Precip

BENCHMARK WHITEPAPER

Evaluation of Precip's 10-Day Forecast

How Precip's custom forecast blend improves accuracy over the raw output of the world's leading forecast models

Precip's hourly forecast compared with ECMWF IFS HRES

Temperature	Relative humidity	Wind speed	Heavy rain
26%	**24%**	**19%**	**2.9×**
more accurate, days 1–10	more accurate, days 1–10	more accurate, days 1–10	better on downpours, days 1–2

Accuracy is mean absolute error over all 10 days; heavy-rain skill is the critical success index for rain over 4 mm an hour. Scored against approximately 2,200 ASOS airport weather stations across 29 forecast runs; see How the forecast was tested.

Introduction

Precip's hourly forecast powers the 10-day outlook in the Precip app and API, delivering hourly predictions of temperature, humidity, wind, cloud cover, precipitation, and the chance of rain. This report evaluates the production forecast Precip serves against two leading global benchmarks:

- **ECMWF IFS HRES**, the European Centre for Medium-Range Weather Forecasts' high-resolution deterministic model, widely regarded as the most skillful physics-based medium-range forecast.
- **Google WeatherNext 3**, Google DeepMind's global AI ensemble model, evaluated on its raw published output.

Over the contiguous United States (CONUS), Precip builds on WeatherNext 3 with two post-processing stages:

- **Downscaling to 2.5 km.** Three lightweight neural networks combine WeatherNext 3's ensemble mean (its 0.05° (~5 km) station-calibrated fields for temperature and humidity, its 0.1° (~10 km) fields for wind) with local terrain, coastline geometry, coordinates, and time of day to downscale temperature, relative humidity, and wind speed onto the 2.54 km grid of the National Weather Service's Unrestricted Mesoscale Analysis (URMA). This restores sharp local gradients across mountain valleys, coastlines, and urban areas that a 10 km grid smooths out.
- **A multi-model precipitation blend.** For the chance of rain, a statistical blend combines the WeatherNext 3 and ECMWF IFS ensembles across all 10 days, incorporating NOAA's 3 km High-Resolution Rapid Refresh (HRRR) during the first two days. Over those first two days, the blend also sets hourly rainfall amounts, calibrated to match the intensity distribution of observed precipitation.

Key findings:

- **Against IFS HRES**, Precip is more accurate for temperature, relative humidity, and wind speed in every forecast window, verified against both the URMA analysis and surface weather stations. Over the full 10-day horizon, Precip's temperature MAE at stations is 26% lower, and its 10-day temperature forecast is as accurate as IFS HRES's 8-day forecast.
- **Against raw WeatherNext 3**, neural downscaling improves temperature, humidity, and wind speed in every forecast window, while Precip's precipitation blend achieves a lower Brier score for the chance of rain across every window and reference dataset.
- **For heavy rain in the first two days**, Precip's critical success index (CSI) for hourly rainfall exceeding 4 mm at U.S. Climate Reference Network (USCRN) gauges is 3.0× that of IFS HRES and 4.3× that of raw WeatherNext 3.

How the forecast was tested

Precip, IFS HRES, and WeatherNext 3 were evaluated across the same 29 forecast runs, initialized at 00Z and 12Z between 2026-07-27 and 2026-10-08, for every hour out to 240 hours (10 days). Each forecast was verified against three independent observational references:

- **URMA**, the National Weather Service's 2.54 km gridded surface analysis across CONUS, providing complete spatial coverage.
- **ASOS**, approximately 2,200 Automated Surface Observing System airport stations reporting hourly temperature, dew point, humidity, wind, and precipitation.
- **USCRN**, approximately 115 research-grade U.S. Climate Reference Network stations used as a reference for precipitation.

At station locations, each forecast is sampled from its grid at the station's exact coordinates, and all models are verified across the identical set of stations and valid hours. Precipitation is evaluated every hour; temperature, humidity, and wind are evaluated every 3 hours through hour 144 and every 6 hours thereafter.

Results are pooled into four lead-time windows: days 1–2 (1–48 hours), days 3–5 (49–120 hours), days 6–10 (121–240 hours), and the full 10-day horizon (1–240 hours).

