https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

Physics vs. AI in Weather Prediction: Evaluating GraphCast, AIFS, and FuXi against an Observation-Corrected WRF Model for Flash Floods

Haseeb ur Rehman[1] and Félicia N. R. Teferle[1]

[1]University of Luxembourg, 2 Avenue de l'Université, L-4365 Esch-sur-Alzette, Luxembourg

Correspondence: Haseeb ur Rehman (haseeb.rehman@uni.lu)

Abstract. Data-driven artificial intelligence (AI) weather models are increasingly positioned as alternatives or complements to physics-based numerical weather prediction (NWP) systems, yet systematic model evaluation studies that compare both paradigms under controlled experimental conditions remain limited. Here we present a structured model evaluation framework applying standardised verification metrics to assess a regional WRF configuration against three publicly available AI models – GraphCast, AIFS Single 1.0, and FuXi – over three high-impact flood events in Luxembourg and the Greater Region (2016, 2018, and 2021), comprising 90 simulation days and over 7 300 matched forecast–observation pairs. The WRF model employs a three-dimensional variational (3D-VAR) data assimilation scheme ingesting GNSS Zenith Total Delay and conventional observations on a 6-hourly Rapid Update Cycle. All systems are initialised from ERA5 reanalysis to ensure consistent initial conditions. Categorical precipitation scores at the 1 mm threshold and continuous temperature metrics at surface stations serve as verification targets. Data assimilation measurably improves WRF's categorical precipitation skill (Critical Success Index: 0.306 to 0.341) while leaving near-surface temperature largely unchanged, demonstrating the value of the assimilation scheme. AIFS achieves the highest detection rate (POD 0.765) and net categorical skill (CSI 0.370), GraphCast and AIFS reduce temperature RMSE by $\sim$10 % relative to assimilated WRF, and WRF alone reproduces the observed mesoscale precipitation structure of the catastrophic July 2021 flood. The per-event breakdown reveals no skill degradation for AI models on out-of-sample events, suggesting that meteorological regime rather than training-data overlap governs model performance. These results provide a replicable model evaluation methodology for benchmarking emerging data-driven NWP systems against observation-corrected regional models.

1 Introduction

Accurate weather forecasting remains a critical challenge, particularly for predicting extreme events that can lead to catastrophic impacts such as the July 2021 European floods. For decades, conventional numerical weather prediction (NWP) models like the Weather Research and Forecasting (WRF) model have been the cornerstone of meteorological forecasting (Skamarock et al., 2021). These physics-based models rely on solving complex differential equations that govern atmospheric dynamics. However, they are computationally intensive and highly sensitive to initial conditions and parameterisation schemes (Prein et al., 2017; Nguyễn et al., 2025; Mekawy et al., 2023; Fosser et al., 2024).

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

Recently, artificial intelligence (AI) models have emerged as a disruptive technology in weather forecasting. Models such as GraphCast (Lam et al., 2023), Pangu-Weather (Bi et al., 2023), FourCastNet (Pathak et al., 2022), FuXi (Chen et al., 2023), the Artificial Intelligence Forecasting System (AIFS) (Lang et al., 2024), and Aurora (Bodnar et al., 2025), developed by leading research groups, leverage deep learning architectures and are trained on decades of historical reanalysis data. These data-driven models have demonstrated remarkable skill and computational efficiency (Kashinath et al., 2021; Wu and Xue, 2024; Bodnar et al., 2025; Weyn et al., 2021), often matching or exceeding the performance of conventional NWP models at global scales (Sham et al., 2025; Johny et al., 2026).

Early comparative studies have demonstrated the potential of AI to complement or outperform physics-based NWP models for short-term forecasting. Shirali et al. (2020) evaluated the WRF model against AI techniques for short-term rainfall, temperature, and flood forecasting in a regional case study. Their results showed that while WRF provided physically consistent simulations of atmospheric processes, AI-based models achieved higher accuracy in predicting rainfall intensity and flood peaks during extreme events. Zhang et al. (2026) showed that physics-based NWP models still outperform state-of-the-art AI systems for record-breaking extreme events, finding that AI models systematically underestimate the frequency and intensity of unprecedented extremes.

This study aims to conduct a comparative analysis between the physics-based WRF model (configured and tuned for Luxembourg and the Greater Region) and three AI-driven weather models: GraphCast, AIFS, and FuXi. By evaluating their performance across three high-impact flood events—the flash flood of 21–23 July 2016 in north-eastern Luxembourg, the Mullerthal flash flood of 31 May–2 June 2018, and the catastrophic 14–15 July 2021 European flood—we investigate the respective strengths and limitations of both approaches. Each event is simulated over a 30-day window (10 July–6 August 2016, 20 May–20 June 2018, and 20 June–20 July 2021). A single-domain (12 km) WRF setup with data assimilation, configured and tuned specifically for Luxembourg and the Greater Region, is used to provide a physics-based benchmark against the rapid inference capabilities of the AI models.

