Author response to Referee Comment RC1

Manuscript egusphere-2026-3655 — Physics vs. AI in Weather Prediction: Evaluating GraphCast, AIFS, and FuXi against an Observation-Corrected WRF Model for Flash Floods
Referee: Víctor Galván Fraile, 31 July 2026

What has changed since our first response

Three changes affect numbers throughout the manuscript.

GraphCast input. The checkpoint takes the six-hour precipitation accumulation as an input; we had supplied the one-hour value. All GraphCast forecasts were regenerated with the correct input. AIFS takes no precipitation input and FuXi ignores its precipitation channel, so neither is affected.

Verification network. Rebuilt from the raw provider archives — DWD, Météo-France, AgriMeteo and the Walloon SPW — giving 259, 280 and 264 stations for 2016, 2018 and 2021, each at least 12 km from any observation assimilated in that event.

2018 assimilation. Our PREPBUFR conversion had dropped the surface reports for the 2018 event. The observations were rebuilt — 43,338 SYNOP reports across the 125 analysis cycles — and the assimilation repeated. Only the 2018 assimilated run changed.

This letter replaces our earlier response in full. No revised manuscript was submitted at that stage, so this is the first revised version. Referee comments are shown in shaded italics and our reply follows each one.

Comments and replies

> ***Referee comment.*** *The 1 mm per 6 hours threshold is appropriate for evaluating the spatial coherence of precipitation but is less informative regarding rainfall intensity. Would the main conclusions still hold when considering higher precipitation thresholds that are more representative of extreme rainfall (e.g., 5, 10 or 20 mm per 6 hours)?*

They do not, and this is now a main result (Sect. 6.2.5, Fig. 10). The ranking reverses with intensity: at 1 mm GraphCast leads (CSI 0.424) ahead of AIFS (0.415) and the assimilated WRF (0.389); at 20 mm WRF leads (0.160) ahead of AIFS (0.141) and GraphCast (0.114), and FuXi produces no exceedance anywhere.

> ***Referee comment.*** *I recommend including confidence intervals to support statements about differences in model performance. In addition, please clarify if the verification metrics are computed just in those days where the events occur or in the entire forecasted cycle of 90 days. In the case of the latter, I would recommend analysing the metrics just during the event.*

Added (Sect. 6.2.3). We use a paired moving-block bootstrap over whole analysis cycles, 10,000 resamples at 6-, 12-, 24- and 48-hour block lengths, with intervals computed directly on differences. Figure 10 shows the 24-hour individual-model CSI intervals as error bars. The sample is 92,119 precipitation pairs and 45,925 temperature pairs per model.

At 24-hour blocks and 1 mm, GraphCast and AIFS exceed the assimilated WRF by +0.035 (+0.012 to +0.057) and +0.026 (+0.003 to +0.050), both excluding zero at every block length; GraphCast and AIFS are not separated (+0.009, –0.003 to +0.020). At 20 mm WRF exceeds GraphCast by +0.046 (+0.009 to +0.096) at every block length, and AIFS by +0.019 (+0.002 to +0.046) at 24 and 48 hours only. The assimilation gain at 1 mm (+0.0246) includes zero at 24-hour blocks and we report it as a sample behaviour.

For temperature, the AIFS and GraphCast reductions in RMSE against the assimilated WRF are –0.364 °C (–0.423 to –0.315) and –0.337 °C (–0.392 to –0.293), 19.0 % and 17.6 %. Assimilation improves the temperature bias by +0.167 °C (+0.115 to +0.222); its effect on RMSE, +0.036 °C (–0.008 to +0.083), is not resolved.

Scores are computed over the full windows. The event-only analysis is added in Sect. 6.2.6 and Fig. 11: scores are higher on the flood days in 2018 and 2021, and AIFS overtakes GraphCast in 2018 (0.466 against 0.452).

