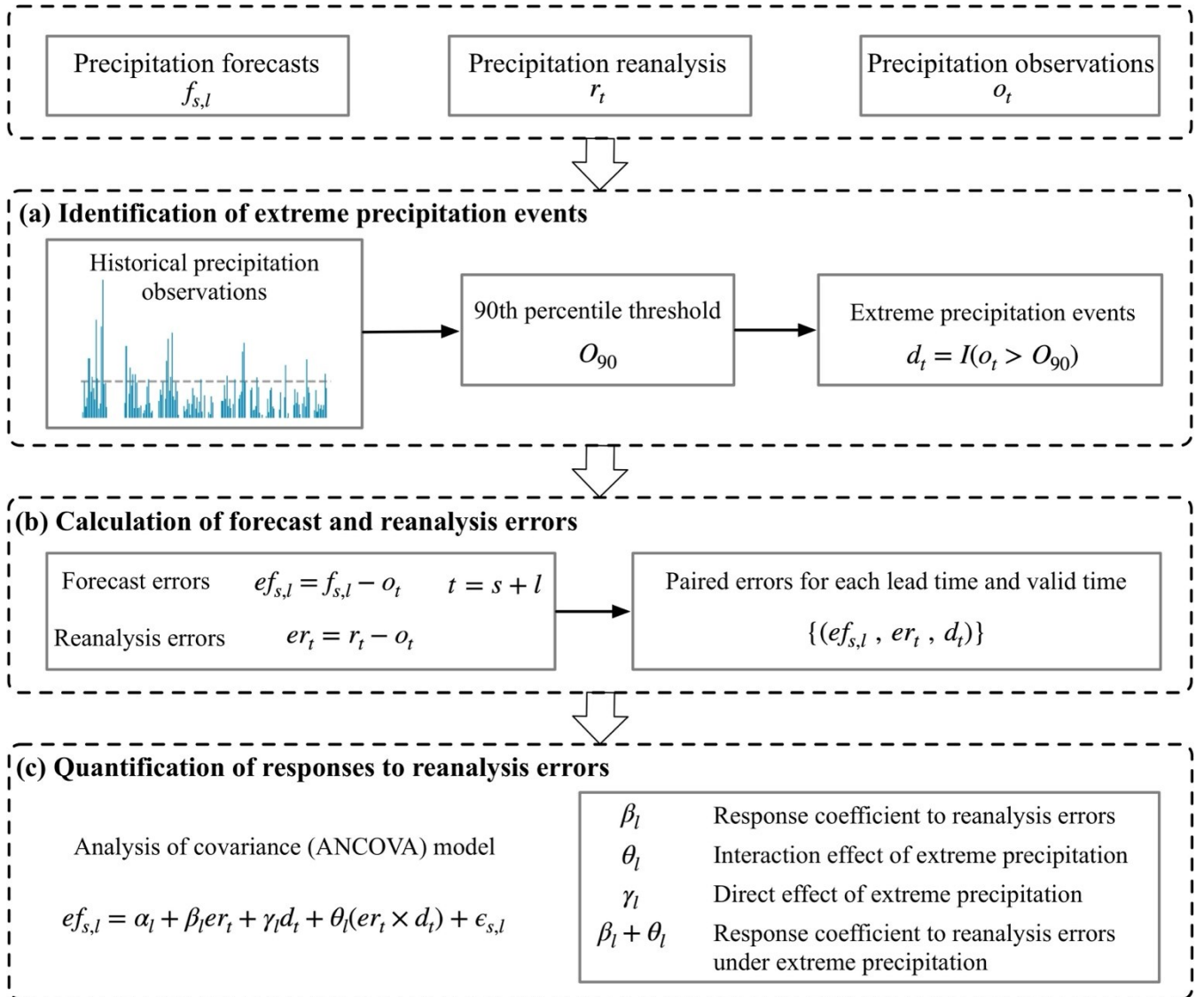


Figure 1. Schematic view of the ANCOVA method to quantify the responses of forecast errors to reanalysis errors.

2.2 Identification of extreme precipitation events

Precipitation observations from the training period of the specific AI weather model are used to calculate the percentile threshold for identifying extreme precipitation events. Using the training period ensures that the threshold is defined independently of the evaluation period, thereby avoiding information leakage from the evaluation data and maintaining consistency with the precipitation climatology represented during model training (Zhang et al., 2026). Because precipitation climatology varies substantially among catchments, the threshold is determined separately for each catchment



rather than using a single value across all catchments (Olivetti and Messori, 2024; Zhao et al., 2025). This catchment-specific definition accounts for spatial heterogeneity in precipitation regimes and avoids classifying the same precipitation intensity as equally extreme events across regions with contrasting climate conditions.

The catchment-specific threshold derived from the training period is then applied to the evaluation period. For each
 90 catchment, extreme precipitation event is defined as:

$$d_t = I(o_t > O_{90}) \quad (1)$$

where o_t represents the precipitation observation at time t ; O_{90} is the 90th percentile of precipitation observations during the training period; $I(\cdot)$ is the indicator function that equals 1 when the specified condition is satisfied and 0 otherwise. Thus, $d_t = 1$ denotes an extreme precipitation event, while $d_t = 0$ denotes a non-extreme event.

95 2.3 Calculation of forecast and reanalysis errors

Forecast and reanalysis errors are calculated as the differences between precipitation forecasts and observations and between reanalysis data and observations, respectively. Using the same observational reference ensures that the two error terms are directly comparable and that their relationship is not confounded by differences in verification data. For each catchment, the forecast error $ef_{s,l}$ is calculated as:

$$ef_{s,l} = f_{s,l} - o_t, \quad t = s + l \quad (2)$$

100 where $f_{s,l}$ denotes the precipitation forecast initialized at time s with lead time l ; o_t is the precipitation observation at the corresponding valid time t . The reanalysis error er_t at the same valid time is calculated as:

$$er_t = r_t - o_t \quad (3)$$

where r_t is the precipitation reanalysis at valid time t . Positive and negative values indicate overestimation and underestimation relative to the observations, respectively. Signed errors are retained rather than converted to absolute errors because the sign preserves information on the direction of the precipitation error, which is required to quantify how forecast
 105 errors vary with reanalysis errors. For each forecast sample, $ef_{s,l}$ is paired with er_t and the extreme precipitation indicator d_t at the corresponding catchment and valid time. The paired samples are then grouped by lead time for the subsequent analysis.

2.4 Quantification of responses to reanalysis errors

The ANCOVA method is utilized to quantify the responses of forecast errors to reanalysis errors while accounting
 110 for the influence of extreme precipitation. By incorporating the binary term of extreme precipitation and its interaction with reanalysis error in a single regression, the ANCOVA enables both the baseline response and its change under extreme precipitation to be estimated within a unified model (Gao et al., 2023; Fauchaux et al., 2022). For each catchment and lead time, the model is fitted using the ordinary least squares:



$$ef_{s,l} = \alpha_l + \beta_l er_t + \gamma_l d_t + \theta_l(er_t \times d_t) + \epsilon_{s,l} \quad (4)$$

where α_l is the intercept for non-extreme precipitation conditions; β_l represents the response coefficient of forecast errors to reanalysis errors under non-extreme conditions; γ_l represents the offset in the intercept between extreme and non-extreme precipitation conditions; θ_l represents the interaction effect between reanalysis errors and extreme precipitation; and $\epsilon_{s,l}$ is the residual term.

The interaction term determines whether the responses of forecast errors to reanalysis errors differ between the non-extreme and extreme precipitation conditions. Under non-extreme conditions ($d_t = 0$), Eq. (4) reduces to:

$$ef_{s,l} = \alpha_l + \beta_l er_t + \epsilon_{s,l} \quad (5)$$

where the response coefficient is β_l . Under extreme conditions ($d_t = 1$), the model becomes:

$$ef_{s,l} = (\alpha_l + \gamma_l) + (\beta_l + \theta_l)er_t + \epsilon_{s,l} \quad (6)$$

where the corresponding response coefficient becomes $\beta_l + \theta_l$. The estimated coefficients are interpreted as measures of how forecast errors respond to reanalysis errors rather than strictly causal evidence of error propagation.

To quantify the goodness-of-fit for the ANCOVA method, the coefficient of determination (R^2) is calculated (Wang et al., 2026; Over et al., 2025), as follows:

$$R^2 = 1 - \frac{\sum (ef_{s,l} - \widehat{ef}_{s,l})^2}{\sum (ef_{s,l} - \overline{ef}_l)^2} \quad (7)$$

where $\widehat{ef}_{s,l}$ and $\overline{ef}_l$ denote the fitted value and mean value of $ef_{s,l}$, respectively. Statistical significance of the estimated coefficients is evaluated using two-sided t-tests at the significance level of 0.05 (Fauchaux et al., 2022; McEachran et al., 2021). Specifically, the significance of β_l is tested to assess whether the response of forecast errors to reanalysis errors is significant under non-extreme conditions. The significance of θ_l is tested to determine whether extreme precipitation significantly modifies the corresponding response coefficient. The significance of γ_l is tested to determine whether extreme precipitation introduces a significant additional offset in forecast errors after accounting for reanalysis errors.

3 Case study

3.1 Forecast, reanalysis and observation data

The forecast, reanalysis and observation data used in this study are summarized in Table 1. Precipitation forecasts are generated by running GraphCast autoregressively. GraphCast is an AI weather model based on the graph neural network and trained by ERA5 data from 1979 to 2017 (Lam et al., 2023). Initialized with ERA5 data of the two preceding time steps, GraphCast produces global forecasts of 6 atmospheric variables at 37 pressure levels and 5 surface variables. For the evaluation period from 2020 to 2025, GraphCast is initialized once daily at 00:00 UTC to generate 10d forecasts at 6h temporal resolution on two NVIDIA A800 GPUs, each with 80 GB of memory.