2 Study area

This study focuses on Luxembourg and the surrounding Greater Region (Fig. 1), a transboundary area whose contrasting terrain exerts a strong control on the hydrological response. Within Luxembourg itself, two physiographic units dominate the landscape. To the north lies the Oesling, the Luxembourg portion of the forested Ardennes massif, where mean elevations of roughly 450–500 m and steep, incised valleys favour rapid runoff. To the south, the Gutland comprises gentler, fertile lowlands with mean elevations near 200 m (Lucius, 1948; Müller, 1980; Waterlot and Beugnies, 1973). The region is drained by an interconnected river network: the Moselle, together with its tributaries the Sauer and the Our, forms the backbone of the system. Because many basins are small and steep, they respond quickly to convective rainfall, leaving only short lead times for warning.

The hazard was illustrated on 1 June 2018, when a near-stationary convective cell dropped extreme rainfall over the steep slopes of the Mullerthal and triggered severe local floods (Mathias et al., 2017). The July 2021 event was more catastrophic still,

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

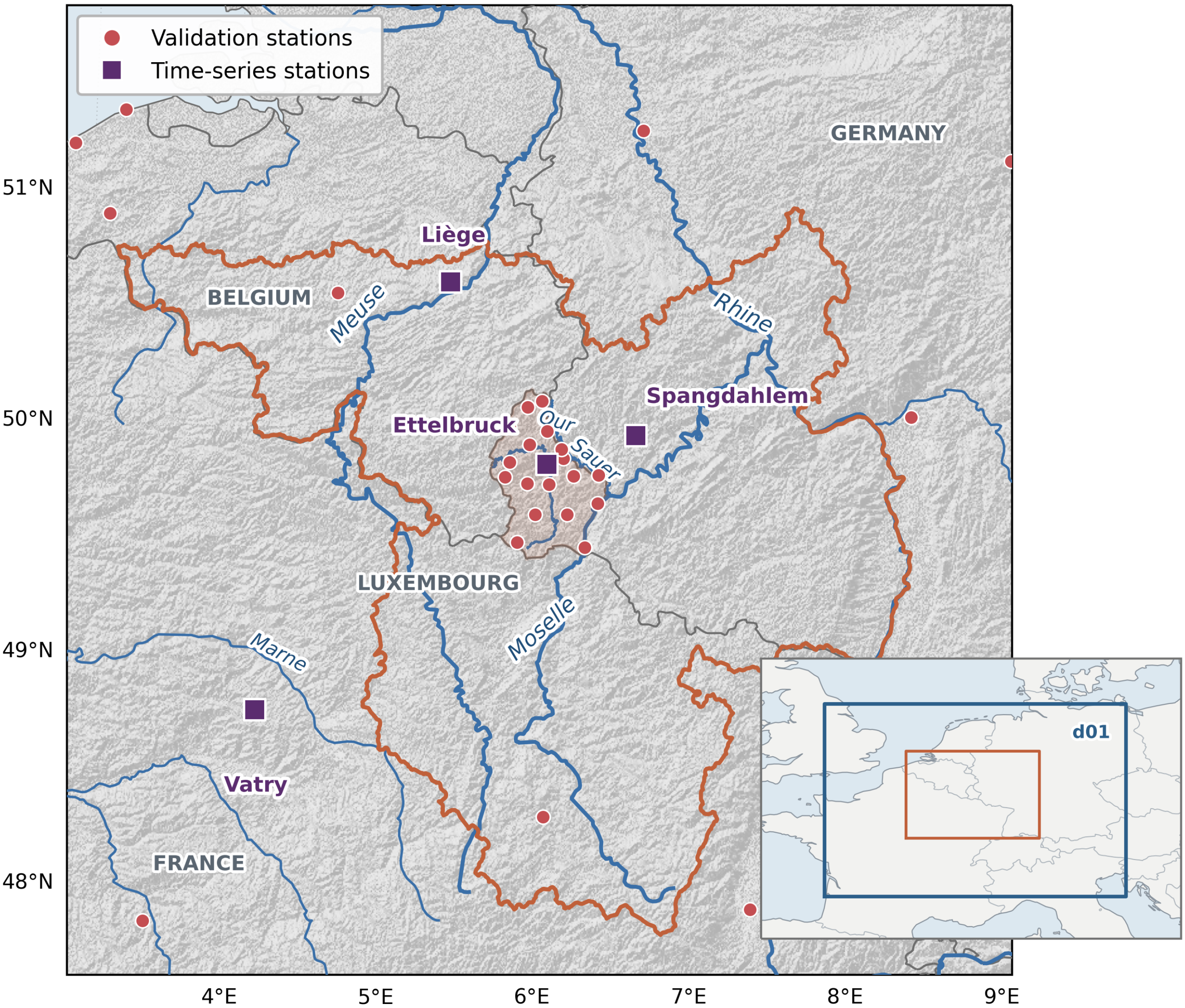

Figure 1. The study area. The main panel zooms on Luxembourg and the Greater Region (orange outline, with Luxembourg lightly shaded) over SRTM-derived hillshade topography. The principal rivers are drawn in blue and labelled along their courses. Red circles mark the surface stations used for precipitation validation; purple squares mark the four stations used for the event time series in Sect. 6.3 (Ettelbruck, Spangdahlem, Liège, and Vatry). The inset shows the single WRF domain (d01, blue, 12 km grid spacing) and the extent of the main panel over Western Europe.

with specific discharges exceeding $1300\,\mathrm{l\,s^{-1}\,km^{-2}}$ in small basins such as the Bibeschbach (Luxembourg Institute of Science and Technology, 2021). Such episodes underline the need for forecasting systems able to resolve localised, high-intensity
60 convection and the rapid runoff it generates.