> ***Referee comment.*** *Have the authors assessed the sensitivity of the results to the data assimilation cycling frequency (e.g., 3-hourly or 1-hourly instead of 6-hourly)?*

We have not. The 6-hour step is fixed in all three AI models and cannot be changed without retraining, and cycling WRF more often would give it a shorter effective update interval than the products it is compared against. Sect. 5.1 states this and identifies the sensitivity as future work.

> ***Referee comment.*** *Could the authors clarify the independence of the verification dataset? It is not entirely clear whether all surface observations used for verification were also assimilated into WRF. This is particularly important when comparing WRF with AI models, since the latter ones do not ingest these observations. Indeed, comparing them with the WRF simulation without DA may provide a fairer benchmark.*

Precipitation is never assimilated here: the observation files contain pressure, wind, temperature, humidity and precipitable water only, and radar assimilation is disabled. For the station network we took the positions of every surface observation actually assimilated in each event and excluded any candidate within 12 km; the smallest remaining separation is 12.04 km (Sect. 5.3, Fig. 1).

We have softened “independent by construction”, since assimilated temperature, humidity and wind can influence rainfall indirectly and ERA5 has itself assimilated many of the same networks. The unassimilated WRF run is now included as a fifth system throughout, including the radar comparison, as the referee suggests.

> ***Referee comment.*** *The time series at the four selected stations reveal some differences between the radar estimates and the station measurements, especially at the Vatry station. Have the authors examined nearby stations to determine whether this discrepancy reflects a localised measurement issue or a limitation of the radar product? In addition, I suggest overlaying the observed station precipitation measurements on the model fields.*

Vatry is no longer in the paper. It came from the NOAA Integrated Surface Database, whose holdings for this region are sparse and several of whose stations, Vatry included, no longer carry a usable 6-hourly record. The network was therefore rebuilt from the national archives. Mirecourt replaces Vatry as the distant control, roughly 150 km south in France.

At the three Luxembourg gauges of Fig. 13 the gauge and radar totals agree closely over the episode (130.6, 121.8 and 131.8 mm against 129.3, 116.2 and 116.3 mm), so the radar is a reliable reference there. The station observations are now overlaid on the model fields in Fig. 12 on the same colour scale, as suggested.

> ***Referee comment.*** *Is the behaviour observed for the AI models during the July 2021 event (e.g., smoothing in AIFS and GraphCast, or the weak precipitation signal in FuXi) also found for the 2016 and 2018 events? Although radar data are unavailable for those cases, the comparison could still be performed using the station observations.*

Yes (Sect. 6.2.4, Fig. 8). GraphCast and AIFS lead in every event and FuXi is weakest in 2018 and 2021. The ordering is not identical throughout: in 2016 AIFS and GraphCast are tied and FuXi exceeds the unassimilated WRF, and the text says so.

> ***Referee comment.*** *It should be acknowledged that the AI models evaluated here are global, whereas WRF is a regional dynamical model. This fundamental difference may partly explain the smoother precipitation fields produced by the AI models. Related to this, while the WRF simulations had a horizontal resolution of 12 km, the AI models had around 25 km. Will the results still hold if the WRF simulation were run at 25 km?*

The global-versus-regional difference is now acknowledged in the Conclusions. We did not run WRF at 25 km: our configuration is a 12, 4 and 1.33 km nest chosen to respect the 1:3 nesting ratio, with the 1.33 km domain intended to drive hydraulic simulation in later work.

We tested resolution in the opposite direction, scoring all three nest members against the radar on identical geography and the same 22 six-hour periods (Sect. 6.5, Table 5). Refinement improves amplitude — the maximum 6-hour accumulation rises from 39.97 mm at 12 km to 53.40 and 58.12 mm, against 53.97 mm from radar — but not placement: FSS at 10 mm falls from 0.307 to 0.274. Sect. 6.5 states that this does not test WRF against the AI models at a matched 25–28 km grid.