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

We thank the second referee for a detailed and technically demanding review. Four of the six comments prompted new analyses, two of which produced results we consider important additions to the paper: a quantitative spatial verification against radar, and a resolution experiment that separates the effect of grid spacing from that of the modelling approach. One of these analyses also revealed that the benefit of data assimilation is threshold-dependent, which has led us to qualify the first conclusion of the manuscript. That change is described under comment 1.

Page and line numbers refer to the revised manuscript. Where a point overlaps with the first referee's review, we indicate the corresponding item in our response to RC1 so that the two replies can be read together.

Note on a correction reported in our response to RC1

For completeness we repeat here a correction described in our response to the first referee. While addressing a question on verification independence we identified an error in the pre-processing of NOAA Integrated Surface Database records, in which overlapping accumulation periods were summed rather than being selected according to the reported period. This over-estimated observed rainfall at some non-Luxembourgish stations. The extraction has been rewritten and the entire verification repeated. The Luxembourg AgriMeteo observations, which form the majority of the verification network, are unaffected. The correction removed an apparent dry bias of 0.5–1.1 mm per 6 h in all four systems; no model ranking changed.

A second correction, found after the RC1 response was submitted

While looking more closely at the assimilation results for 2018, we found a fault in our own pre-processing. The step that converts the NCEP PREPBUFR archive into the little_r format had silently dropped every surface report for that event. The 2018 analyses had therefore been computed from an observation set made up almost entirely of satellite wind vectors. The 2016 and 2021 events were not affected.

We rebuilt the 2018 observations from the original PREPBUFR files, which recovered 42,997 SYNOP reports, and repeated the full 124-cycle assimilation experiment. Only the 2018 assimilated WRF run changed. The observations, the unassimilated run and all three AI forecasts are exactly as before, so the comparison between the four systems is unaffected.

Because this fault came to light only during review, we then audited all three events the same way: the platform counts in every cycle header were tabulated and checked against the reports actually present, with 20 cycles per event inspected in detail. The 2016 and 2021 sets carry their full complement of surface reports throughout and no further discrepancy was found. This audit is described in Sect. 3.4.

The effect on the numbers is small. Pooled CSI for the assimilated WRF moves from 0.372 to 0.374 at 1 mm and from 0.233 to 0.243 at 5 mm, is unchanged at 0.193 at 10 mm, and moves from 0.205 to 0.204 at 20 mm. One consequence is that the figure 0.205 quoted in our response to the first referee now reads 0.204 in the manuscript. We mention this so that the small difference between the two responses is not mistaken for an inconsistency. No ranking and no conclusion changes.

Importantly, the rebuild did not remove the loss of skill at high thresholds described under comment 1. That result survived the correction, and understanding why led us to the explanation given there.

Summary of new analyses

Comment	New analysis	Location
1	WRF without DA added as a fifth system; observation counts per event	Table 6 and Table 7, p19-20
2	Justification of the single-step evaluation design	Sect. 5.2, p10
3	Resolution experiment, WRF 12 / 4 / 1.33 km nest	Sect. 6.5 and Table 10, p25
4	AIFS fine-tuning overlap, scope of claims restricted	Sect. 6.2.4, p14
5	Model versions, inference settings, post-processing	Sect. 4.1 and Table 3, p9
6	Fractions Skill Score and spatial correlation vs radar	Sect. 6.4 and Table 9, p24

Comment 1 — Imbalance in the initial conditions

> ***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 AI forecasting approaches. I suggest including a more explicit comparison between WRF without DA and AI models, or quantifying the contribution of improved initial conditions relative to the forecasting model itself.*

This is a fair objection and we have addressed it directly. The design of the study was to compare the best available configuration of each paradigm, and a regional NWP system reaches that state only when it assimilates observations, which is why the assimilated run was used as the reference. The referee is nonetheless right that this conflates two effects, and they can be separated because the unassimilated run exists for all three events. It has now been added as a fifth system throughout the threshold analysis.

The decomposition at the 1 mm threshold is unambiguous. Assimilation raises the WRF Critical Success Index from 0.336 to 0.374, a gain of 0.038. The best AI system exceeds the *unassimilated* WRF by 0.063, which is larger than the entire contribution of the assimilation. The advantage of the AI systems at low thresholds therefore cannot be explained by the difference in initial conditions.