140 Precipitation reanalysis data are obtained from the ERA5 dataset (Hersbach et al., 2020). It combines observations with numerical model simulations through data assimilation to provide the spatially and temporally complete representation of the global atmospheric, ocean-wave and land-surface variables. In ERA5, precipitation is a diagnostic variable rather than a directly assimilated quantity and is derived from the forecast model component of the reanalysis system. Its accuracy is therefore influenced by the assimilated atmospheric state, model physics and parameterization schemes.

145 Precipitation observations are sourced from the Daymet dataset (Thornton et al., 2022). It provides long-term, continuous, gridded estimates of daily meteorological variables over continental North America, Hawaii and Puerto Rico by interpolating and extrapolating ground-based observations. With a spatial resolution of approximately $1\text{km} \times 1\text{km}$, Daymet provides a high-resolution observation-based reference dataset that has been widely used for the calibration and evaluation of weather and climate products (Eldardiry et al., 2025; Wootten et al., 2024; Kim et al., 2024).

150 GraphCast forecasts and ERA5 reanalysis are aggregated to daily temporal resolution and Daymet precipitation is resampled to $0.25^\circ \times 0.25^\circ$ grid using bilinear interpolation (Li and Zhao, 2026a; Price et al., 2025). Daymet daily precipitation is defined over accumulation windows ending at approximately 7:00 local standard time for most stations (Thornton et al., 2021). To achieve the closest temporal alignment, GraphCast and ERA5 precipitation data are aggregated over 24h periods ending at 12:00 UTC. Because each daily forecast requires a complete 24-h accumulation window, this aggregation reduces the maximum lead time of GraphCast forecasts from 10 to 9 days. The gridded precipitation data are subsequently averaged to catchment-mean values based on the 1426 catchments in the CAMELS-SPAT dataset (Knoben et al., 2025). For each catchment, Daymet precipitation from 1980 to 2017 is used to determine the 90th percentile threshold for identifying extreme precipitation events, whereas data from 2020 to 2025 serve as the observational reference for calculating precipitation errors in GraphCast forecasts and ERA5 reanalysis.

160

Table 1. Forecast, reanalysis and observation data used in the case study.

Data source	Spatial resolution	Temporal resolution	Coverage	Time period	Reference
GraphCast	$0.25^\circ \times 0.25^\circ$	6h	Global	2020-2025	(Lam et al., 2023)
ERA5	$0.25^\circ \times 0.25^\circ$	1h	Global	1940-present	(Hersbach et al., 2020)
Daymet V4R1	$1\text{km} \times 1\text{km}$	1d	North America	1980-2025	(Thornton et al., 2022)

3.2 Experimental design

Experiment 1. Distribution of precipitation data and extreme events

165 The spatial patterns and statistical distributions of daily precipitation from GraphCast, ERA5 and Daymet are compared across the 1426 catchments. Mean precipitation and probability density distribution are examined to identify

systematic differences among the three datasets. The intensity and frequency of extreme precipitation during the evaluation are then analysed. This experiment provides the characterization of precipitation climatology and extreme precipitation.

Experiment 2. Relationship between forecast errors and reanalysis errors

Mean GraphCast forecast errors and ERA5 reanalysis errors are compared across catchments and lead times. The comparisons are performed for all precipitation samples and for extreme events to assess their spatial patterns and systematic biases. Pearson correlation coefficients between forecast and reanalysis errors are then calculated for each catchment and lead time. This experiment examines how their statistical association changes as lead time increases.

Experiment 3. Responses of forecast errors to reanalysis errors

To provide the catchment-scale interpretation of the ANCOVA method, the catchment USA_14137000 is selected for detailed analysis. Time series of precipitation from GraphCast, ERA5 and Daymet are compared across lead times to examine their temporal correspondence under non-extreme and extreme precipitation conditions. The ANCOVA method is then applied to quantify the response of forecast errors to reanalysis errors.

Experiment 4. Comparison across different lead times and catchments

The ANCOVA method is fitted separately for each of the 1426 catchments at each lead time from 1d to 9d. The statistical and spatial distributions of the estimated coefficients are compared to assess how the responses vary with lead time and across catchments. Differences between non-extreme and extreme precipitation conditions are also examined. This experiment evaluates the consistency and applicability of the ANCOVA method across a range of lead times and catchments.

4 Results

4.1 Distribution of precipitation data and extreme events

The spatial and statistical distributions of daily precipitation across the 1426 catchments from 2020 to 2025 are shown in Fig. 2. Overall, GraphCast forecasts, ERA5 reanalysis and Daymet observations exhibit broadly consistent spatial patterns. Higher mean precipitation occurs along the Pacific coast and across parts of the southeastern United States, whereas drier conditions are observed over much of the interior western region. GraphCast forecasts reproduce these large-scale spatial patterns at both 1d and 5d lead times, although deviations from Daymet become more pronounced when the lead time increases. The probability density functions further reveal systematic discrepancies in precipitation intensity. All datasets show similar distributions for low to moderate daily precipitation below 20 mm/d, but their differences increase toward the upper tail. Daymet exhibits the highest probability density for daily precipitation exceeding 50 mm/d, while GraphCast shows a lighter upper tail, particularly at the 5d lead time. ERA5 generally lies between Daymet and GraphCast over much of the high-intensity range. These results suggest that GraphCast captures the overall precipitation climatology but tends to underestimate the frequency of heavy precipitation, with this discrepancy increasing with lead time.

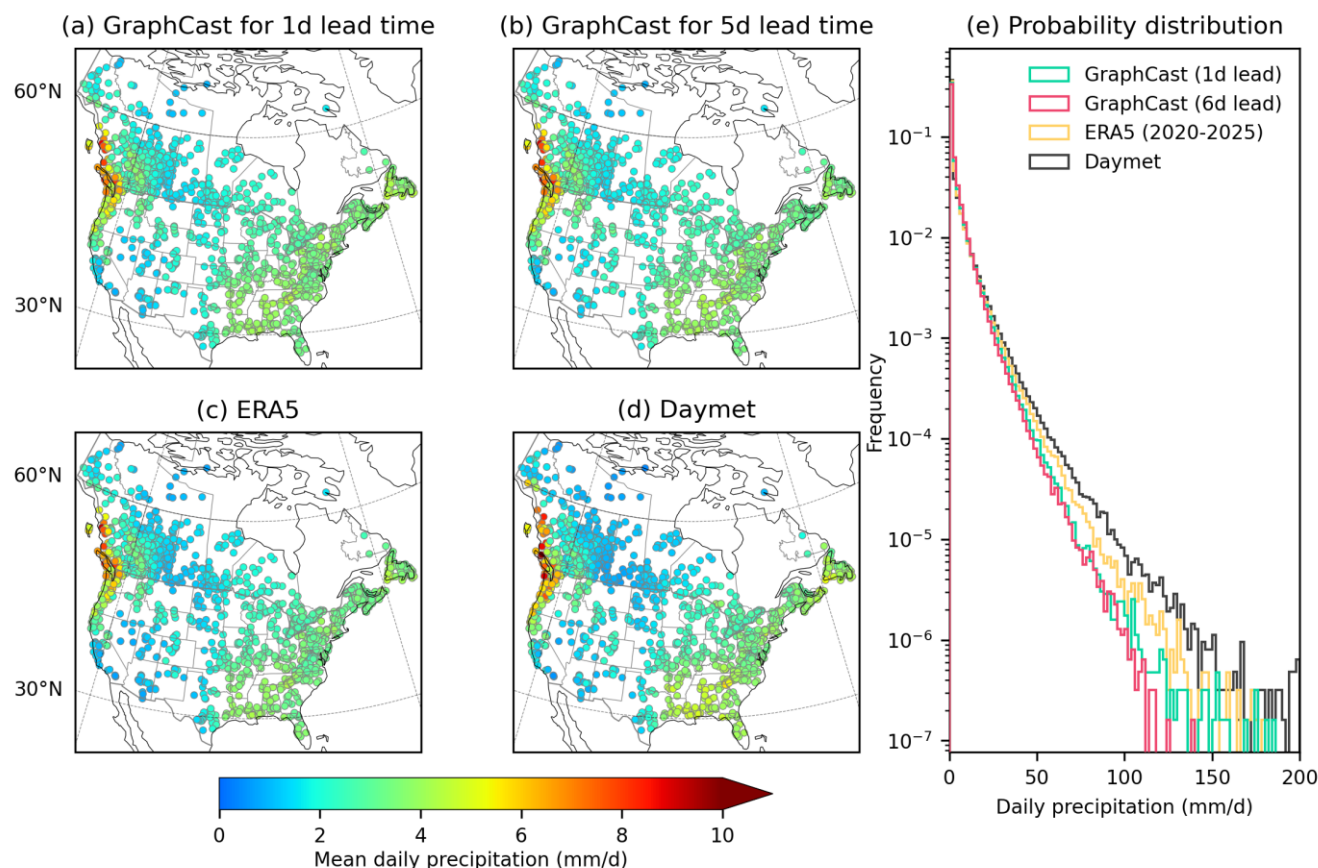


Figure 2. Spatial and statistical distributions of daily precipitation across 1426 catchments from 2020 to 2025. Spatial distribution of mean daily precipitation from (a-b) GraphCast forecasts at 1d and 5d lead times, (c) ERA5 and (d) Daymet. (e) Probability density distributions of daily precipitation.