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

Table 1. WRF physics parameterisation options used in this study.

Physics option	**Scheme**
Microphysics	Thompson et al. (Thompson et al., 2008)
Longwave radiation	RRTM (Mlawer et al., 1997)
Shortwave radiation	Dudhia (Dudhia, 1989)
Surface layer	Revised MM5 (Jiménez et al., 2012)
Land surface model	Noah LSM (Tewari et al., 2004)
Planetary boundary layer	YSU (Hong et al., 2006)
Cumulus parameterisation	Grell–Freitas GF (Grell and Freitas, 2014)

3 Atmospheric modelling system

3.1 Weather Research and Forecasting model

The WRF model is a community mesoscale NWP system developed by the National Center for Atmospheric Research (NCAR) and partner institutions (Skamarock et al., 2019). The present study uses the Advanced Research WRF (ARW) core, configured as a single domain at 12 km horizontal grid spacing covering Luxembourg and the Greater Region (Fig. 1). The domain comprises 120×120 horizontal grid points with 33 terrain-following vertical levels and a model top at 50 hPa, integrated with a 60 s time step. All simulations were performed with WRF version 4.5. The physics suite is summarised in Table 1.

3.2 WRF data assimilation strategy

The 3D-VAR assimilation mode within the WRFDA system was employed. The primary objective of 3D-VAR is to minimise a quadratic cost function representing the discrepancy between the model analysis and both the background field and observations (Lorenc, 2003; Barker et al., 2004):

$$J(x) = (x - x_b)^T B^{-1}(x - x_b) + (H(x) - y)^T O^{-1}(H(x) - y), \quad (1)$$

where x_b is the background field, B is the background error covariance matrix, H is the observation operator, y represents the observations, and O is the observation error covariance matrix. The background error covariance was configured using the `CV5` option with domain-specific statistics generated via the NMC method. The cost-function minimisation used two outer loops, and observations were accepted within a ± 1 h window centred on each analysis time.

3.3 NWP data and pre-processing

Initial and lateral boundary conditions were taken from the ERA5 reanalysis (Hersbach et al., 2020) at 0.25° resolution, sampled every 6 h across each 30-day window (Hersbach et al., 2023b, a). Three classes of conventional observations were

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

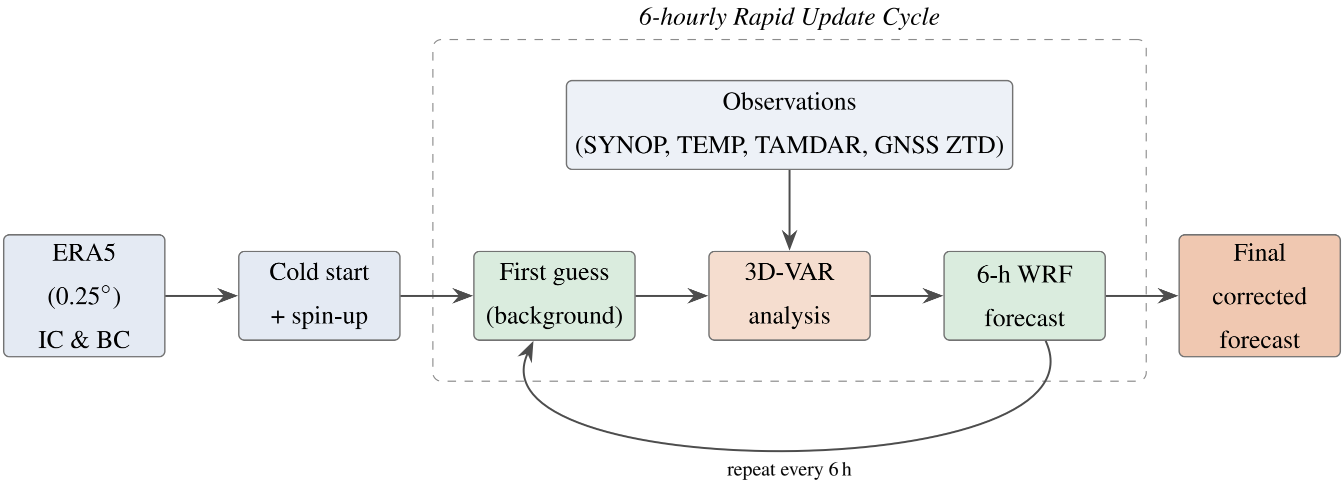

Figure 2. The WRF/WRFDA Rapid Update Cycle (RUC). The model is cold-started from ERA5 and spun up; thereafter, every 6 h the short-range forecast (first guess) is combined with conventional and GNSS ZTD observations in a 3D-VAR analysis that initialises the next 6-hour forecast. After repeated cycling, the final observation-corrected forecast is launched.

assimilated: SYNOP surface reports, TEMP radiosonde profiles, and TAMDAR aircraft measurements (Daniels et al., 2006). In addition, GNSS ZTD observations were obtained from the Nevada Geodetic Laboratory (Blewitt et al., 2018).