Definitions

- **Mean absolute error (MAE)** is the average magnitude of forecast error in the variable's native units. **Root mean square error (RMSE)** weights larger errors more heavily by squaring residuals before averaging. Lower is better for both.
- **Equivalent lead time** ("Precip's 10-day forecast is as accurate as IFS HRES's 8-day") locates Precip's mean absolute error at hour 240 on the IFS HRES error curve (Figure 1), interpolating between 12-hour bins and rounding up to a whole day.
- **Critical success index (CSI)**, or threat score, is the fraction of observed or forecast rain events above a given threshold that were predicted correctly, penalizing both missed events and false alarms: hits ÷ (hits + misses + false alarms). Higher is better.
- **Brier score** is the mean squared error of a probability forecast; lower is better. **Brier skill score (BSS)** measures relative improvement over a climatological baseline that always forecasts the period's observed frequency of rain; higher is better, and 0 indicates no improvement.
- **Bold** marks the best competing forecast in each row.
- **Blue figures** beside Precip's score show how much better it is than the best other forecast in that row (IFS HRES or WeatherNext 3). For error scores (MAE, RMSE, Brier) it is the percentage by which Precip's error is lower; for skill scores (CSI, Brier skill score) it is the percentage by which Precip's score is higher, shown as a multiple such as 3.0× once Precip's score is more than double.
- **URMA analysis** columns show the discrepancy between the 2.54 km URMA grid and point stations at the valid hour: the share of error attributable to comparing a gridded field with a single point sensor. It serves as a representativeness floor rather than a competing forecast and is never bolded.

Temperature, humidity, and wind

Raw WeatherNext 3 already outperforms IFS HRES on surface temperature, relative humidity, and wind speed; downscaling to 2.5 km widens that lead across every forecast window. Against the URMA analysis over the full 10-day horizon, Precip reduces temperature MAE by 31% relative to IFS HRES and by 17% relative to raw WeatherNext 3. At stations, Precip's 10-day forecast matches IFS HRES's 8-day for temperature and 6-day for humidity. The gain from downscaling is largest in the first two days, where synoptic-scale uncertainty is smallest and local topographic detail accounts for a larger share of total error.

