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

We thank the referee for a careful and constructive review. The comments prompted us to re-examine our verification procedure in detail and, in doing so, we identified and corrected an error in the pre-processing of the observations (described immediately below). All conclusions of the original manuscript are unchanged, and one new result has been added that we believe substantially strengthens the paper.

Page and line numbers refer to the revised manuscript. Referee comments are shown in shaded italics; our response follows, with the resulting manuscript changes marked in green.

Correction identified during revision

While addressing General Comment 4 we re-examined the pre-processing of the NOAA Integrated Surface Database (ISD) records used for verification at non-Luxembourgish stations. ISD reports precipitation in overlapping accumulation periods (1, 3, 6, 12 or a running 24 h) within the same AA field. Our original routine parsed the accumulation period but did not use it, and summed all reports falling within each 6-hour window. At stations reporting a running 24-hour total every hour this substantially over-estimated the observed rainfall.

We have rewritten the extraction to respect the reported accumulation period and have repeated the entire verification. The Luxembourg AgriMeteo observations, which constitute the majority of the verification network and are processed by a separate routine, are unaffected: their totals are identical before and after the correction.

The effect is confined to the mean precipitation bias. Because the observations were previously over-estimated, all four systems appeared to carry a dry bias of roughly 0.5–1.1 mm per 6 h; with the corrected observations the biases are small (−0.13 to −0.38 mm, with AIFS marginally wet at +0.11 mm). The statement that all models systematically underestimate total precipitation has therefore been removed. **No model ranking and no other conclusion changed.** As set out under General Comment 1, the underestimation of extreme rainfall is real, but it resides in the upper tail of the distribution rather than in the mean, and is now demonstrated directly by the threshold analysis.

Summary of principal additions

Referee point	Addition	Location
GC1	Threshold sensitivity (1/5/10/20 mm) + Table 4	Sect. 6.2.5, p15 l255
GC2	Bootstrap 95% confidence intervals + Table 3	Sect. 6.2.3, p13 l226
GC2	Flood-day-only verification + Table 5	Sect. 6.3, p18 l261
GC3	Fixed 6-hour AI time step as design constraint	Sect. 5.1, p9 l148
GC4	Verification network and its independence	Sect. 5.3, p9 l157
GC5	Gauge markers on radar comparison	Fig. 9, p19
GC7	Global vs. regional discussion; new citations	Conclusions, p22 l346
L44	Physics provenance; Stergiou, Knist references	Sect. 3.1, p4 l81
L212	Quantitative runtime comparison	Sect. 5.1, p9 l153

General comments

GC1 — Higher precipitation thresholds

> ***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)?*

We agree, and we thank the referee for this suggestion: it has produced what we now regard as the most important result in the paper. The categorical scores have been recomputed at 5, 10 and 20 mm $(6\ \mathrm{h})^{-1}$.

The conclusions do **not** hold uniformly across thresholds — the ranking reverses. AIFS remains the most skilful system up to 10 mm (CSI 0.324 against 0.193 for WRF), but at 20 mm $(6\ \mathrm{h})^{-1}$ the assimilated WRF becomes the best system (CSI 0.205 against 0.172 for AIFS), GraphCast retains almost no skill (0.032), and FuXi does not produce a single exceedance of 20 mm anywhere in the verification set. Restricting the comparison to the flood windows sharpens the contrast further (WRF 0.315 versus GraphCast 0.037).

This is precisely the behaviour the referee anticipated from the spatial comparison, and it is now quantified. It also explains the smoothing visible in Fig. 9: models trained with pixel-wise losses are rewarded for predicting the conditional mean, which suppresses the upper tail of the distribution.

> **Changes in the manuscript.** New Sect. 6.2.5 ‘Sensitivity to the precipitation threshold’ and Table 4 (p15, l255–285). Abstract (p1, l10–17) and Conclusions (p21, l338–346) revised to state that the apparent advantage of the AI systems at low thresholds should not be extrapolated to flash-flood-relevant rainfall rates.

GC2 — Confidence intervals; event-only versus full-window metrics

> ***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.*

Confidence intervals. Added: 95 % intervals obtained from 1000 bootstrap resamples of the matched forecast–observation pairs, for all pooled precipitation and temperature scores. These clarify which differences are resolved by the sample and which are not. The AIFS detection advantage is unambiguous (POD interval 0.762–0.811, disjoint from all others), as is FuXi’s inferiority and the temperature advantage of GraphCast and AIFS. However, the CSI intervals of WRF (0.352–0.399), GraphCast (0.358–0.407) and AIFS (0.382–0.423) overlap substantially, and we now state explicitly that the ranking of those three systems is not resolved at the 1 mm threshold. This is an additional motivation for the threshold analysis of GC1, where the separation greatly exceeds the sampling uncertainty.