Observations were assimilated on a 6-hourly cycle following the Rapid Update Cycle (RUC) strategy (Benjamin et al., 2004). The model is first cold-started from ERA5 to produce a spin-up; thereafter, each analysis updates the short-range first guess with the available observations, and the corrected state initialises the next cycle (Fig. 2).

Precipitation and temperature observations were collected from the AgriMeteo network (Administration des Services Techniques de l'Agriculture (ASTA), 2025) and the NOAA Integrated Surface Database (National Centers for Environmental Information, 2021). Hourly composite weather-radar rainfall accumulations for the July 2021 event were obtained from the RADFLOOD21 product of the Royal Meteorological Institute of Belgium (Goudenhoofdt et al., 2023) and aggregated to 6-hour totals.

3.4 Post-processing

Meteorological variables were extracted from the WRF output using a custom Python workflow built on GDAL (GDAL/OGR contributors, 2023). Horizontal interpolation to station locations used the SciPy (Virtanen et al., 2020) `griddata` function, combining inverse-distance-weighting and linear schemes to preserve spatial consistency.

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

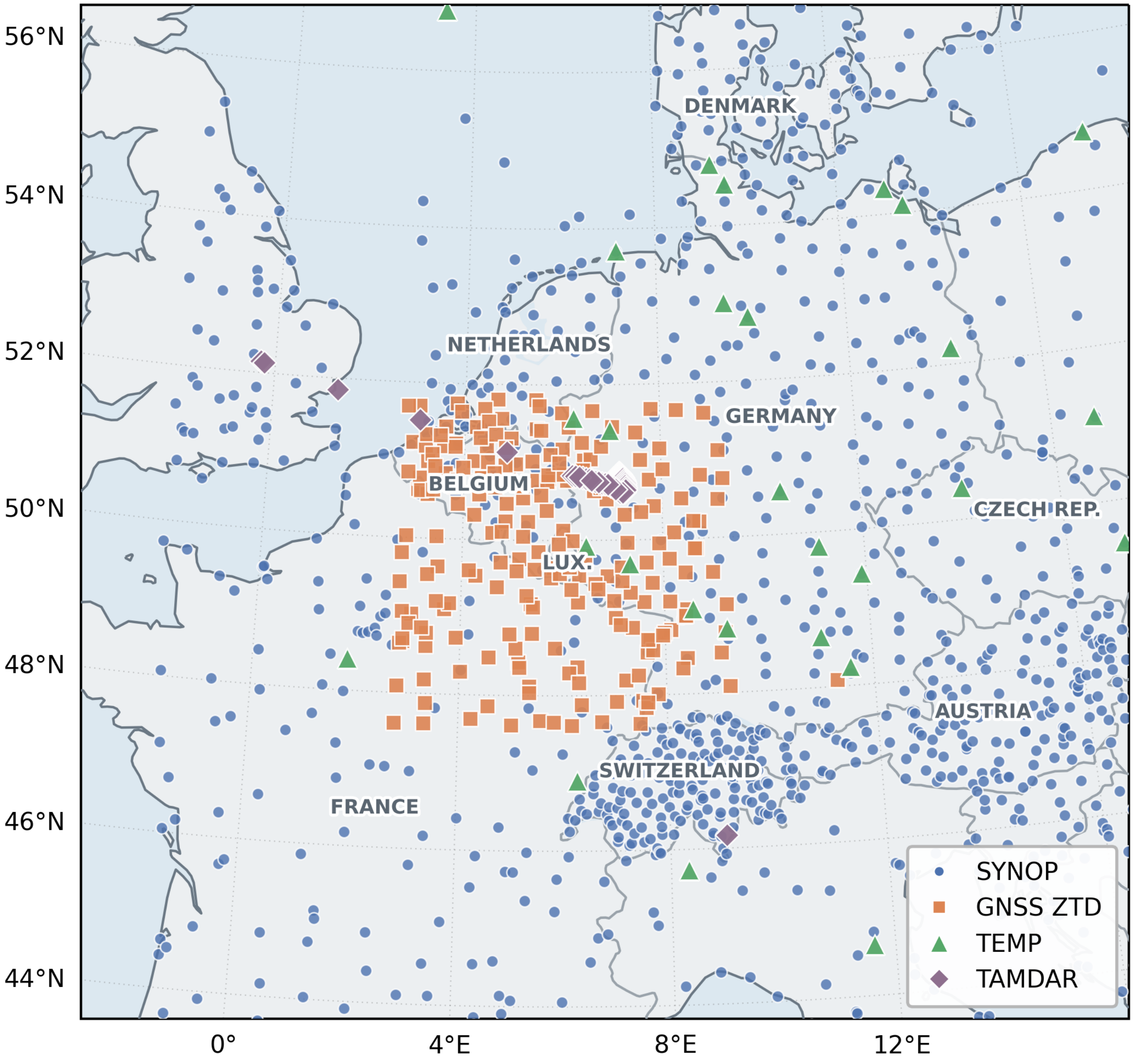

Figure 3. Spatial distribution of observations assimilated into the WRF 3D-VAR system: SYNOP surface stations (blue circles), GNSS ZTD sites (orange squares), TEMP radiosondes (green triangles), and TAMDAR aircraft reports (purple diamonds).