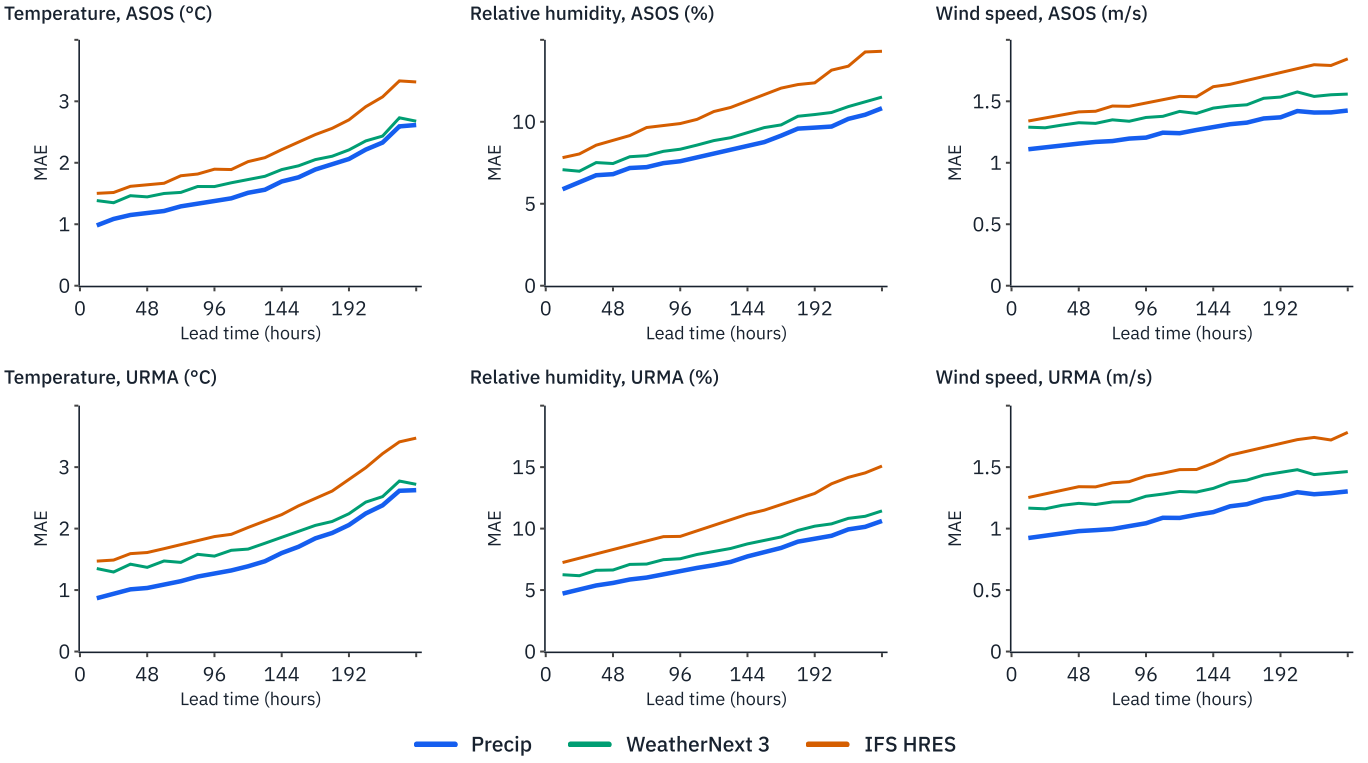

Figure 1. Mean absolute error by lead time, pooled in 12-hour bins, at ASOS stations (top) and against the URMA analysis (bottom). Lower is better.

Table 1. Temperature.

Error by forecast window against the URMA analysis and ASOS stations. Blue figures are Precip's margin over the next-best forecast (definitions).

Mean absolute error °C, lower is better	**URMA** analysis floor	**IFS HRES** ECMWF	**WeatherNext 3** Google	**Our Blend** Precip
Against the URMA 2.54 km analysis 2,417,664,576 grid-point hours				
Days 1–2	—	1.54	1.36	**0.97** +29%
Days 3–5	—	1.84	1.56	**1.24** +21%
Days 6–10	—	2.67	2.17	**1.96** +10%
Days 1–10	—	2.08	1.74	**1.44** +17%
Days 1–10, root mean square error	—	2.84	2.33	**2.01** +14%
Against ASOS stations 4,100,164 station hours				
Days 1–2	0.64	1.57	1.41	**1.10** +22%
Days 3–5	0.64	1.85	1.61	**1.36** +16%
Days 6–10	0.63	2.61	2.15	**2.00** +7%
Days 1–10	0.63	2.06	1.76	**1.53** +13%
Days 1–10, root mean square error	0.90	2.82	2.36	**2.11** +11%

Table 2. Relative humidity.

Error by forecast window against the URMA analysis and ASOS stations. Blue figures are Precip's margin over the next-best forecast (definitions).

Mean absolute error %, lower is better	URMA analysis floor	IFS HRES ECMWF	WeatherNext 3 Google	Our Blend Precip	
Against the URMA 2.54 km analysis 2,417,664,576 grid-point hours					
Days 1–2	—	7.8	6.4	**5.2**	+19%
Days 3–5	—	9.4	7.5	**6.4**	+15%
Days 6–10	—	12.5	9.7	**8.7**	+10%
Days 1–10	—	10.2	8.1	**7.0**	+14%
Days 1–10, root mean square error	—	14.0	10.8	**9.6**	+11%
Against ASOS stations 4,080,817 station hours					
Days 1–2	4.2	8.3	7.3	**6.4**	+12%
Days 3–5	4.2	9.9	8.3	**7.6**	+8%
Days 6–10	4.2	12.3	10.1	**9.3**	+8%
Days 1–10	4.2	10.4	8.7	**7.9**	+9%
Days 1–10, root mean square error	5.7	14.2	11.5	**10.5**	+9%

Table 3. Wind speed at 10 m.

Error by forecast window against the URMA analysis and ASOS stations. Blue figures are Precip's margin over the next-best forecast (definitions).