CSI	1 mm	5 mm	10 mm	20 mm
WRF (Before DA)	0.336	0.267	0.306	0.234
WRF (After DA)	0.374	0.243	0.193	0.204
GraphCast	0.379	0.210	0.176	0.032
FuXi	0.245	0.083	0.005	0.000
AIFS	0.399	0.260	0.324	0.172

The analysis also produced a result we had not anticipated, which we report for completeness. The benefit of assimilation is confined to light rainfall. At 1 mm the improvement is significant (+0.038, 95 % bootstrap interval +0.018 to +0.060), but at 10 mm the assimilated run scores *lower* than the cold-start run (0.193 against 0.306; difference −0.112, interval −0.174 to −0.051), and at 20 mm the unassimilated run is the most skilful of all five systems. The degradation at 10 mm is concentrated in the localised convective event of 2018 (−0.282), is small for the large-scale 2021 episode (−0.022), and is absent in 2016 where neither configuration produces any exceedance.

We looked into this carefully, because a loss of skill this large needed an explanation before we were willing to publish it. What we found is that the three events did not have comparable observations available to assimilate. The table below gives the counts averaged over 20 cycles of each event.

Event	Total	SYNOP	METAR	GPSZD	SATOB	TAMDAR
2016	924	293	210	32	351	0
2018	1209	349	207	27	594	0
2021	3285	1335	0	223	0	1710

The 2021 analysis has about three times more observations, including 223 GNSS ZTD retrievals and roughly 1700 aircraft profiles, so the moisture and temperature structure is constrained well above the surface. The 2016 and 2018 analyses have fewer than 35 ZTD retrievals, no aircraft data, and are dominated by satellite wind vectors. Over Luxembourg itself the difference is 45 GNSS sites in 2021 against two in 2016 and one in 2018. This is not a design choice on our part. Aircraft coverage depends on commercial traffic, and the denser 2021 GNSS network came from regional providers whose archives do not reach back to the earlier events. It is a constraint an operational system would face in the same way.

The contingency tables show this clearly. For 2018 at 10 mm the unassimilated run turns 22 of its 34 heavy forecasts into hits, while the assimilated run turns only 3 of 28 into hits, which is close to the base rate of the sample. The assimilated run does not produce less heavy rain. It produces a similar amount in the wrong places.

Two further points guard against the obvious objections. The difference is resolved by the data for 2018 only: resampling the stations gives -0.282 with a 95 % interval of -0.400 to -0.178, against -0.022 (-0.091 to +0.045) for 2021 and no difference for 2016, so the pooled effect is carried by a single event. And it is not a matter of the analysis simply adding or removing rain: the frequency bias at 10 mm is 0.60 before assimilation and 0.49 after, so both configurations under-forecast the number of heavy events. We should also be clear that with roughly 1200 reports per cycle these analyses are not near the no-observation limit, where 3D-VAR would return the background and the two runs would coincide; what differs is the composition of the observation set, not its size alone. A complete attribution would need the observation-minus-background and observation-minus-analysis statistics per cycle, which we did not retain and which we now recommend archiving in any repetition of this work.

Two checks support this. First, AIFS, which shares nothing with our WRF chain, turns 22 of 36 heavy forecasts into hits for the same event, so the 2018 rainfall really was predictable at this threshold and the result is not caused by our verification network. Second, rebuilding the 2018 observations and restoring the full surface network (see the note on page 1) did not repair the high-threshold scores, which confirms that surface observations alone cannot supply the moisture information aloft that these convective cases need.

We are careful about how strongly we state this. The three events differ at the same time in observation density, observation type and synoptic character: the densely observed 2021 episode is also the only large-scale forced case, while 2016 and 2018 are weakly forced convective events in which rainfall placement is intrinsically harder to predict. With three events these factors cannot be separated, so we present the link as an association and not as a demonstrated cause. Establishing the cause would require repeating one event with its observation set varied systematically while the weather situation is held fixed, and we now state this in the manuscript as the priority for future work.

We would also stress what this heterogeneity does not affect. It concerns the input to one configuration, WRF (After DA). In every event all systems are initialised from the same ERA5 fields, re-initialised on the same 6-hourly cycle, and verified against the same station network with the same metrics. The physics-versus-AI comparison that the paper is built on is therefore uniform across all three events, and we have added a paragraph to Sect. 5.2 saying so explicitly. We considered restricting the paper to the 2021 event, where the observing network is densest, but rejected this: it would remove the only events lying inside the GraphCast and FuXi training periods, and with them the in-sample versus out-of-sample test that the referees asked us to address, and it would reduce the verification sample by roughly 60 percent.