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

Table 2. Characteristics of the AI-based weather forecasting models evaluated in this study.

Characteristic	GraphCast	FuXi	AIFS Single 1.0
Institution	Google DeepMind	Fudan Univ. / Shanghai AI Lab	ECMWF
Architecture	Multi-scale GNN	Cascaded Transformer	Hybrid GNN + Transformer
Forecast rollout	Autoregressive 6-h	Cascaded autoregressive	Autoregressive 6-h
Pre-training	ERA5 (1979–2017)	ERA5 (∼1979–2017)	ERA5 (1979–2022)
Fine-tuning	None	None	IFS analyses (2016–2022)
Resolution	0.25° (∼25 km)	0.25° (∼25 km)	N320 (∼28 km)

4 AI-based weather forecasting models

This study evaluates three distinct data-driven weather forecasting models: GraphCast, AIFS, and FuXi. These models were selected because their publicly available implementations output precipitation, unlike other models such as Pangu-Weather or Aurora whose open-source versions do not expose precipitation fields, precluding a like-for-like comparison with our precipitation-focused WRF configuration.

GraphCast, developed by Google DeepMind, utilises a multi-scale graph neural network (GNN) trained on ERA5 reanalysis data from 1979 to 2017. It generates global forecasts at a horizontal resolution of 0.25° with a native 6-hour time step (Lam et al., 2023).

AIFS is the data-driven forecasting system developed by ECMWF (Lang et al., 2024). We use AIFS Single 1.0, the most recent AIFS release that can be initialised entirely from ERA5. The subsequent AIFS 2.0 release requires ocean-wave spectral variables available only in ECMWF operational analyses, which would break the consistency of initial conditions across the compared models. AIFS Single 1.0 operates on an N320 (∼28 km) grid and was fine-tuned using operational IFS analyses from 2016–2022.

FuXi, developed by Fudan University and the Shanghai AI Lab, is a cascaded transformer-based forecasting system (Chen et al., 2023). Like GraphCast, it was trained on ERA5 and operates at 0.25° horizontal resolution with 6-hourly steps. The key characteristics of the three AI models are summarised in Table 2.

5 Experimental design and verification

5.1 Case studies and simulation configuration

The two modelling systems were evaluated on three historic high-impact flood events over Luxembourg and the Greater Region: 2016, 2018, and 2021. Each event was simulated over a 30-day window (10 July–6 August 2016, 20 May–20 June 2018, and 20 June–20 July 2021), with forecasts initialised every 6 h (120 analysis–forecast cycles per event), giving 90 simulation days per

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

configuration across the three events. For WRF, two configurations were produced: a cold-start run without data assimilation (*Before DA*) and an observation-corrected run (*After DA*) using the 6-hourly RUC described in Sect. 3.3. WRF simulations were executed on 52 CPU cores of a high-performance computing cluster.

The three AI models were run on the same three events and 30-day windows. Forecasts were initialised from ERA5 analyses (at $t-6$ h and t) and advanced a single 6-hour step to the verification time; each model was re-initialised from ERA5 every 6 hours rather than rolled out freely. Inference was performed using a single GPU with 40 GB of VRAM.

5.2 Verification metrics

Because precipitation is intermittent and highly skewed, categorical scores derived from a 2×2 contingency table of forecast/observed rainfall occurrence above a $1.0\,\mathrm{mm}$ threshold are used, with hits a, false alarms b, misses c, correct negatives d, and total $n = a+b+c+d$:

$$\mathrm{POD} = \frac{a}{a+c}, \tag{2}$$

$$\mathrm{FAR} = \frac{b}{a+b}, \tag{3}$$

$$\mathrm{CSI} = \frac{a}{a+b+c}, \tag{4}$$

$$\mathrm{ETS} = \frac{a-a_r}{a+b+c-a_r}, \qquad a_r = \frac{(a+b)(a+c)}{n}, \tag{5}$$

where POD is the probability of detection, FAR the false-alarm ratio, CSI the critical success index, and ETS the equitable threat score. Near-surface temperature and precipitation amount were verified with continuous statistics over N forecast–observation pairs (F_i, O_i):

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i), \tag{6}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i|, \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}, \tag{8}$$

$$r = \frac{\sum_i (F_i - \bar{F})(O_i - \bar{O})}{\sqrt{\sum_i (F_i - \bar{F})^2 \sum_i (O_i - \bar{O})^2}}. \tag{9}$$

For POD, CSI, ETS, and r, higher is better; for FAR, MAE, and RMSE, lower is better; the bias is best near zero.