The spatial and statistical distributions of the catchment-specific 90th percentile thresholds and the numbers of extreme precipitation events are illustrated in Fig. 3. The spatial patterns of the thresholds are generally consistent with those of the mean precipitation climatology, with higher values in wetter coastal and southeastern regions and lower values across many interior and northern catchments. The frequency distribution is positively skewed, with thresholds of most catchments being approximately 5–15 mm/d and only 52 catchments exceeding 20 mm/d. Applying these catchment-specific thresholds to the 2020–2025 evaluation period reveals clear spatial heterogeneity in the number of identified extreme precipitation events. More extreme precipitation events are identified in parts of the Pacific Northwest, southeastern United States and several coastal regions, while fewer events are observed in some interior and northern catchments. The frequency distribution is relatively concentrated, with most catchments experiencing approximately 150–250 events and a peak around 200 events. These results indicate that the percentile-based approach retains regional differences in precipitation climatology while providing a comparable criterion for identifying extreme precipitation across the 1426 catchments.



215

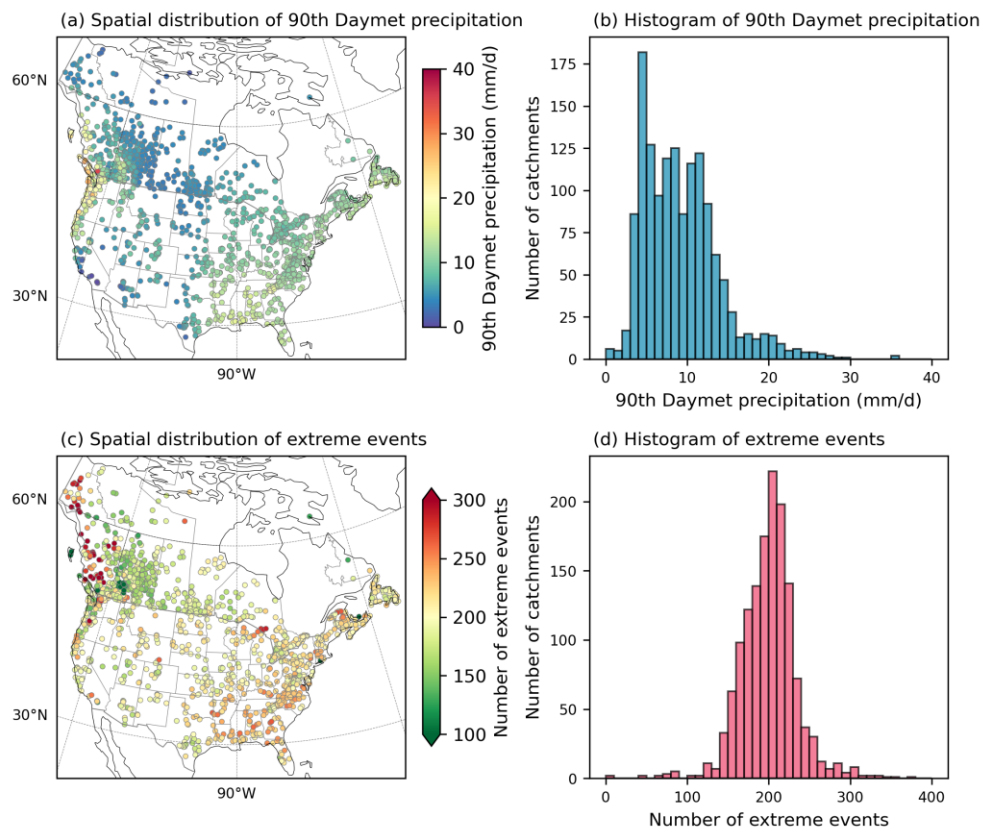


Figure 3. Identification of extreme precipitation events across 1426 catchments. Spatial and statistical distributions of (a-b) catchment-specific 90th percentile thresholds derived from Daymet precipitation during 1980–2017 and (c-d) the numbers of extreme precipitation events identified during 2020–2025.

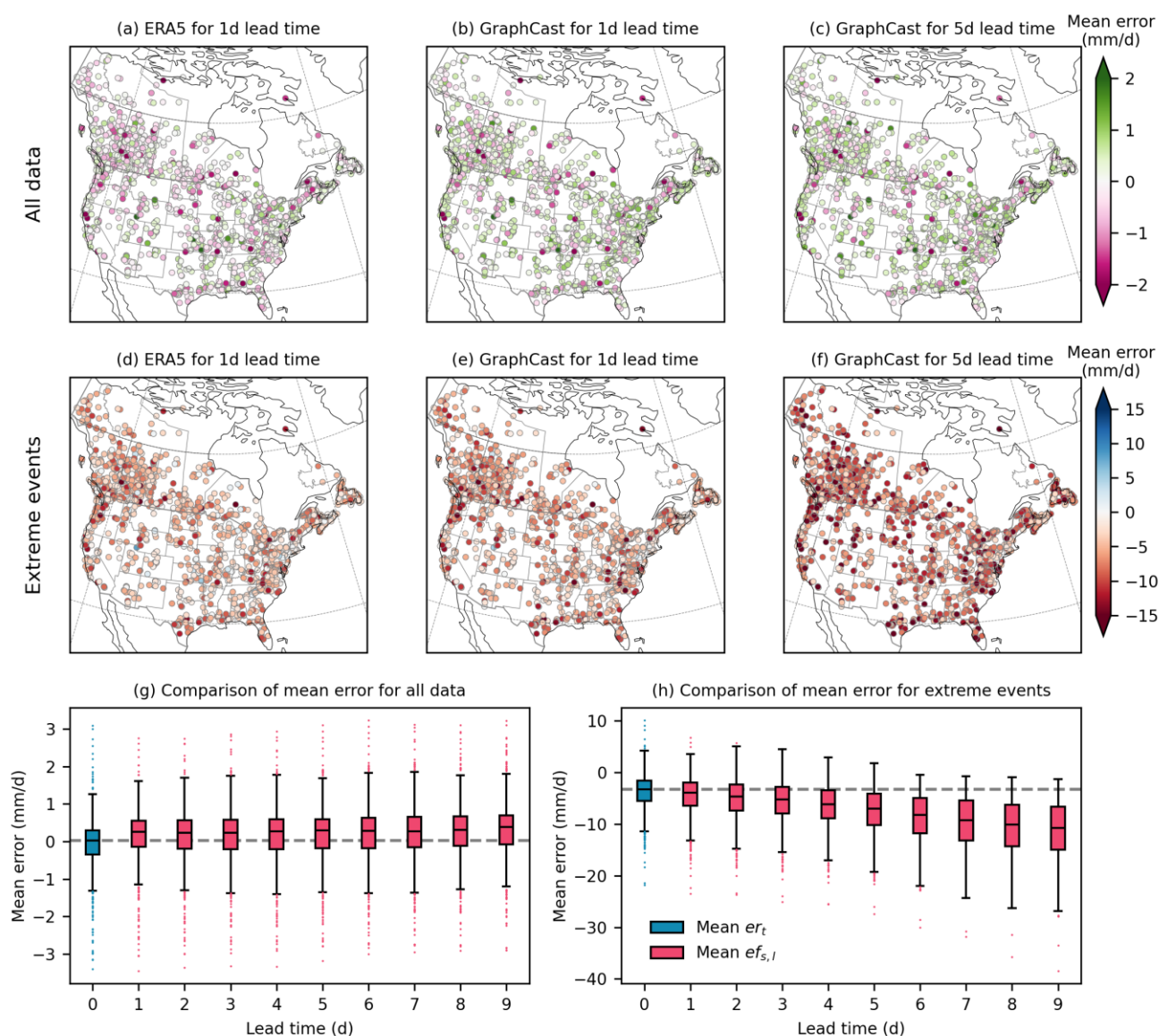
220

4.2 Relationship between forecast and reanalysis errors

The mean precipitation errors of GraphCast forecasts and ERA5 reanalysis from 2020 to 2025 are shown in Fig. 4. When all precipitation samples are considered, both ERA5 and GraphCast exhibit relatively small mean errors across most catchments. Specifically, GraphCast forecasts show a slight tendency toward positive bias and the overall spatial patterns change little between 1d and 5d lead times. The median GraphCast error increases from 0.26 mm/d to 0.38 mm/d when lead time increases from 1d to 9d, while the median ERA5 error is 0.03 mm/d across the 1426 catchments. By contrast, substantially larger negative errors are observed for extreme precipitation events, indicating systematic underestimation by both ERA5 and GraphCast. The spatial coverage and magnitude of negative GraphCast biases increase with lead time, accompanied by a pronounced downward shift in the error distribution. The median GraphCast error decreases from −3.96 mm/d to −10.75 mm/d when lead time increases from 1d to 9d, whereas the median ERA5 error is −3.31 mm/d. These

230

results indicate that precipitation errors differ markedly between the full precipitation samples and extreme events, with the underestimation of extreme precipitation becoming substantially larger for GraphCast as the lead time increases.



235 **Figure 4. Mean precipitation errors of GraphCast forecasts and ERA5 reanalysis during 2020–2025. Spatial distributions of mean ERA5 errors and mean GraphCast errors under (a–c) all precipitation samples and (d–f) extreme precipitation events only. Statistical distributions of mean errors for ERA5 and GraphCast across different lead times under (g) all precipitation samples and (h) extreme precipitation.**



240 The relationships between GraphCast forecast errors and ERA5 reanalysis errors across the 1426 catchments are illustrated in Fig. 5. At the 1d lead time, the two errors are positively correlated across all catchments, with the median Pearson correlation coefficient (r) of 0.84. As the lead time increases, the correlations weaken progressively and become more spatially heterogeneous. Specifically, high correlations are widespread across catchments at 1d lead time, whereas lower and more heterogeneous values are observed at 5d and 9d lead times. The median r across the 1426 catchments
 245 declines to 0.52 and 0.38 at 5d and 9d lead times, respectively. The correlations remain statistically significant across all catchments at 1d and 5d lead times. The correlations remain significant across 1408 catchments, while only 18 catchments show nonsignificant relationships at the 9d lead time. The correlation coefficients are centered above zero across lead times from 1d to 9d. The linear relationships between forecast errors and reanalysis errors weaken steadily with increasing lead time. This result may be due to the increasing contribution of accumulated forecast uncertainty as lead time increases.