Verification period. The scores in the original manuscript were computed over the full 30-day windows. We have added the event-only analysis the referee suggests (689 matched pairs spanning the flood days alone) and we now justify the 30-day choice explicitly. The first objective of the study is to establish whether assimilation improves forecast skill, and a two- to four-day sample does not contain enough independent precipitation events to separate that signal from sampling noise; the longer window also captures the pre-event synoptic evolution and provides the spin-up required by the Rapid Update Cycle.

The two are consistent: the relative ordering of the systems is unchanged, confirming that the full-window conclusions are not an artefact of including quiescent days. Two features are specific to the flood days. All systems show a pronounced dry bias during the events (−1.5 to −3.7 mm) in contrast to the near-zero full-window biases, confirming that the underestimation is confined to heavy-rainfall periods. And during the July 2021 flood the assimilated WRF achieves a false-alarm ratio of 0.095, the lowest value recorded anywhere in this study.

> **Changes in the manuscript.** New Sect. 6.2.3 and Table 3 (p13, l226–250); new Sect. 6.3 and Table 5 (p18, l261–300); 30-day justification added to Sect. 5.1 (p8, l137–144).

GC3 — Sensitivity to the assimilation cycling frequency

> ***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, and we would like to explain why this is a design constraint rather than an omission. The 6-hour interval is imposed by the AI models: GraphCast, FuXi and AIFS all operate on a fixed 6-hour native time step that is built into their architecture and cannot be altered without retraining. Cycling WRF at 1 or 3 hours would therefore have given the physics-based system an update frequency that the data-driven systems are structurally unable to match, and would have removed the like-for-like basis of the comparison. Six hours is also the interval at which ERA5 initial conditions are available to all systems.

Higher-frequency cycling is certainly of interest for WRF considered on its own, and we note it as future work, but it falls outside the controlled comparison this paper is designed to make.

> **Changes in the manuscript.** Reasoning added to Sect. 5.1 (p9, l148–152).

GC4 — Independence of the verification dataset

> ***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.*

This was the most consequential comment and we have addressed it in three ways.

1. Precipitation was never assimilated. The WRFDA observation files used in this study contain only pressure, wind speed and direction, height, temperature, dew point and humidity, together with sea-level pressure and precipitable water; radar reflectivity and rain-rate assimilation are disabled in our configuration. There is no precipitation field in the observation format, so the precipitation verification — the primary variable of the paper — is independent of the analysis by construction.

2. The verification network has been screened. We extracted the positions of every SYNOP report from all 244 `ob.ascii` files actually used by the Rapid Update Cycle, rather than assuming which stations were involved, and excluded every verification station within 3 km of an assimilated report. We additionally excluded stations whose ISD records cannot support a 6-hour accumulation, retaining only those supplying at least 95 % of the 6-hourly windows. The verification network is now 16, 17 and 20 stations for 2016, 2018 and 2021 respectively. It is dominated by the Luxembourg AgriMeteo gauges, which are agro-meteorological stations that do not report on the Global Telecommunication System and therefore cannot enter an operational analysis: they are independent by construction rather than by selection.

3. The unassimilated WRF run is now included in the spatial comparison, following the referee's suggestion (see the response to the specific comment on Fig. 9).

We note that the retained network is concentrated in Luxembourg. This follows from the study design: the domain covers the Greater Region because the responsible synoptic systems develop at that scale, but the hydrological interest is in the small, fast-responding Luxembourgish catchments, and Luxembourg is where the densest independent observations exist. The 2021 event additionally retains one German and one French station, which allow the domain-scale behaviour to be checked.

> **Changes in the manuscript.** New Sect. 5.3 'Verification network and its independence' (p9, l157–185). Fig. 9 extended with a WRF (Before DA) row (p19). Abstract updated to state that verification uses stations independent of the assimilated network (p1, l6).

GC5 — Vatry discrepancy and station markers on Fig. 9

> ***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 as coloured markers on the maps in Fig. 9.*

Vatry. Our investigation gave a definite answer, and it is not the one the original manuscript implied. The NOAA ISD archive contains **no precipitation field at all** for Vatry during the study period; the apparent zeros were missing data recorded as zero. The speculative explanation offered in the original manuscript — a localised miss or an unreliable gauge — was therefore incorrect, and we apologise for it. Vatry is also co-located with an assimilated SYNOP report and has been removed from the verification network under the criteria described in GC4, so the discrepancy no longer arises.