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

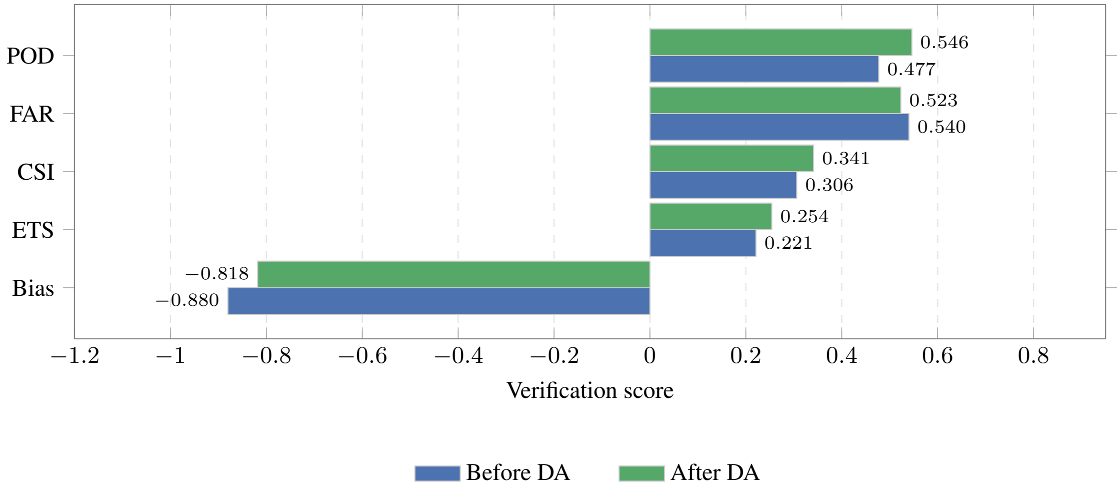

Figure 4. WRF precipitation verification before and after data assimilation, aggregated over the 2016, 2018, and 2021 events (90 simulation days). Higher POD, CSI, and ETS and lower FAR indicate better skill; the bias is the mean precipitation bias (best near zero).

6 Results and discussion

The results are presented in three stages. We first quantify the effect of data assimilation (Sect. 6.1), then benchmark WRF (After DA) against the three AI models (Sect. 6.2), and finally evaluate all systems against weather radar for the July 2021 flood (Sect. 6.3).

6.1 Impact of data assimilation on WRF forecasts

6.1.1 Precipitation

Assimilation improves every categorical precipitation score. POD rises from 0.477 to 0.546 ($+14.5\,\%$), FAR edges down from 0.540 to 0.523, CSI rises from 0.306 to 0.341 ($+11.4\,\%$), and ETS from 0.221 to 0.254 ($+14.9\,\%$). The mean precipitation bias moves toward zero ($-0.880 \rightarrow -0.818$), indicating that the assimilated moisture information—notably GNSS ZTD—partially relieves the model's tendency to under-forecast totals.

6.1.2 Temperature

Near-surface temperature is largely insensitive to assimilation. RMSE and MAE shift only marginally ($2.471 \rightarrow 2.455\,°$C and $1.846 \rightarrow 1.828\,°$C), the correlation dips slightly from 0.809 to 0.806, and the mean bias flips sign from $-0.076\,°$C to $+0.070\,°$C while remaining within $\pm 0.1\,°$C. These negligible changes are expected: the assimilated network is dominated by moisture-sensitive observations, so 6-hourly cycling chiefly reshapes the humidity and mass fields rather than the surface temperature.

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

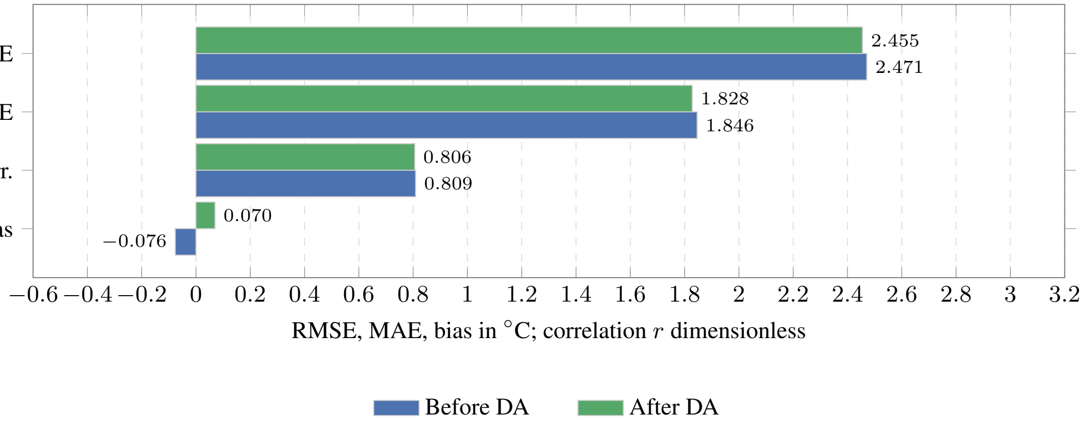

Figure 5. WRF 2 m temperature verification before and after data assimilation, aggregated over the 2016, 2018, and 2021 events. Lower RMSE and MAE are better, higher correlation is better, and the bias is best near zero.