What we do state is conditional and, we believe, useful: assimilation improved where rain falls in all three events, but improved how hard it falls only in the event whose observing network constrained moisture aloft. For flash-flood forecasting this is worth reporting, because it indicates that the benefit of a regional assimilation system at flood-relevant rainfall rates cannot be assumed from its benefit at low thresholds.

Changes in the manuscript. WRF (Before DA) added to Table 6 at all four thresholds (p19). New discussion of the threshold-dependence of the assimilation benefit, stated as an association with the confounding factors named, plus the new Table 7 of assimilated observation counts per event (p19–20). New Sect. 3.4 describing the reconstruction of the 2018 observations (p9). Abstract (p1, l11–14) and first conclusion (p26) restated in the conditional form above. Sample sizes at the highest thresholds are stated so that the per-event attribution is read as indicative.

Comment 2 — Re-initialisation every 6 hours rather than free running

> ***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 evaluation strategy follows from the object of the comparison. WRF is run here in a 6-hourly Rapid Update Cycle, in which each forecast is re-initialised from a fresh analysis. Running the AI models autoregressively while cycling WRF every 6 hours would have compared a free forecast against a repeatedly re-initialised one, so that any difference would reflect the cycling strategy rather than the models. Both systems are therefore re-initialised at the same interval, and the comparison isolates the single-step forecast operator. Six hours is additionally the native time step of all three AI systems, which is fixed by their architectures and cannot be altered without retraining, and the interval at which ERA5 initial conditions are available. This is the same reasoning we gave to the first referee in response to a question on the assimilation cycling frequency.

On how the results would change, we prefer to state the expectation rather than assert an untested result. Free-running AI forecasts accumulate error with lead time and are known to smooth further as the rollout proceeds, so we would expect the deficit at high thresholds documented here to grow rather than diminish. Testing this properly requires a matching free-running WRF experiment, which is a different experimental design addressing lead-time degradation rather than model comparison, and we have not attempted it here.

We accept the referee's implicit criticism that a claim of computational efficiency for long-range forecasting is not supported by a single-step evaluation. The efficiency figures quoted in the manuscript have been scoped explicitly to single-step inference.

> **Changes in the manuscript.** Justification added to Sect. 5.2 (p10, l150–156). Multi-step rollout identified as future work in the Conclusions. Efficiency claim restricted to single-step inference (p26).

Comment 3 — Sufficiency of the 12 km domain

> ***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 for rainfall structure is related to physical modelling itself or simply to the spatial resolution difference.*

We have carried out the experiment the referee proposes. The 12 km domain used throughout the study is the outermost member of a 12–4–1.33 km nest configured on the recommended 1:3 ratio, whose innermost member is intended to drive hydraulic simulation of the Luxembourgish catchments at approximately 1 km. That nest was integrated for the flood peak of 14–16 July 2021, and all three members have now been scored against the radar composite over the footprint of the innermost domain, so that the three resolutions are compared on identical geography.

Refining the grid does not improve the position-based scores. The Fractions Skill Score at 10 mm falls from 0.792 at 12 km to 0.661 at 4 km and 0.639 at 1.33 km, and the spatial correlation declines in the same order. This is the double-penalty behaviour familiar from convection-permitting verification: a finer model produces sharper and more numerous small-scale features, and any displacement of them is penalised twice. The amplitude statistics move in the opposite direction, as convection-permitting simulation should: the maximum 6-hour accumulation rises from 39.97 mm at 12 km to 45.10 mm at 4 km and 47.66 mm at 1.33 km, approaching the 53.97 mm recorded by radar.

This bears directly on the referee's question. The advantage of WRF over the AI systems cannot be attributed to horizontal resolution alone, because increasing the resolution of WRF by an order of magnitude does not increase its FSS, whereas the gap between the 12 km WRF and the 25 km AI systems at a common threshold is

large. We are careful not to overstate this: resolution does contribute to the sharpness of the WRF fields, as we acknowledged in response to the first referee, and the experiment covers one event and one three-panel window. But it does establish that resolution is not the origin of the contrast between the two paradigms.