250

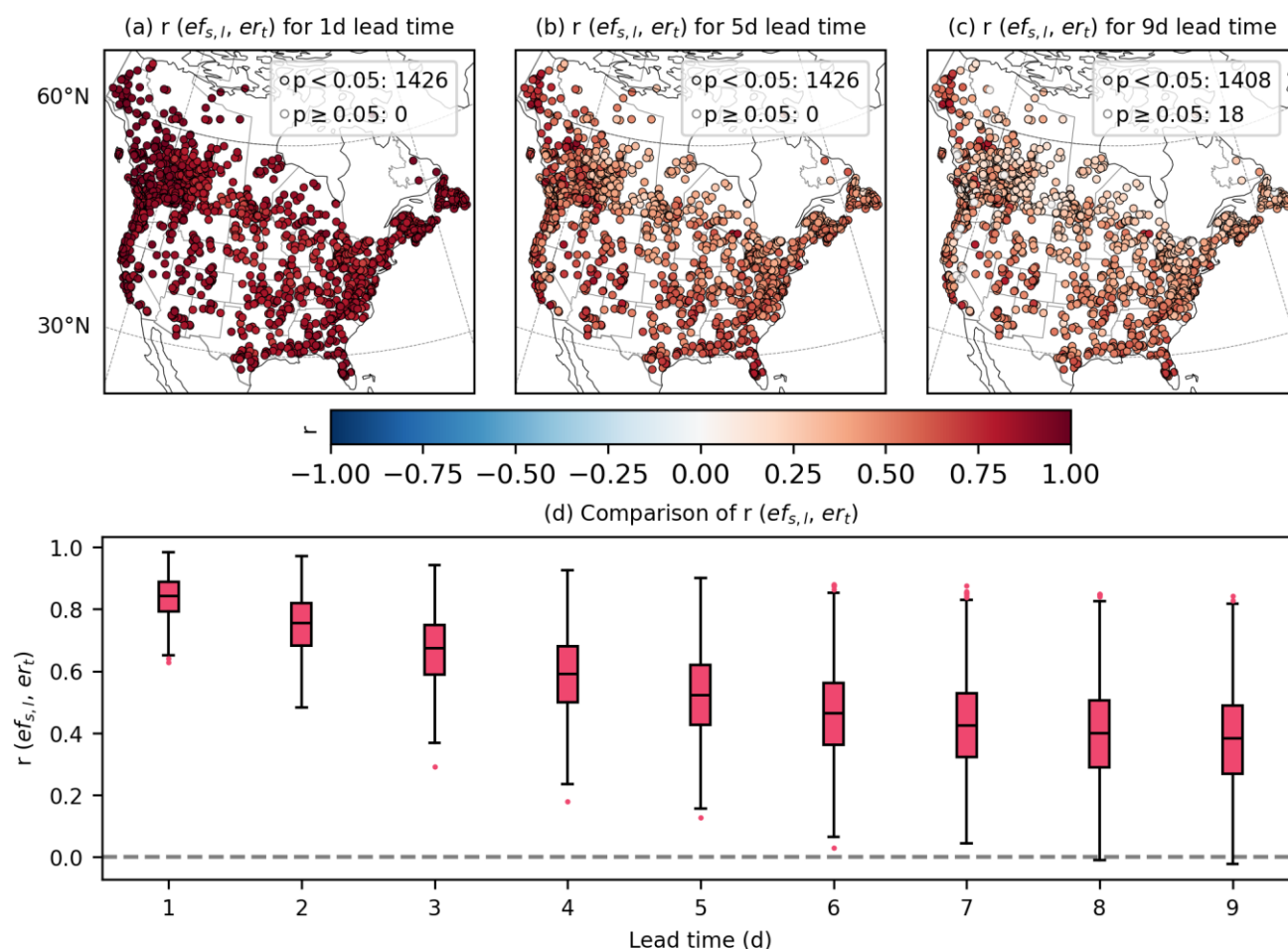


Figure 5. Pearson correlation coefficients (r) between GraphCast forecast errors and ERA5 reanalysis errors. (a-c) Spatial distribution of r for 1d, 5d and 9d lead times. (d) Comparison of r across lead times from 1d to 9d. Filled and open circles indicate



255 statistically significant ($p < 0.05$) and nonsignificant correlations, respectively; inset values show the numbers of corresponding catchments.

4.3 Responses of forecast errors to reanalysis errors

260 The time series of daily precipitation from GraphCast forecasts, ERA5 reanalysis and Daymet observations for catchment USA_14137000 are shown in Fig. 6. At the 1d lead time, GraphCast forecasts generally reproduce the temporal variations observed in ERA5 and Daymet, capturing many precipitation peaks. Discrepancies become more pronounced for events exceeding the 90th percentile threshold of 17.88 mm/d. Both GraphCast forecasts and ERA5 reanalysis frequently underestimate the magnitude of the extreme precipitation identified from Daymet observations. Specifically, the mean ERA5 reanalysis error is -5.32 mm/d for extreme precipitation, while the corresponding mean GraphCast errors are -6.62 mm/d and -6.59 mm/d at lead times of 1d and 3d, respectively. As the lead time increases, the underestimation of extreme precipitation by GraphCast forecasts becomes progressively larger. The mean forecast error decreases to -8.65 , -11.73 and -14.63 mm/d at 5d, 7d and 9d lead times, respectively. In addition to the increasing intensity errors, isolated GraphCast precipitation peaks occur when only little precipitation is observed in Daymet. These results show that the deterioration of GraphCast forecasts with increasing lead time reflects both increasing errors in precipitation magnitude and weakening temporal correspondence in event occurrence.

270

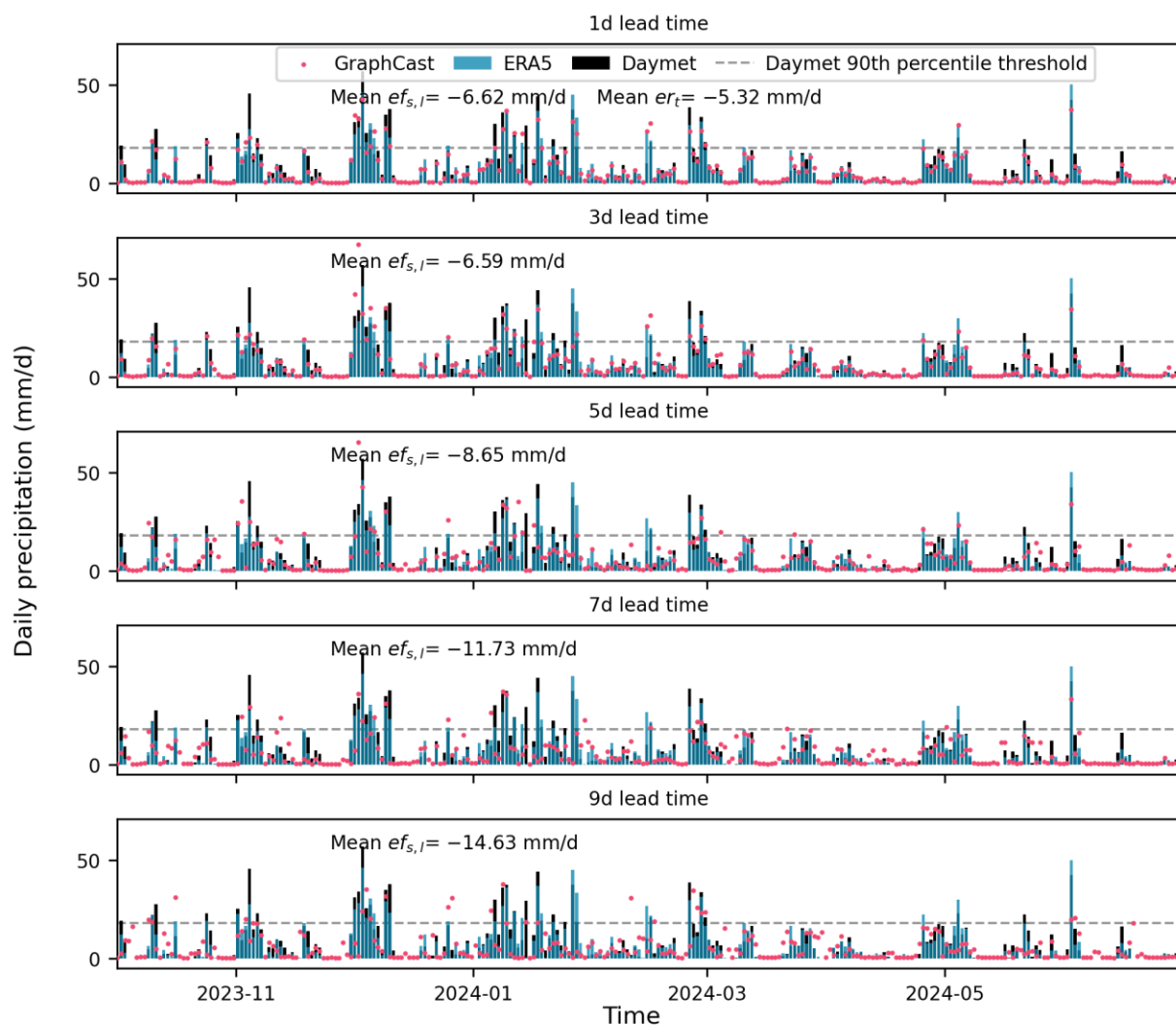


Figure 6. Time series of daily precipitation from GraphCast forecasts, ERA5 reanalysis and Daymet observations for catchment USA_14137000 at lead times of 1d, 3d, 5d, 7d and 9d. The $Mean\ ef_{s,l}$ represents the mean forecast error for extreme precipitation from 2020 to 2025 and the $Mean\ er_t$ represents the corresponding mean reanalysis error.