Mean absolute error m/s, lower is better	URMA analysis floor	IFS HRES ECMWF	WeatherNext 3 Google	Our Blend Precip	
Against the URMA 2.54 km analysis 2,417,664,576 grid-point hours					
Days 1–2	—	1.30	1.18	**0.95**	+19%
Days 3–5	—	1.41	1.25	**1.04**	+17%
Days 6–10	—	1.63	1.40	**1.21**	+14%
Days 1–10	—	1.46	1.29	**1.08**	+16%
Days 1–10, root mean square error	—	1.95	1.69	**1.46**	+14%
Against ASOS stations 4,077,841 station hours					
Days 1–2	0.69	1.38	1.30	**1.13**	+13%
Days 3–5	0.69	1.48	1.36	**1.21**	+11%
Days 6–10	0.69	1.69	1.49	**1.35**	+9%
Days 1–10	0.69	1.53	1.40	**1.24**	+11%
Days 1–10, root mean square error	0.94	1.98	1.78	**1.62**	+9%

Precipitation

Hourly rainfall is the most difficult surface variable to forecast at a point, and the one where raw ensemble averaging hurts most: an ensemble mean smears a storm's rainfall thinly across every location it might pass, suppressing peak rates. Conversely, a single deterministic run captures storm intensity but incurs double penalties whenever a storm is slightly displaced in space or time. Precip's blend is designed to predict both where rain will fall and how hard.

In the first two days against the URMA analysis, Precip's critical success index is 3.0× that of IFS HRES for hourly rainfall above 4 mm and 6.3× above 8 mm, while also achieving the lowest mean absolute error in every lead-time window across all three reference datasets.

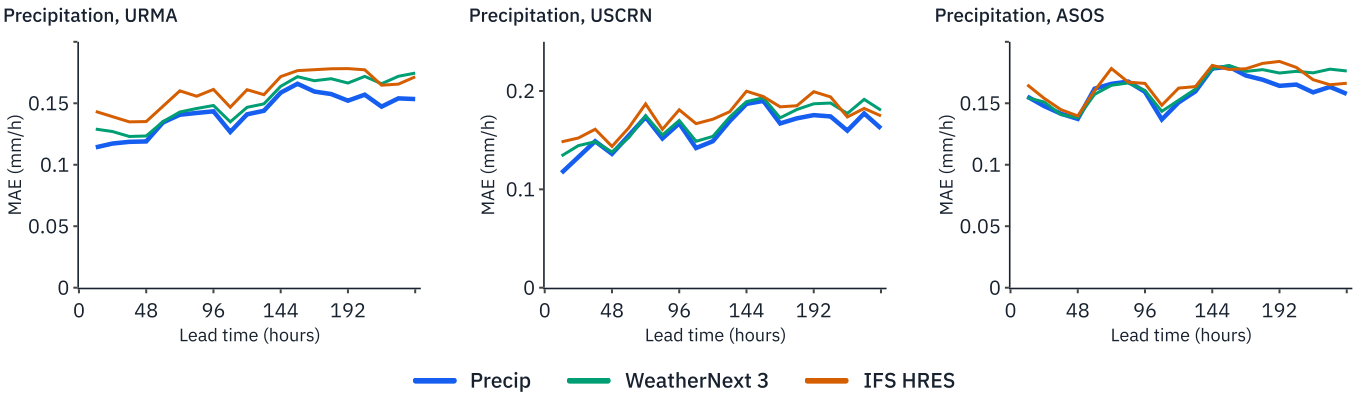

Figure 2. Hourly precipitation mean absolute error by lead time, pooled in 12-hour bins. Lower is better.

Table 4. Hourly precipitation amount.

Error by forecast window against the URMA analysis, USCRN gauges, and ASOS stations. Blue figures are Precip's margin over the next-best forecast (definitions).

Mean absolute error mm/h, lower is better	URMA analysis floor	IFS HRES ECMWF	WeatherNext 3 Google	Our Blend Precip	
Against the URMA 2.54 km analysis 9,066,237,606 grid-point hours					
Days 1–2	—	0.138	0.126	**0.117**	+7%
Days 3–5	—	0.156	0.142	**0.138**	+3%
Days 6–10	—	0.172	0.167	**0.155**	+7%
Days 1–10	—	0.160	0.151	**0.142**	+6%
Against USCRN gauges 757,314 gauge hours					
Days 1–2	0.051	0.151	0.141	**0.133**	+6%
Days 3–5	0.051	0.172	0.159	**0.156**	+2%
Days 6–10	0.053	0.187	0.183	**0.173**	+5%
Days 1–10	0.052	0.175	0.168	**0.160**	+5%
Against ASOS stations 15,794,126 station hours					
Days 1–2	0.072	0.151	0.146	**0.145**	+1%
Days 3–5	0.072	0.164	0.158	**0.157**	+1%
Days 6–10	0.072	0.175	0.175	**0.167**	+5%
Days 1–10	0.072	0.167	0.164	**0.160**	+2%

Table 5. Hourly rain events.