Station markers. Implemented. Fig. 9 now overlays the gauge observations valid at each panel time as filled circles on the same colour scale as the fields, so that station, radar and model values can be compared directly. This proves informative in its own right: within Luxembourg the gauges and the radar composite agree closely on both the location and the magnitude of the heaviest accumulations, which supports treating the radar field as a reliable spatial reference for judging the models.

Changes in the manuscript. Fig. 9 regenerated with gauge markers and a WRF (Before DA) row (p19). Fig. 10 now shows four stations of the independent network — Briedfeld, Ettelbruck, Oberkorn and Mirecourt — spanning the Oesling–Gutland gradient (p20). Vatry discussion removed; new discussion of gauge–radar agreement added (Sect. 6.4, p19).

GC6 — Is the behaviour also found for 2016 and 2018?

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. No open weather-radar composite equivalent to RADFLOOD21 exists for 2016 and 2018, so a spatial comparison of the kind shown in Fig. 9 is not possible for those events. The station-based comparison the referee suggests is, however, already reported in the per-event breakdown and it confirms that the behaviour is not specific to 2021: the model ranking established by the pooled scores holds in every individual event. FuXi in particular is the weakest system in all three events, with a false-alarm ratio of 0.754 in 2018 and 0.644 in 2021, and its ETS falls to 0.076 in 2018. The event-only analysis added under GC2 shows the same for the flood days considered separately.

While revising this section we also tightened a related claim. The original manuscript stated that the AI models show no skill degradation on out-of-sample events. That test is in fact only available for GraphCast and FuXi, whose training data end in 2017: for them the 2021 event is genuinely unseen, and GraphCast's skill is actually higher there (CSI 0.441) than for the in-training 2016 event (0.314). AIFS Single 1.0, however, was fine-tuned on IFS analyses covering 2016–2022, so all three case studies fall inside its fine-tuning window and none is out-of-sample for that model.

Changes in the manuscript. Generalisation claim restricted to GraphCast and FuXi in the Abstract (p1, l18–20) and Sect. 6.2.4 (p14, l246–251), where the AIFS results are now described as an upper bound on performance for genuinely unseen events.

GC7 — Global versus regional models, resolution, limited-area AI

> ***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? It would also strengthen the discussion to compare these results with recent studies on regional or limited-area AI weather models (e.g., Xu et al., 2025; Baño-Medina et al., 2025).*

We agree that this asymmetry should be stated explicitly and have added a paragraph to the conclusions acknowledging that the three AI systems are global models at 0.25° or N320 (≈25–28 km) whereas WRF is a limited-area model at 12 km, and that part of the smoothing in the AI fields is therefore attributable to resolution rather than to the data-driven formulation itself.

We have not repeated the WRF simulations at 25 km. The 12 km configuration is the outermost member of a 12–4–1.33 km nest built following the recommended 1:3 nesting ratio and designed to drive hydraulic simulation of the Luxembourgish catchments; the 1.33 km member is already being used for hydraulic modelling of the July 2021 flood, to be reported separately. Degrading the outer domain to 25 km would not correspond to any configuration of operational or hydrological interest to us, and would not isolate the resolution effect cleanly in any case, since it would simultaneously alter the physics behaviour of the model.

We have added the comparison with limited-area data-driven models that the referee suggests, citing both recommended references, and we identify regional AI models trained specifically for extremes as the natural route to separating the resolution effect from the data-driven formulation. We are grateful for these two references, which we had not included.

> **Changes in the manuscript.** Paragraph added to the Conclusions (p22, l340–352), citing Xu et al. (2025) and Baño-Medina et al. (2025).

Specific comments

Line 32 — reframe the paragraph to motivate the analysis

> ***Referee comment.*** *I suggest reframing this paragraph to motivate your analysis, using the differing results to clarify the strengths and weaknesses of each modelling framework.*

Done. The paragraph now presents the two studies as pointing in different directions, notes that they may simply describe different parts of the precipitation distribution, and uses that tension to motivate a comparison across a range of rainfall intensities rather than at a single threshold. This connects directly to the new threshold analysis of GC1.

> **Changes in the manuscript.** Introduction rewritten (p2, l30–42).