275

The responses of GraphCast forecast errors to ERA5 reanalysis errors for catchment USA_14137000 are quantified using the ANCOVA method, as shown in Fig. 7. Overall, the responses weaken as lead time increases, accompanied by increasing negative bias associated with extreme precipitation. At the 1d lead time, the ANCOVA method explains 80% of the variance in forecast errors, with the fitted values closely following the 1:1 line. The goodness-of-fit generally declines with increasing lead time, with R^2 decreasing from 0.80 at 1d to 0.29 at 8d and 9d lead times. Under non-extreme precipitation conditions, the response coefficient (β_l) remains positive and statistically significant ($p < 0.01$) across lead times. This coefficient generally decreases from 0.81 at 1d to 0.47 at 9d lead time, indicating that the responses of forecast errors to

280

reanalysis errors weaken with increasing lead time. By contrast, the direct effect of extreme precipitation (γ_l) is negative and statistically significant. Its magnitude decreases from -1.78 at 1d to -12.28 at 9d lead time, showing an increasingly negative offset in forecast errors during extreme precipitation events. The interaction coefficient (θ_l) is generally positive, indicating steeper response to reanalysis errors under extreme precipitation, although its statistical significance varies across lead times.

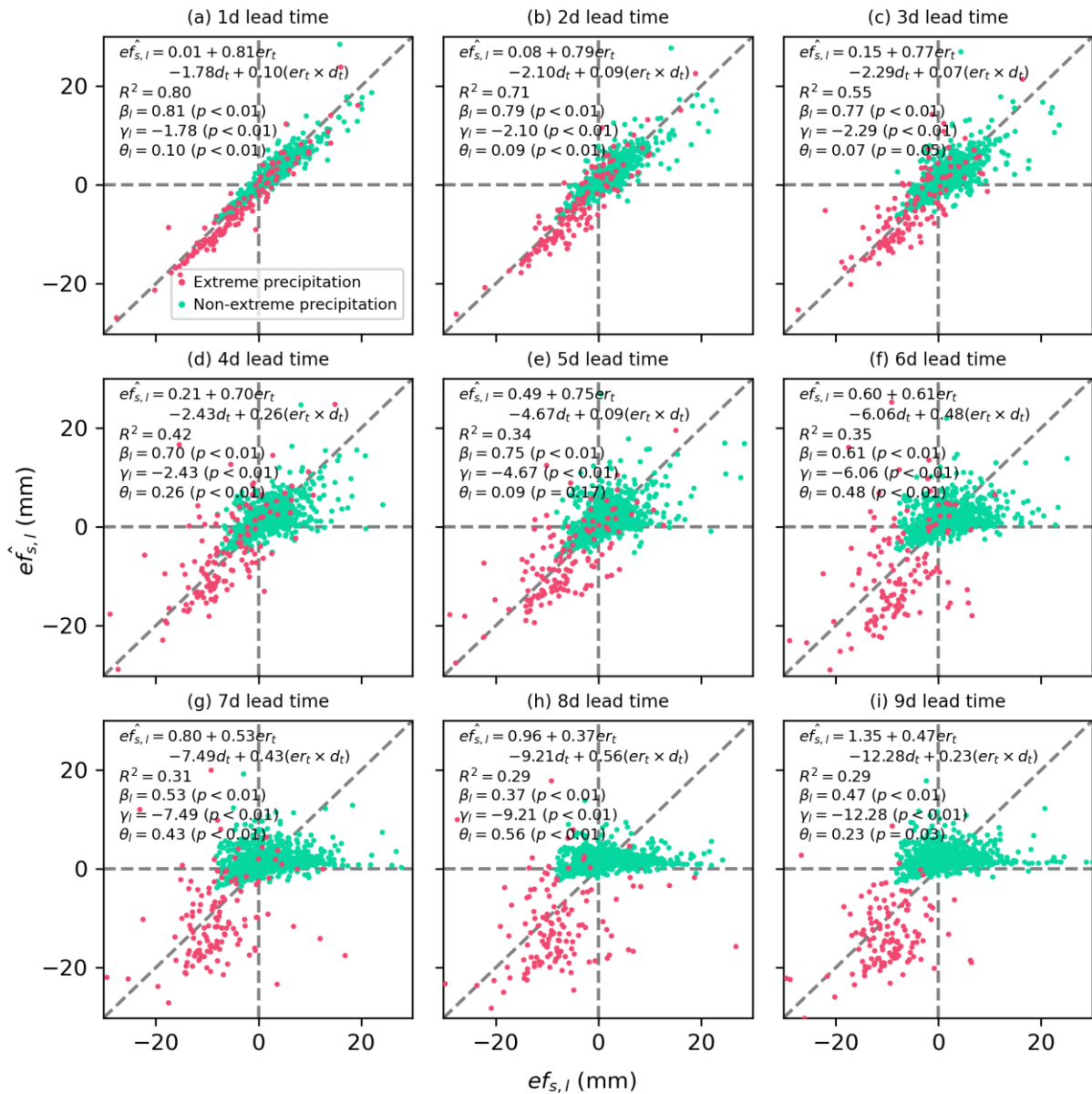
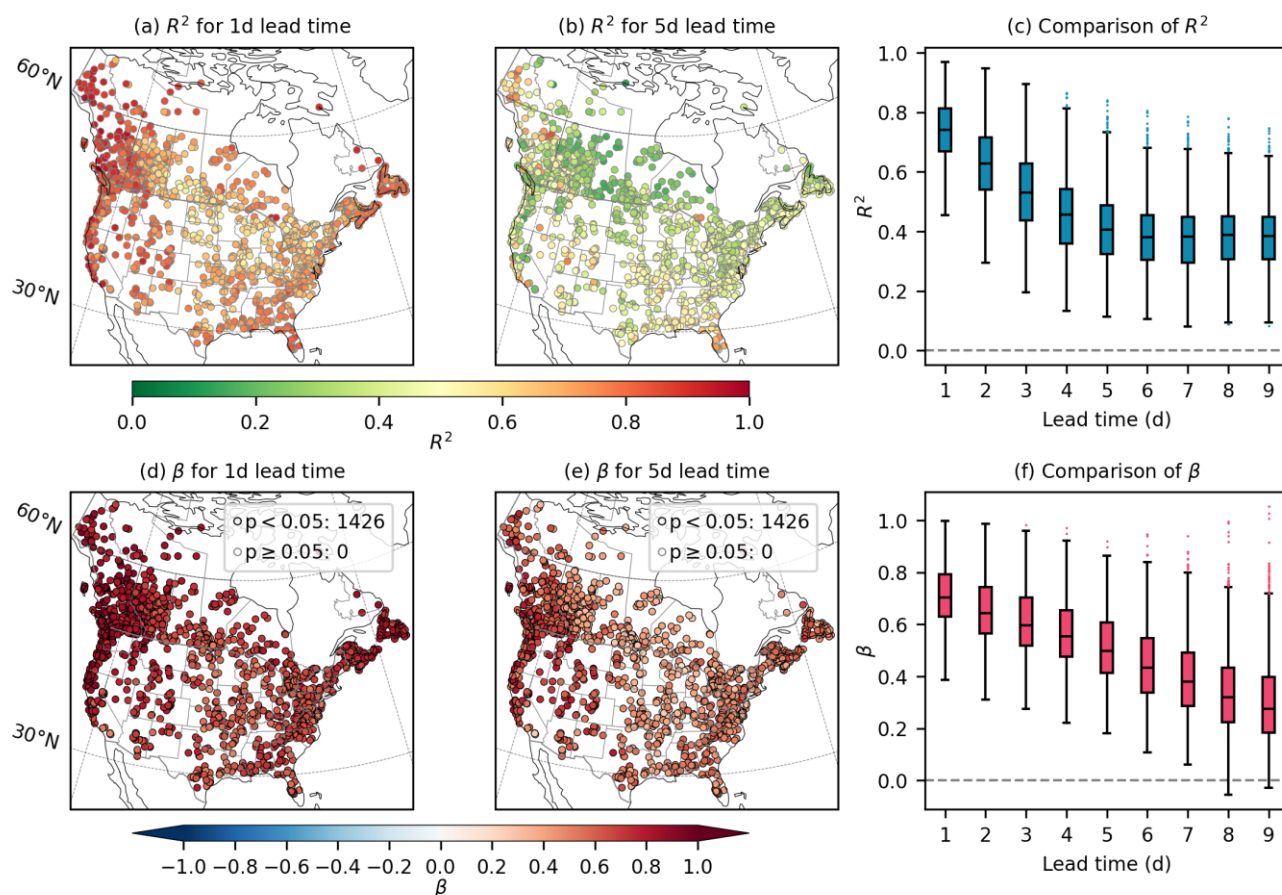


Figure 7. Quantified responses of forecast errors to reanalysis errors for catchment USA_14137000 at lead times from 1d to 9d.