Skill at forecasting hours with rain above 1, 4, and 8 mm. Blue figures are Precip's margin over the next-best forecast (definitions).

Critical success index higher is better	URMA analysis floor	IFS HRES ECMWF	WeatherNext 3 Google	Our Blend Precip	
Against the URMA 2.54 km analysis					
Days 1–2, above 1 mm/h	—	0.167	0.242	**0.286**	+18%
Days 1–2, above 4 mm/h	—	0.052	0.036	**0.155**	3.0×
Days 1–2, above 8 mm/h	—	0.014	0.003	**0.088**	6.3×
Days 3–5, above 1 mm/h	—	0.103	0.152	**0.175**	+15%
Days 3–5, above 4 mm/h	—	0.024	0.009	**0.047**	+96%
Against USCRN gauges					
Days 1–2, above 1 mm/h	0.638	0.156	0.241	**0.264**	+10%
Days 1–2, above 4 mm/h	0.627	0.048	0.033	**0.142**	3.0×
Days 1–2, above 8 mm/h	0.608	0.008	0.000	**0.073**	9.1×
Days 3–5, above 1 mm/h	0.641	0.099	0.158	**0.171**	+8%
Days 3–5, above 4 mm/h	0.643	0.020	0.011	**0.058**	2.9×
Against ASOS stations					
Days 1–2, above 1 mm/h	0.497	0.131	0.191	**0.215**	+13%
Days 1–2, above 4 mm/h	0.476	0.047	0.044	**0.134**	2.9×
Days 1–2, above 8 mm/h	0.447	0.014	0.006	**0.080**	5.7×
Days 3–5, above 1 mm/h	0.499	0.076	0.132	**0.136**	+3%
Days 3–5, above 4 mm/h	0.478	0.020	0.009	**0.048**	2.4×

Chance of rain

The chance of rain is the probability of at least 0.1 mm of precipitation within a given hour. Because IFS HRES is a single deterministic run that can only issue binary yes-or-no predictions, probabilistic skill is compared here against the WeatherNext 3 ensemble, derived from its published precipitation quantiles. Precip's blend achieves a lower Brier score in every lead-time window across all three reference datasets, including a 34% reduction in days 1–2 Brier score at USCRN gauges.

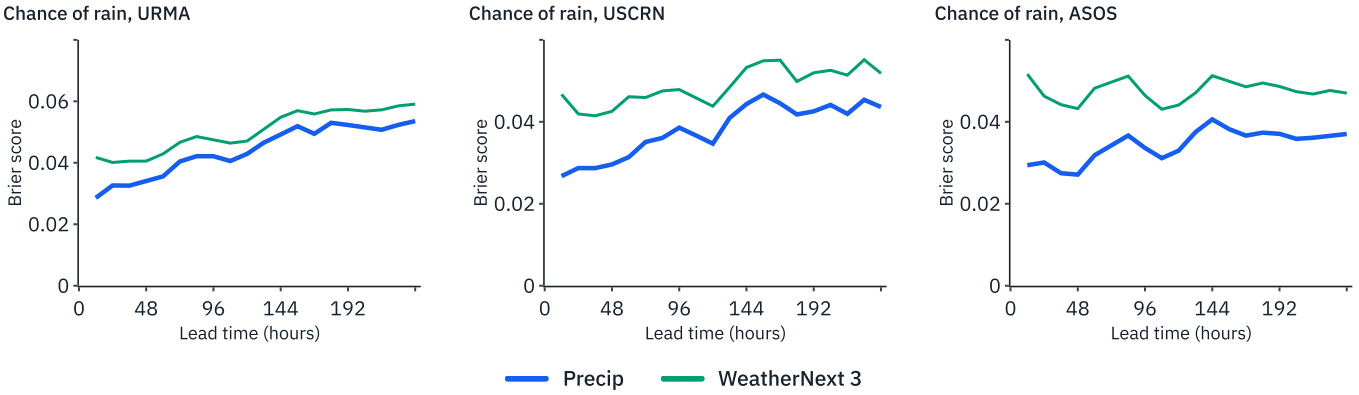

Figure 3. Brier score for the chance of rain (at least 0.1 mm in the hour) by lead time, pooled in 12-hour bins. Lower is better.