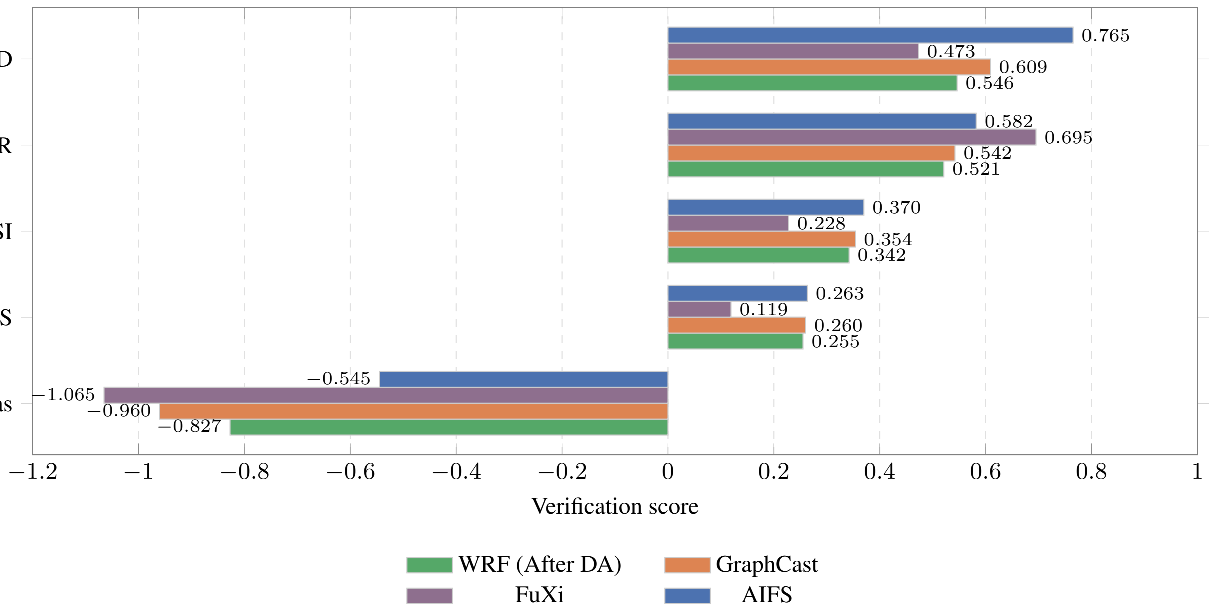

Figure 6. Precipitation verification of WRF (After DA) versus AI models over the common observation forecast pairs of the 2016, 2018, and 2021 events. Higher POD, CSI, and ETS and lower FAR indicate better skill; the bias is the mean precipitation bias (best near zero).

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

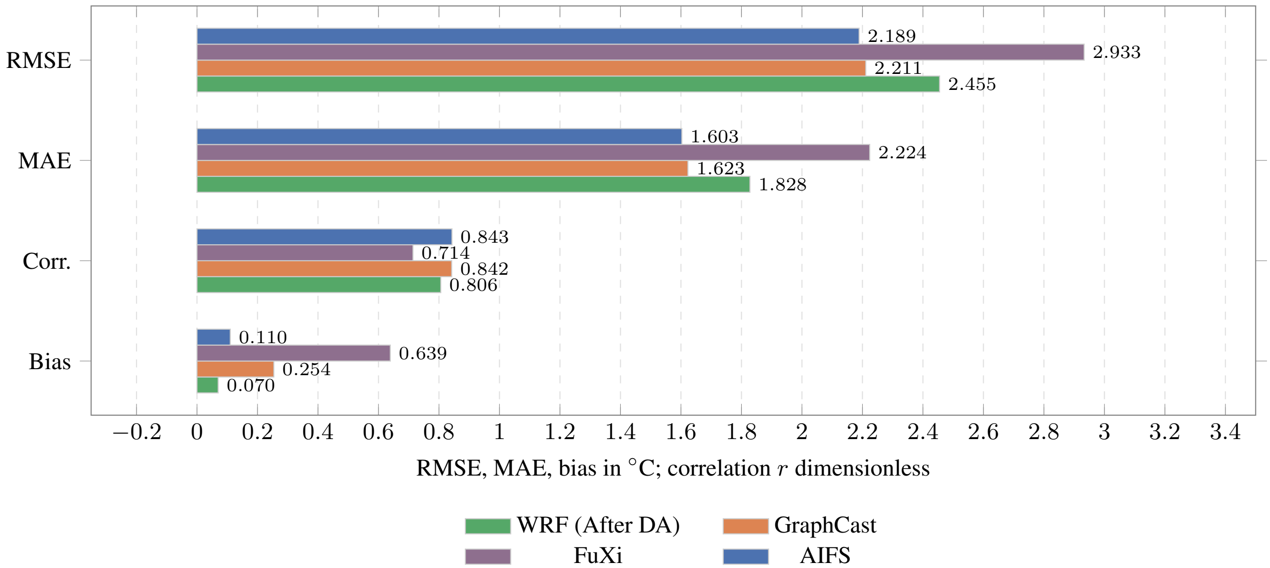

Figure 7. 2 m temperature verification of WRF (After DA) versus AI models over the common pairs of the 2016, 2018, and 2021 events. Lower RMSE and MAE are better, higher correlation is better, and the bias is best near zero.