4.4 Comparison across different lead times and catchments

The R^2 values of the ANCOVA method and response coefficients to reanalysis errors under non-extreme conditions across lead times from 1d to 9d and 1426 catchments are shown in Fig. 8. Overall, both the proportion of variance explained by the ANCOVA method and the estimated response coefficients decrease with increasing lead time. At the 1d lead time, R^2 is relatively high across most catchments, with the median value of 0.74, indicating that the ANCOVA method explains a substantial proportion of the variance in forecast errors. The median R^2 declines rapidly to 0.41 by the 5d lead time and then changes slightly, reaching to 0.38 at 9d lead time. Under non-extreme precipitation conditions, the response coefficient (β_l) is positive and statistically significant across all catchments at 1d and 5d lead times, indicating consistent positive responses of forecast errors to reanalysis errors. The median coefficient decreases steadily from 0.70 at 1d lead time to 0.27 at 9d, indicating that, under non-extreme conditions, a 1 mm/d increase in ERA5 precipitation errors leads to a median increase of 0.70 mm/d and 0.27 mm/d in GraphCast precipitation errors across the 1426 catchments, respectively.





305 **Figure 8. Responses of precipitation forecast errors quantified by the ANCOVA method across 1426 catchments. (a-c) R^2 of the ANCOVA method and (d-f) response coefficients to reanalysis errors under non-extreme conditions across lead times from 1d to 9d.**

The effects of extreme precipitation on the responses of GraphCast forecast errors to ERA5 reanalysis errors are illustrated across lead times from 1d to 9d and 1426 catchments, as shown in Fig. 9. Overall, extreme precipitation modifies both the magnitude of forecast errors and their responses to reanalysis errors. Under extreme precipitation, the response coefficient ($\beta_l + \theta_l$) remains predominantly positive but decreases with increasing lead time. Its median value decreases from 0.76 to 0.45 when the lead time increases from 1d to 9d, indicating that a 1 mm/d increase in ERA5 errors leads to median increases of 0.76 mm/d and 0.45 mm/d in GraphCast errors, respectively. The increasing spread of $\beta_l + \theta_l$ across the 1426 catchments further indicates larger spatial heterogeneity with lead time. The median interaction coefficient (θ_l) increases from 0.04 to 0.16, suggesting that extreme precipitation generally steepens the responses of forecast errors to reanalysis errors relative to non-extreme conditions. This effect varies across catchments and is not always statistically significant. The direct effect of extreme precipitation (γ_l), representing the offset in intercept relative to non-extreme conditions, is negative. The median decreases from -1.75 mm/d at 1d lead time to -10.43 mm/d at 9d. For a given reanalysis error, the increasingly negative offsets indicate that the underestimation of extreme precipitation by GraphCast is amplified with lead time.

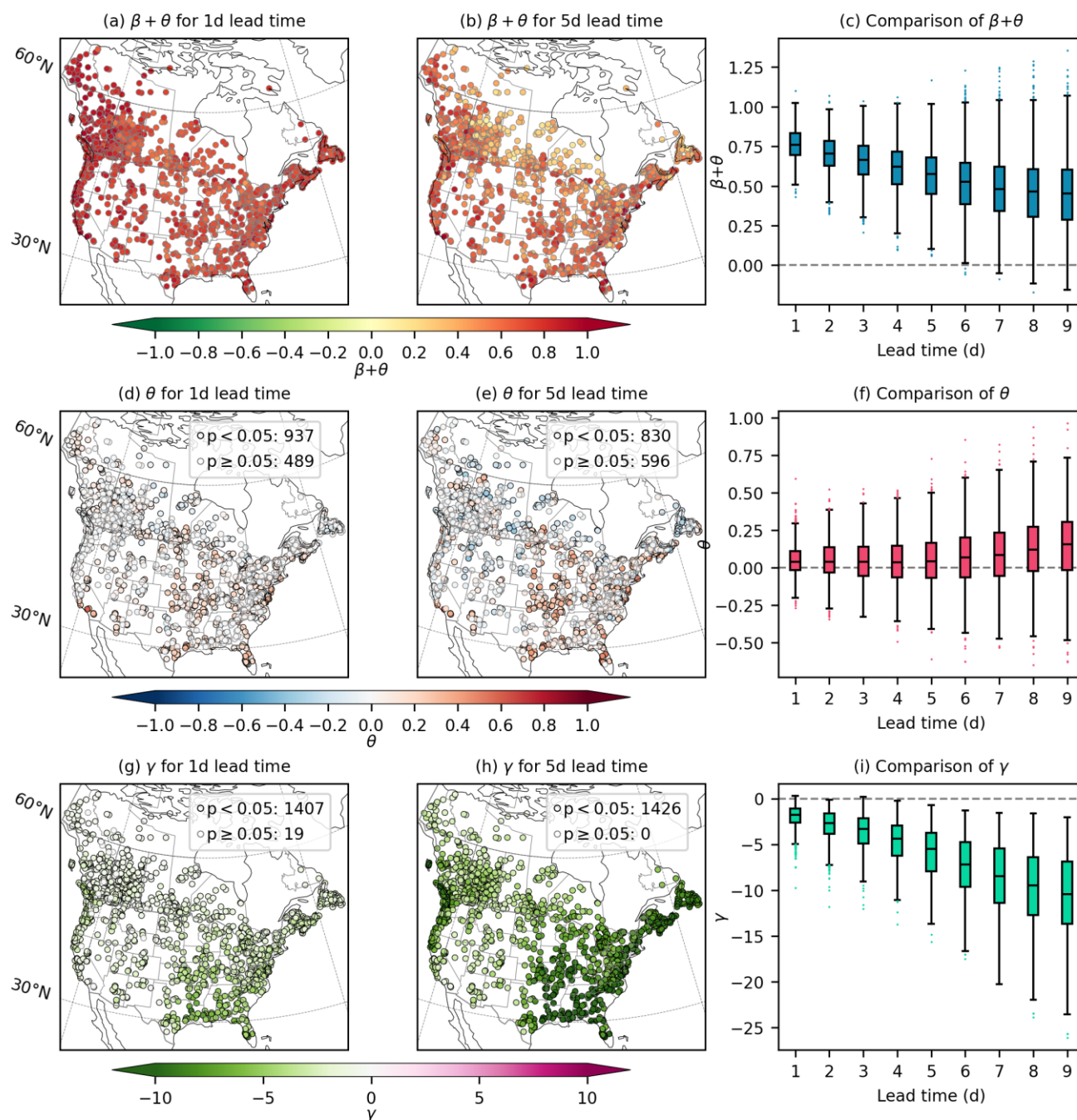


Figure 9. Effects of extreme precipitation events on the responses of GraphCast forecast errors to ERA5 reanalysis errors. (a-c) Response coefficients, (d-f) interaction effects and (g-i) direct effects of extreme precipitation across different lead times.



5 Discussion

Precipitation forecast errors generally increase and become more spatially heterogeneous with increasing lead time, particularly for extreme events (Oh et al., 2026; Li and Zhao, 2026a; Zhang et al., 2026). This paper shows that the estimated responses of GraphCast forecast errors to ERA5 reanalysis errors weaken progressively with increasing lead time. These responses are consistent with the fact that the general large-scale atmospheric relationships represented in ERA5 are learned by GraphCast (Gentile et al., 2026), together with systematic discrepancies relative to reference data (Jin et al., 2024; Gupta et al., 2026). As the lead time increases, other sources of forecast errors, including accumulated atmospheric uncertainty, displacement of precipitation systems and errors in event timing, become more important (Radford et al., 2025; Selz and Craig, 2023). Their increasing contribution can reduce the proportion of GraphCast forecast errors explained by ERA5 reanalysis errors. Under extreme precipitation, the increasingly negative offsets persist in forecast errors after accounting for reanalysis errors, indicating amplified underestimation by GraphCast with increasing lead time (Zhao et al., 2025; Oh et al., 2026). Such additional underestimation is related to the limited representation of localized convective and orographic processes, the scarcity of extreme precipitation events in the training dataset and the spatial smoothing inherent in AI weather models (Baño-Medina et al., 2025a; Liu et al., 2024; Gentile et al., 2026).

Reanalysis datasets are indispensable for training and evaluating AI weather models, yet errors inherent in these datasets can complicate the attribution and interpretation of forecast errors (Abel et al., 2026; Mu et al., 2026; Rasp et al., 2024). The ANCOVA method provides a computationally efficient method to separate the response of forecast errors to reanalysis errors under non-extreme conditions, the intercept offset under extreme precipitation and the corresponding change in the response coefficient. This decomposition enables to quantify how AI precipitation forecast errors respond to reanalysis errors under different conditions. Compared with initial-condition perturbation experiments (Guo et al., 2026; Baño-Medina et al., 2025b; Bülte et al., 2026) and retraining AI models using alternative datasets (Mu et al., 2026; Sun et al., 2025; Yuval et al., 2026), the ANCOVA method is more computationally efficient because it does not require repeated forecast integrations or model retraining. Although the estimated coefficients of the ANCOVA method do not by themselves establish strict causality (Gao et al., 2023; McEachran et al., 2021; Fauchaux et al., 2022), they provide quantitative evidence consistent with the inheritance and propagation of reanalysis errors into AI weather forecasts (Gentile et al., 2026). In addition, the assumption of linear responses may not fully capture the behaviors (Gupta et al., 2015; Gao et al., 2023).

6 Conclusions

This paper has quantified the responses of precipitation errors in AI weather forecasts to errors in climate reanalysis using the ANCOVA method. Specifically, extreme precipitation events are first identified using percentile thresholds derived from historical observations over the training period of the specific AI weather model; forecast and reanalysis errors are then calculated against the observations during the evaluation period; and the ANCOVA method is fitted to quantify the



responses of forecast errors to reanalysis errors under non-extreme and extreme precipitation conditions. For GraphCast precipitation forecasts and ERA5 reanalysis, a case study is devised using Daymet dataset across 1426 catchments from the CAMELS-SPAT dataset. The results show that GraphCast forecasts increasingly underestimate extreme precipitation with lead time and the median of the bias decreases from -3.96 mm/d at 1d lead time to -10.75 mm/d at 9d. The responses of GraphCast errors to ERA5 errors weaken with lead time and the median response coefficients under non-extreme conditions decrease from 0.70 at 1d lead time to 0.27 at 9d. When the lead time is 1d, the median coefficient of 0.70 indicates that a 1 mm/d increase in ERA5 precipitation errors leads to a median increase of 0.70 mm/d in GraphCast precipitation errors across the 1426 catchments. Under extreme precipitation, GraphCast forecasts exhibit additional underestimation after accounting for reanalysis errors, with the median offset decreasing from -1.75 mm/d at 1d lead time to -10.43 mm/d at 9d. Overall, the ANCOVA method contributes to understandings of factors governing AI precipitation forecast errors, highlighting the need to consider inherent uncertainties in climate reanalysis when training and evaluating AI weather models.