On why the assimilation was not repeated on the finer grids: the nest is run with two-way feedback, so the increment introduced on the 12 km domain is propagated to all three members rather than staying on the parent. Assimilating the same reports again on the inner grids would reuse observations already applied on the parent domain and count the same information twice. The experiment is therefore a test of resolution at fixed analysis, which is the quantity the referee is asking about. Assimilation carried out directly at convection-permitting scale needs its own error statistics and background error covariance, and we identify it as future work.

Changes in the manuscript. New Sect. 6.5 ‘Effect of horizontal resolution’ and Table 10 (p25). Sixth conclusion added (p25). The scope of the experiment is stated explicitly in the new Sect. 5.1 (p10, l165–180).

Comment 4 — AIFS is not purely comparable with GraphCast and FuXi

> ***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.*

We agree entirely, and this point was raised independently by the first referee. The revised manuscript states it in three places. The generalisation argument, that skill does not degrade on events outside the training period, applies only to GraphCast and FuXi, whose training data end in 2017; for those models the 2021 event is genuinely unseen and GraphCast in fact scores higher there (CSI 0.441) than on the in-training 2016 event (0.314). Because the AIFS fine-tuning window of 2016–2022 encompasses all three case studies, none is out-of-sample for that model, and its results are now described as an upper bound on what it would achieve on genuinely unseen events. The conclusions additionally note that AIFS inherits information from the full ECMWF operational assimilation system, which ingests far more observations than the network assimilated into WRF here.

> **Changes in the manuscript.** Abstract (p1, l18–20), Sect. 6.2.4 (p14) and Conclusions (p25) restrict the generalisation claim to GraphCast and FuXi and describe the AIFS results as an upper bound.

Comment 5 — Reproducibility details for the AI models

> ***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.*

All five items have been added as a dedicated table. It records the checkpoint or release used for each model, the framework through which it was run, the native grid, the time step and rollout length, the source of initial conditions, the exact names of the precipitation and temperature variables extracted, the unit conversion applied, and the interpolation method. We also state explicitly that no bias correction, calibration or downscaling was applied to any AI output, that ERA5 fields were used as distributed without regridding, and that WRF fields were extracted with the same nearest-neighbour method so that all four systems are sampled identically.

> **Changes in the manuscript.** New Sect. 4.1 and Table 3 (p9). The processing scripts are archived on Zenodo and referenced in the code availability statement.

Comment 6 — Quantitative spatial verification

> ***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.*

We are grateful for this suggestion, which has produced one of the strongest results in the revised manuscript. The three peak periods have been scored against the radar composite using the Fractions Skill Score at neighbourhood widths of 12, 25, 50 and 100 km and thresholds of 1, 5, 10 and 20 mm per 6 hours, with the spatial correlation of the accumulation fields as a secondary measure. All fields were interpolated to a common 0.02 degree mesh so that the scores reflect the forecasts rather than the native grids.

The quantitative scores confirm the visual impression and sharpen it. The assimilated WRF attains the highest FSS at every threshold and every neighbourhood width, and its spatial correlation with the radar field, 0.578, is more than twice that of the best AI system. The margin widens markedly with intensity: at 1 mm the systems are close (0.796 for WRF against 0.752 for AIFS and 0.746 for GraphCast), whereas at 20 mm WRF retains an FSS of 0.506 while AIFS falls to 0.177, GraphCast to 0.076 and FuXi produces no exceedances at all. FSS increases monotonically with neighbourhood width for every system, which indicates that the AI deficit is one of amplitude rather than of gross displacement, since enlarging the neighbourhood cannot recover intensities that

were never produced.

Two caveats are stated in the text. The AI systems have native grid spacings of 25–28 km, so the 12 km neighbourhood lies below the scale they can resolve and their scores in that column are a lower bound; and the sample is three 6-hour periods within a single event.

> **Changes in the manuscript.** New Sect. 6.4 ‘Spatial verification against radar’ and Table 9 (p24), citing Roberts and Lean (2008). Fifth conclusion extended with the FSS result (p25).

We thank the referee for a review that has substantially improved the manuscript. The requests for a quantitative spatial comparison and for an explicit treatment of the initial-condition imbalance both led to results that we consider central to the paper, and the second of them corrected a claim we would otherwise have stated too broadly.