Author response to Referee Comment RC2

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 2

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 WRF experiments use ERA5 as boundary and initial conditions followed by a regional data assimilation system incorporating additional observations, whereas the AI models are initialized directly from ERA5 analyses. The comparison between observation-corrected WRF and AI models may partly reflect differences in the quality of the initial atmospheric state rather than differences between dynamical and data-driven approaches.*

We agree. The unassimilated WRF run is now reported as a fifth system throughout, so the effect of assimilation can be read directly: at 1 mm it raises WRF from CSI 0.364 to 0.389, while the best AI system exceeds the unassimilated run by 0.060. Sect. 5.2 now frames the study as a comparison of ERA5-driven 6-hour hindcast systems with an explicit assimilation sensitivity, not as an isolation of physics from data-driven formulation under identical initial states.

> ***Referee comment.*** *The AI models are re-initialised from ERA5 every 6 hours rather than being run in a fully autoregressive mode. This design avoids error accumulation but does not represent the operational use case of multi-step AI forecasting. Please clarify why this evaluation strategy was selected and discuss how the results would change under free-running forecasts.*

The 6-hour re-initialisation matches the WRF Rapid Update Cycle, so both systems are scored at the same forecast lead time. A free-running AI forecast would be scored at longer leads than WRF and would confound the comparison. The reasoning and its limitation are stated in Sect. 5.1.

> ***Referee comment.*** *The WRF configuration uses a single 12 km domain. However, flash floods are often strongly influenced by convection-permitting processes and local topography. The manuscript should discuss whether the chosen WRF resolution is sufficient to represent convective rainfall mechanisms in this region. A comparison with higher-resolution convection-permitting WRF simulations would help determine whether the observed advantage of WRF is related to physical modelling itself or simply to spatial resolution.*

We ran the nested 12, 4 and 1.33 km configuration and scored all three against the radar on the footprint of the innermost domain, over the same 22 six-hour periods (Sect. 6.5, Table 5). Refinement does not improve

placement — FSS at 10 mm falls from 0.307 to 0.278 and 0.274 — but it does improve amplitude in the tail: the maximum 6-hour accumulation rises from 39.97 mm to 53.40 and 58.12 mm, against 53.97 mm from radar. The bulk does not follow: all three members are 20–24 % low in the domain mean.

> ***Referee comment.*** *AIFS uses fine-tuning with IFS analyses from 2016–2022, which overlaps with all three evaluation events. Therefore the comparison does not only represent architecture differences but also differences in training data and operational assimilation systems.*

Agreed. The manuscript now states that no event is out-of-sample for AIFS and that its scores should be read as potentially optimistic. The out-of-sample argument is restricted to GraphCast, whose training data end in 2017 and whose skill is highest on the unseen 2021 event (CSI 0.493 against 0.430 for 2016).

> ***Referee comment.*** *Table 2 provides basic information about AI models, but important details for reproducibility are missing: exact model checkpoints/releases used; inference settings; precipitation variable definition; any preprocessing applied to ERA5 inputs; interpolation method from model grid to station locations.*

Table 3 has been added, giving the checkpoint identity for each model, the variable names, units and conversions, the ERA5 pre-processing and the nearest-grid-point extraction. Checkpoint file names, checksums and package versions are in the archived repository, now Zenodo v1.3 (concept DOI 10.5281/zenodo.20794936; this version 10.5281/zenodo.22664141), which also holds the four ECMWF HPC scripts that generated the AI forecasts.

> ***Referee comment.*** *Figure 9 provides qualitative comparison of rainfall fields. However, visual comparison alone is subjective. Please consider adding spatial verification metrics, such as Fractions Skill Score (FSS), object-based precipitation verification, spatial correlation. This would strengthen the conclusion that WRF better represents rainfall structures.*

The Fractions Skill Score and the spatial correlation against the radar composite have been added (Sect. 6.4, Fig. 14) at four neighbourhood widths and four thresholds, over the 27 six-hour periods for which the radar total is complete and every system has a field. They show the same reversal as the station scores: at 1 mm AIFS leads at every width (0.574 against 0.530 for GraphCast and 0.519 for the assimilated WRF at 12 km), whereas at 20 mm a WRF configuration leads at every width (assimilated 0.178 at 12 km against 0.126 and 0.122; cold-start 0.324 at 100 km against 0.197 and 0.198).

Two points of method are stated explicitly in Sect. 6.4. A 6-hour radar total is formed only where all six hourly composites exist, which removes one of the 28 periods. Scores are reported both as a panel mean and as a pooled aggregate, because the two can rank systems differently: at 1 mm and 12 km the aggregate leader is GraphCast (0.748) whereas AIFS leads the panel mean (0.574). Both favour WRF at 20 mm.