Code and data availability

The code and scripts used in this paper are archived in Zenodo under <https://doi.org/10.5281/zenodo.22699488> (Li and Zhao, 2026b). This archive contains the configuration files and scripts used to prepare the GraphCast forecasts, aggregate precipitation to the catchment scale, identify extreme precipitation, calculate forecast and reanalysis errors, fit the ANCOVA method and reproduce all figures. These files are licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0). This archive also contains a frozen copy of the exact GraphCast version used in this study, including its source code, model weights and normalization statistics. The original GraphCast code is available from the project website <https://github.com/google-deepmind/weathernext> under the Apache License, Version 2.0 (Apache 2.0). The model weights and normalization statistics are licensed under the CC BY 4.0.

The GraphCast precipitation forecasts, ERA5 reanalysis, Daymet observations and catchments shapefiles used in this paper are archived in Zenodo under <https://doi.org/10.5281/zenodo.22686404> (Li and Zhao, 2026c). The GraphCast forecasts generated by running the model autoregressively are licensed under the CC BY 4.0. The ERA5 dataset is under the Copernicus data licence (<https://climate.copernicus.eu/sites/default/files/2025-04/licence-to-use-copernicus-products.pdf>, last access: 10 September 2026). The Daymet V4 R1 dataset is openly shared, without restriction, in accordance with the NASA Earth Science Data and Information System (ESDIS) Project Data Use Policy (<https://www.earthdata.nasa.gov/learn/use-data/data-use-policy>, last access: 10 September 2026). The shapefiles of the 1426 catchments in the CAMELS-SPAT are in the U.S. Public Domain (<https://www.usgs.gov/information-policies-and-instructions/copyrights-and-credits>, last access: 10 September 2026) and available under the Open Government License–Canada (<https://open.canada.ca/en/open-government-licence-canada>, last access: 10 September 2026).



390 Author contributions

T.Z. and Q.L. designed the research. Q.L. collected the data and developed the codes. T.Z. and Q.L. discussed and interpreted the results. Q.L. wrote the first draft of the paper. T.Z. supervised the research and revised the manuscript.

395 Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China and the Guangdong Provincial Department of
 400 Science and Technology.

Financial support

This paper is supported by the National Natural Science Foundation of China (2023YFF0804900 and 52379033).

405 References

- Abbas, A., Yang, Y., Pan, M., Trambly, Y., Shen, C., Ji, H., Gebrechorkos, S. H., Pappenberger, F., Pyo, J., Feng, D., Huffman, G., Nguyen, P., Massari, C., Brocca, L., Tan, J., and Beck, H. E.: Comprehensive Global Assessment of 24 Gridded Precipitation Datasets Across 18428 Catchments Using Hydrological Modeling, *Hydrol. Earth Syst. Sci.*, 30, 3399–3423, <https://doi.org/10.5194/hess-30-3399-2026>, 2026.
- 410 Abel, M. R., Thompson, A. J., Gutmann, E. D., Mahoney, K., McCrary, R. R., Schumacher, R. S., and Slivinski, L. C.: Reanalyses in the Age of Machine Learning: Why Dataset Curation Matters Now More than Ever, *Bull. Am. Meteorol. Soc.*, 107, E922–E931, <https://doi.org/10.1175/BAMS-D-25-0149.1>, 2026.
- Alet, F., Andersson, T. R., Price, I., Markou, S., El-Kadi, A., Masters, D., Li, A., Merchant, S., Williams, N., Thornton, G., MacKay, K., Graham, O., Uddin, A., Gaiarin, B., Shah, D., Kruse, E., Hogsett, W., Zelinsky, D., Cangialosi, J.,
 415 Martinez, J., Franklin, J., DeMaria, M., Musgrave, K., Bain, C. L., Titley, H., Stott, J., Lam, R., Bell, A., Komarek, P., Willson, M., Sanchez-Gonzalez, A., and Battaglia, P.: Operational Tropical Cyclone Forecasting with AI, *Nature*, 1–3, <https://doi.org/10.1038/s41586-026-10953-2>, 2026.
- Baño-Medina, J., Sengupta, A., Steinhoff, D., Mulrooney, P., Nipen, T., Santa-Cruz, M., Nie, Y., and Delle Monache, L.: A regional high resolution AI weather model for the prediction of atmospheric rivers and extreme precipitation, *Npj Clim. Atmos. Sci.*, 8, 385, <https://doi.org/10.1038/s41612-025-01265-9>, 2025a.
- 420



- Baño-Medina, J., Sengupta, A., Doyle, J. D., Reynolds, C. A., Watson-Parris, D., and Monache, L. D.: Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia, *Npj Clim. Atmos. Sci.*, 8, 1–9, <https://doi.org/10.1038/s41612-025-00949-6>, 2025b.
- Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q.: Accurate medium-range global weather forecasting with 3D neural networks, *Nature*, 619, 533–538, <https://doi.org/10.1038/s41586-023-06185-3>, 2023.
- 425 Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., Gupta, J. K., Thambiratnam, K., Archibald, A. T., Wu, C.-C., Heider, E., Welling, M., Turner, R. E., and Perdikaris, P.: A foundation model for the Earth system, *Nature*, 641, 1180–1187, <https://doi.org/10.1038/s41586-025-09005-y>, 2025.
- 430 Bülte, C., Horat, N., Quinting, J., and Lerch, S.: Uncertainty Quantification for Data-Driven Weather Models, *Artif. Intell. Earth Syst.*, 5, <https://doi.org/10.1175/AIES-D-24-0049.1>, 2026.
- Chen, K., Han, T., Ling, F., Gong, J., Bai, L., Wang, X., Luo, J.-J., Fei, B., Zhang, W., Chen, X., Ma, L., Zhang, T., Su, R., Ci, Y., Li, B., Yang, X., and Ouyang, W.: The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time, *Commun. Earth Environ.*, 6, 1–10, <https://doi.org/10.1038/s43247-025-02502-y>, 2025.
- 435 Eldardiry, H., Sun, N., Yan, H., Reed, P., Thurber, T., and Rice, J.: Characterizing How Meteorological Forcing Selection and Parameter Uncertainty Influence Community Land Model Version 5 Hydrological Applications in the United States, *J. Adv. Model. Earth Syst.*, 17, e2024MS004222, <https://doi.org/10.1029/2024MS004222>, 2025.
- Faucheux, N. M., Sample, A. R., Aldridge, C. A., Norris, D. M., Owens, C., Starnes, V. R., VanderBloemen, S., and 440 Miranda, L. E.: Reservoir Attributes Display Cascading Spatial Patterns Along River Basins, *Water Resour. Res.*, 58, e2021WR029910, <https://doi.org/10.1029/2021WR029910>, 2022.
- Gao, Y., Sahin, A., and Vrugt, J. A.: Probabilistic Sensitivity Analysis With Dependent Variables: Covariance-Based Decomposition of Hydrologic Models, *Water Resour. Res.*, 59, e2022WR032834, <https://doi.org/10.1029/2022WR032834>, 2023.
- 445 Gentile, E. S., Hunt, K. M. R., Tomassini, L., Harvey, B., and Martinez-Alvarado, O.: Global Diurnal Precipitation Cycle in the AI Model GraphCast and a 5-km Unified Model: Challenges and Opportunities, *Geophys. Res. Lett.*, 53, e2025GL120961, <https://doi.org/10.1029/2025GL120961>, 2026.
- Guo, A., Mu, M., Qin, B., Pu, J., and Li, H.: A nonlinear explainable method for data-driven forecasting models in Earth system science, *Npj Clim. Atmos. Sci.*, <https://doi.org/10.1038/s41612-026-01464-y>, 2026.
- 450 Gupta, A., Sheshadri, A., and Suri, D.: MAUSAM: An Observations-Focused Assessment of Global AI Weather Prediction Models During the South Asian Monsoon, *J. Adv. Model. Earth Syst.*, 18, e2025MS005568, <https://doi.org/10.1029/2025MS005568>, 2026.
- Gupta, S. C., Kessler, A. C., Brown, M. K., and Zvomuya, F.: Climate and agricultural land use change impacts on streamflow in the upper midwestern United States, *Water Resour. Res.*, 51, 5301–5317, <https://doi.org/10.1002/2015WR017323>, 2015.
- 455 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, *Q. J. R. Meteorol. Soc.*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- 460 Huang, Z. and Zhao, T.: How suitable are copula models for post-processing global precipitation forecasts?, *J. Hydrol.*, 656, 133005, <https://doi.org/10.1016/j.jhydrol.2025.133005>, 2025.
- Jiang, Q., Li, W., Fan, Z., He, X., Sun, W., Chen, S., Wen, J., Gao, J., and Wang, J.: Evaluation of the ERA5 reanalysis precipitation dataset over Chinese Mainland, *J. Hydrol.*, 595, 125660, <https://doi.org/10.1016/j.jhydrol.2020.125660>, 2021.
- 465 Jin, W., Weyn, J., Zhao, P., Xiang, S., Bian, J., Fang, Z., Dong, H., Sun, H., Thambiratnam, K., and Zhang, Q.: WeatherReal: A Benchmark Based on In-Situ Observations for Evaluating Weather Models, <https://doi.org/10.48550/arXiv.2409.09371>, 14 September 2024.