Line 44 — provenance of the WRF configuration and physics sensitivity

> ***Referee comment.*** *Is the choice of the WRF configuration, described as being tuned for this region, based on a previous study? If so, please provide the corresponding reference. In addition, how sensitive are the results to the selected physical parametrisations, particularly the cumulus parametrisation?*

We have clarified the provenance of the configuration. The domain was defined to cover the Greater Region because that is the scale at which the responsible synoptic systems develop, while the hydrological focus, and hence the density of verification stations, is Luxembourg. The two choices that most directly govern convective rainfall at this grid spacing are corroborated by an independent systematic study: Stergiou et al. (2022) evaluated an ensemble of 30 parameterisation combinations against eight extreme precipitation events over Germany and identified the Grell–Freitas cumulus and YSU boundary-layer schemes as the best-performing members of their respective groups; both are adopted here. Knist et al. (2018) likewise employed the Noah land surface model and YSU boundary layer in convection-permitting simulations over Central Europe. The Thompson

microphysics scheme was retained for its double-moment treatment of ice and rain, well suited to the mixed-phase and orographically enhanced precipitation characteristic of the Ardennes–Eifel uplands. A comprehensive sensitivity analysis of the full physics suite lies beyond the scope of the present comparison, whose object is the contrast between physics-based and data-driven forecasting rather than the tuning of any single model, and we note it as future work. The CV5 background error covariance, by contrast, was not taken from a default dataset but generated specifically for this domain using the NMC method over a one-month period.

Changes in the manuscript. Paragraph added to Sect. 3.1 (p4, l78–86), citing Stergiou et al. (2022) and Knist et al. (2018).

Fig. 1 — add levels to the topography

Referee comment. *I suggest adding levels to the topography to better illustrate to the reader the influence that the terrain can play in the study area.*

Done. The topography is now shaded in 100 m elevation bands with a horizontal colour bar, which makes the contrast between the northern Oesling and the southern Gutland — and hence the terrain control on the hydrological response — much clearer.

Changes in the manuscript. Fig. 1 regenerated (p3); caption updated.

Line 113 — why a 30-day window?

Referee comment. *Why do the WRF simulations use a 30-day window and not a shorter one? If there is a reason for this election, please justify it in the text.*

Justified in the text; see the response to GC2 above.

Changes in the manuscript. Paragraph added to Sect. 5.1 (p8, l137–144).

Fig. 6 — why do the WRF scores differ from Fig. 4?

Referee comment. *Why are the verification scores for WRF (with DA) different from those shown in Fig. 4?*

Because the two figures use different samples. Fig. 4 uses every forecast–observation pair available to WRF, whereas Fig. 6 uses only those pairs available simultaneously for all four systems, so that the comparison is like-for-like. The difference is small, but we agree it was unexplained.

Changes in the manuscript. Caption of Fig. 6 now states this explicitly (p12).

Fig. 9 — include the WRF simulation without data assimilation

Referee comment. *Could the authors include the WRF simulation without data assimilation in this figure? This would help to assess the impact of data assimilation on the spatial distribution and intensity of the predicted precipitation for this event.*

Done. Fig. 9 now has six rows, with WRF (Before DA) and WRF (After DA) shown separately. The comparison is informative: both configurations place the rain band in approximately the correct position, but the assimilated run intensifies it and sharpens its gradients, bringing the peak accumulations closer to the radar and to the co-located gauges. The improvement in the categorical scores therefore corresponds to a visible structural improvement in the precipitation field rather than to a uniform rescaling.

Changes in the manuscript. Fig. 9 regenerated with six rows (p19); discussion added to Sect. 6.4 (p19).

Line 212 — quantitative comparison of computational requirements

Referee comment. *The discussion of the computational efficiency is very valuable. It would be helpful to include a quantitative comparison of the computational requirements (e.g., execution time) for the WRF and AI forecasts.*

Added. On 52 CPU cores, one 6-hour WRF cycle takes approximately 34 min without assimilation and 37 min with the 3D-VAR analysis included, so a 30-day window requires roughly 70 h and 75 h of wall-clock time respectively. On a single GPU with 40 GB of VRAM, one 6-hour AI forecast takes about 1.5 min and a full 30-day window about 30 min — more than two orders of magnitude faster, albeit after a training phase whose cost is borne once by the model developers.

Changes in the manuscript. Paragraph added to Sect. 5.1 (p9, l153–156); Conclusions updated (p22, l356).

We thank the referee once more for a review that has materially improved the manuscript, and in particular for the suggestion to examine higher precipitation thresholds, which revealed a result we consider central to the paper.