Table 6. Chance of rain.

Probability error by forecast window, and skill over climatology for the full 10 days. Blue figures are Precip's margin over the next-best forecast (definitions).

Brier score lower is better	**URMA** analysis floor	**WeatherNext 3** Google	**Our Blend** Precip	
Against the URMA 2.54 km analysis				
Days 1–2	—	0.0407	**0.0319**	+22%
Days 3–5	—	0.0465	**0.0406**	+13%
Days 6–10	—	0.0558	**0.0505**	+9%
Days 1–10	—	0.0486	**0.0422**	+13%
Days 1–10, Brier skill score (higher is better)	—	0.075	**0.197**	2.6×
Against USCRN gauges				
Days 1–2	0.0260	0.0432	**0.0284**	+34%
Days 3–5	0.0270	0.0462	**0.0354**	+23%
Days 6–10	0.0283	0.0522	**0.0434**	+17%
Days 1–10	0.0272	0.0477	**0.0367**	+23%
Days 1–10, Brier skill score (higher is better)	0.360	-0.121	**0.138**	
Against ASOS stations				
Days 1–2	0.0319	0.0463	**0.0285**	+38%
Days 3–5	0.0329	0.0471	**0.0334**	+29%
Days 6–10	0.0340	0.0485	**0.0376**	+22%
Days 1–10	0.0331	0.0474	**0.0337**	+29%
Days 1–10, Brier skill score (higher is better)	0.051	-0.361	**0.032**	

Methods

- **Scope.** The evaluation covers CONUS across 29 forecast cycles initialized at 00Z and 12Z between 2026-07-27 and 2026-10-08, spanning late-summer convective regimes and autumn synoptic systems.
- **Forecasts scored.** IFS HRES is evaluated on its native 9 km deterministic grid, as distributed by Open-Meteo, with no post-processing. Raw WeatherNext 3 is evaluated using its 0.1° ensemble mean for continuous variables and its published precipitation quantiles for the chance of rain. Precip's forecast is scored exactly as served in production for each variable and lead hour: WeatherNext 3 downscaled to 2.5 km for temperature, humidity, and wind speed, starting from its 0.05° station-calibrated fields for temperature and humidity; the precipitation blend for the chance of rain and for rainfall amounts in days 1–2; and probability-matched WeatherNext 3 for rainfall amounts in days 3–10, which keeps the ensemble mean's placement while restoring the peak intensities that averaging smooths away.
- **Held-out data.** All learned components were evaluated strictly out of sample. The 2.5 km downscaling networks were scored using a two-fold temporal split around 1 September 2026, ensuring that no forecast run was evaluated by a network trained on its time period. The precipitation blend was fitted on eight separate initialization cycles from July to October 2026, none of which are included among the 29 runs scored here.
- **Station quality control.** To filter out sensor spikes, ASOS temperature, dew point, and wind speed reports diverging by more than 10 units (°C or m/s) from the concurrent URMA analysis were excluded. Relative humidity observations were retained only when both temperature and dew point passed quality control.
- **Interpolation.** Each gridded forecast is bilinearly interpolated to the exact station coordinates, and all forecasts are verified across the identical set of stations and valid hours.

Data access

Underlying verification data and station-level scores are available upon request at team@precip.ai.

References

- European Centre for Medium-Range Weather Forecasts. IFS documentation.
- Google DeepMind. WeatherNext.
- NOAA National Centers for Environmental Prediction. Real-Time and Unrestricted Mesoscale Analysis (RTMA/URMA).
- NOAA National Centers for Environmental Information. U.S. Climate Reference Network.
- Iowa Environmental Mesonet. ASOS-AWOS-METAR data download.
- NOAA Global Systems Laboratory. High-Resolution Rapid Refresh (HRRR).
- Open-Meteo. Open-Meteo weather API.

Scores pooled across 29 forecast runs issued between 2026-07-27 and 2026-10-08.