- 470 Kim, T., Villarini, G., Done, J. M., Johnson, D. R., Prein, A. F., and Wang, C.: Dominant Sources of Uncertainty for
Downscaled Climate: A Military Installation Perspective, *J. Geophys. Res. Atmos.*, 129, e2024JD040935,
<https://doi.org/10.1029/2024JD040935>, 2024.
- Knoben, W. J. M., Thébault, C., Keshavarz, K., Torres-Rojas, L., Chaney, N. W., Pietroniro, A., and Clark, M. P.:
475 Catchment Attributes and MEteorology for Large-Sample SPATially distributed analysis (CAMELS-SPAT):
streamflow observations, forcing data and geospatial data for hydrologic studies across North America, *Hydrol. Earth
Syst. Sci.*, 29, 5791–5833, <https://doi.org/10.5194/hess-29-5791-2025>, 2025.
- Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirmsberger, 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., and Battaglia, P.:
480 Learning skillful medium-range global weather forecasting, *Science*, 382, 1416–1421,
<https://doi.org/10.1126/science.adi2336>, 2023.
- Li, Q. and Zhao, T.: Code for “Quantifying the responses of AI precipitation forecast errors to reanalysis errors using
analysis of covariance,” <https://doi.org/10.5281/zenodo.22699488>, 2026a.
- Li, Q. and Zhao, T.: Data for “Quantifying the responses of AI precipitation forecast errors to reanalysis errors using analysis
of covariance,” <https://doi.org/10.5281/zenodo.22686404>, 2026b.
- 485 Li, Q. and Zhao, T.: Intercomparison of AI forecasts and numerical weather predictions for the July 2021 extreme
precipitation event in Henan, China, *Stoch. Environ. Res. Risk Assess.*, 40, 141, <https://doi.org/10.1007/s00477-026-03271-6>, 2026c.
- Liu, C.-C., Hsu, K., Chen, D.-S., Chang, P.-L., Hsiao, L.-F., Fong, C.-T., Hong, J.-S., Cheng, C.-P., Lu, K.-C., Chen, C.-R.,
and Kuo, H.-C.: Evaluation of five global AI models for predicting weather in Eastern Asia and Western Pacific, *Npj
Clim. Atmos. Sci.*, 7, 1–12, <https://doi.org/10.1038/s41612-024-00769-0>, 2024.
- 490 McEachran, Z. P., Karwan, D. L., Sebestyen, S. D., Slesak, R. A., and Ng, G.-H. C.: Nonstationary flood-frequency analysis
to assess effects of harvest and cover type conversion on peak flows at the Marcell Experimental Forest, Minnesota,
USA, *J. Hydrol.*, 596, 126054, <https://doi.org/10.1016/j.jhydrol.2021.126054>, 2021.
- Meng, H. and Zhao, T.: Evaluation of the hydrological utility of the GPM IMERG satellite precipitation products, *Atmos.
Res.*, 322, 108139, <https://doi.org/10.1016/j.atmosres.2025.108139>, 2025.
- 495 Mu, M., Qin, B., and Dai, G.: Predictability Study of Weather and Climate Events Related to Artificial Intelligence Models,
Adv. Atmos. Sci., 42, 1–8, <https://doi.org/10.1007/s00376-024-4372-7>, 2025.
- Mu, M., Qin, B., and Dai, G.: AI Model Performances Depend on Adequate Training Data Representing Diverse Physical
Mechanisms, *Geophys. Res. Lett.*, 53, e2025GL121196, <https://doi.org/10.1029/2025GL121196>, 2026.
- 500 Nie, Y., Sengupta, A., Baño-Medina, J., Moore, B. J., and Delle Monache, L.: Diagnosing Forecast Error Propagation and
Large-Scale Dynamics of Weather Extremes with an AI Weather Model, <https://doi.org/10.22541/essoar.15002158/v1>,
2026.
- Oh, S.-G., Bae, Y.-J., Son, S.-W., Hong, D.-C., Kim, J.-Y., Kim, Y., Yoon, H., Yoon, J.-H., Jeong, J.-H., Kim, H., and Lee,
H.: Data-driven forecasts of extreme weather in East Asia: feasibility of operational use, *Weather Clim. Extrem.*, 52,
100875, <https://doi.org/10.1016/j.wace.2026.100875>, 2026.
- 505 Olivetti, L. and Messori, G.: Do data-driven models beat numerical models in forecasting weather extremes? A comparison
of IFS HRES, Pangu-Weather, and GraphCast, *Geosci. Model Dev.*, 17, 7915–7962, <https://doi.org/10.5194/gmd-17-7915-2024>, 2024.
- Over, T., Marti, M., Ortiz, J., and Podzorski, H.: The joint effect of changes in urbanization and climate on trends in floods:
510 A comparison of panel and single-station quantile regression approaches, *J. Hydrol.*, 648, 132281,
<https://doi.org/10.1016/j.jhydrol.2024.132281>, 2025.
- Pan, L., Zhang, H., Liang, M., Li, P., Cao, C., Du, L., and Liu, J.: Comparative Analysis of Precipitation Forecasts between
the ECMWF Artificial Intelligence System (AIFS) and Its Integrated Forecast System (IFS), *Weather Forecast.*, 40,
1653–1670, <https://doi.org/10.1175/WAF-D-24-0227.1>, 2025.
- 515 Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S.,
Battaglia, P., Lam, R., and Willson, M.: Probabilistic weather forecasting with machine learning, *Nature*, 637, 84–90,
<https://doi.org/10.1038/s41586-024-08252-9>, 2025.



- Radford, J. T., Ebert-Uphoff, I., and Stewart, J. Q.: A Comparison of AI Weather Prediction and Numerical Weather Prediction Models for 1–7-Day Precipitation Forecasts, *Weather Forecast.*, 40, 561–575, <https://doi.org/10.1175/WAF-D-24-0081.1>, 2025.
- Rasp, S., Hoyer, S., Meroze, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., and Sha, F.: WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models, *J. Adv. Model. Earth Syst.*, 16, e2023MS004019, <https://doi.org/10.1029/2023MS004019>, 2024.
- Selz, T. and Craig, G. C.: Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect?, *Geophys. Res. Lett.*, 50, e2023GL105747, <https://doi.org/10.1029/2023GL105747>, 2023.
- Sun, Y. Q., Hassanzadeh, P., Zand, M., Chattopadhyay, A., Weare, J., and Abbot, D. S.: Can AI weather models predict out-of-distribution gray swan tropical cyclones?, *Proc. Natl. Acad. Sci.*, 122, e2420914122, <https://doi.org/10.1073/pnas.2420914122>, 2025.
- Tan, J., Huang, Q., and Chen, S.: Deep learning model based on multi-scale feature fusion for precipitation nowcasting, *Geosci. Model Dev.*, 17, 53–69, <https://doi.org/10.5194/gmd-17-53-2024>, 2024.
- Tang, G., Clark, M. P., Papalexiou, S. M., Ma, Z., and Hong, Y.: Have satellite precipitation products improved over last two decades? A comprehensive comparison of GPM IMERG with nine satellite and reanalysis datasets, *Remote Sens. Environ.*, 240, 111697, <https://doi.org/10.1016/j.rse.2020.111697>, 2020.
- Thornton, M. M., Shrestha, R., Wei, Y., Thornton, P. E., Kao, S.-C., and Wilson, B. E.: Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4 R1, , <https://doi.org/10.3334/ORNLDAAAC/2129>, 2022.
- Thornton, P. E., Shrestha, R., Thornton, M., Kao, S.-C., Wei, Y., and Wilson, B. E.: Gridded daily weather data for North America with comprehensive uncertainty quantification, *Sci. Data*, 8, 190, <https://doi.org/10.1038/s41597-021-00973-0>, 2021.
- Wang, F. and Tian, D.: Hourly Evaluation of Eight Gridded Precipitation Datasets over the Contiguous United States: Intercomparison of Satellite, Radar, Reanalysis, and Merged Products, *J. Hydrometeorol.*, 26, 1717–1733, <https://doi.org/10.1175/JHM-D-25-0063.1>, 2025.
- Wang, N., Sun, F., Liu, W., Yang, S., Wang, H., Feng, Y., and Huang, Q.: Climate drives observational changes in hydrological extremes across most global regions, *The Innovation*, 7, 101171, <https://doi.org/10.1016/j.xinn.2025.101171>, 2026.
- Wootten, A. M., Başağaoğlu, H., Bertetti, F. P., Chakraborty, D., Sharma, C., Samimi, M., and Mirchi, A.: Customized Statistically Downscaled CMIP5 and CMIP6 Projections: Application in the Edwards Aquifer Region in South-Central Texas, *Earths Future*, 12, e2024EF004716, <https://doi.org/10.1029/2024EF004716>, 2024.
- Yuval, J., Langmore, I., Kochkov, D., and Hoyer, S.: Neural general circulation models for modeling precipitation, *Sci. Adv.*, 12, <https://doi.org/10.1126/sciadv.adv6891>, 2026.
- Zhang, Z., Fischer, E., Zscheischler, J., and Engelke, S.: Physics-based models outperform AI weather forecasts of record-breaking extremes, *Sci. Adv.*, 12, eaec1433, <https://doi.org/10.1126/sciadv.aec1433>, 2026.
- Zhao, T., Li, Q., Tu, T., and Chen, X.: An extension of WeatherBench 2 to binary hydroclimatic forecasts, *Geosci. Model Dev.*, 18, 5781–5799, <https://doi.org/10.5194/gmd-18-5781-2025>, 2025.