6.2 WRF (After DA) versus AI models

6.2.1 Precipitation

The comparison reveals clear distinctions among the models. AIFS detects rain events far better than any other system (POD 0.765 vs. 0.609 for GraphCast and 0.546 for WRF), achieving the highest net skill (CSI 0.370, ETS 0.263). While all models underestimate total precipitation, AIFS has the smallest dry bias (-0.545 mm vs. -0.827 mm for WRF and over -0.960 mm for GraphCast and FuXi). GraphCast offers balanced performance, surpassing WRF in detection and overall categorical skill but with a stronger dry bias. FuXi consistently underperforms on both categorical metrics and bias. The assimilated WRF maintains the lowest false-alarm ratio (0.521).

6.2.2 Temperature

For near-surface temperature, AIFS and GraphCast show unambiguous advantages over WRF. AIFS attains the lowest RMSE (2.189 °C) and MAE (1.603 °C) alongside the strongest correlation (0.843). Both models track the temporal evolution of temperature better than WRF. However, WRF maintains the smallest mean bias ($+0.070$ °C). FuXi underperforms significantly, with the highest errors and a pronounced warm bias ($+0.639$ °C).

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

Table 3. Verification scores by event and pooled over all events for WRF (After DA) and the three AI models. Precipitation scores (POD, FAR, CSI, ETS, bias in mm) use the $1.0\,\mathrm{mm}\,(6\,\mathrm{h})^{-1}$ threshold; temperature scores (RMSE, MAE in °C, correlation r, bias in °C) are continuous. N is the number of matched forecast–observation pairs. Bold marks the best value per column within each event block.

	Precipitation					2 m temperature			
System	POD	FAR	CSI	ETS	Bias	RMSE	MAE	r	Bias
2016 ($N = 1819$, 17 stations)									
WRF (After DA)	0.568	**0.532**	0.345	0.277	**−0.038**	2.571	2.049	0.829	−0.429
GraphCast	0.534	0.555	0.321	0.251	−0.228	**2.295**	**1.869**	**0.862**	0.103
FuXi	0.483	0.589	0.285	0.215	−0.173	2.830	2.287	0.789	0.590
AIFS	**0.782**	0.566	**0.387**	**0.307**	−0.049	2.348	1.906	0.855	**−0.023**
2018 ($N = 2138$, 19 stations)									
WRF (After DA)	0.524	**0.549**	0.320	0.237	−0.225	2.303	1.836	0.832	0.178
GraphCast	0.573	**0.549**	0.338	0.252	−0.313	**1.991**	**1.594**	**0.871**	0.291
FuXi	0.433	0.755	0.185	0.078	−0.244	2.760	2.204	0.734	0.463
AIFS	**0.710**	0.586	**0.354**	**0.257**	**0.051**	2.011	1.599	0.867	**0.159**
2021 ($N = 3343$, 29 stations)									
WRF (After DA)	0.548	**0.502**	0.353	0.251	−1.642	2.485	1.704	0.767	0.272
GraphCast	0.655	0.534	**0.374**	**0.260**	−1.772	2.297	1.507	0.796	0.312
FuXi	0.491	0.690	0.235	0.105	−2.075	3.091	2.203	0.624	0.778
AIFS	**0.787**	0.586	0.372	0.242	**−1.196**	**2.209**	**1.440**	**0.807**	**0.151**
All events ($N = 7300$, 29 stations)									
WRF (After DA)	0.546	**0.521**	0.342	0.255	−0.827	2.455	1.828	0.806	**0.070**
GraphCast	0.609	0.542	0.354	0.260	−0.960	2.211	1.623	0.842	0.254
FuXi	0.473	0.695	0.228	0.119	−1.065	2.933	2.224	0.714	0.639
AIFS	**0.765**	0.582	**0.370**	**0.263**	**−0.545**	**2.189**	**1.603**	**0.843**	0.110

6.2.3 Skill by event

Because the three events hold different positions relative to each AI model's training period, the pooled scores are broken down by event in Table 3. The 2016 event lies inside GraphCast's training period (1979–2017) and FuXi's validation years (2016–2017); the 2021 event is entirely unseen by GraphCast and FuXi; for AIFS, all three events fall within its 2016–2022 fine-tuning window.

The per-event breakdown is remarkably stable: the model ranking established by the pooled scores holds in every individual event. Notably, GraphCast's categorical skill is in fact highest for the fully out-of-sample 2021 event (CSI 0.374 vs. 0.321 for the in-training 2016 event). This indicates that inter-event differences are governed primarily by the meteorological character of each episode rather than by training-data overlap, and tempers the concern that AI models' skill merely reflects memorisation of training samples. All systems show a much larger dry bias for 2021 (-1.2 to -2.1 mm), confirming systematic underestimation of the extreme rainfall amounts of that flood.

Figure 8 condenses the categorical precipitation results into a single performance diagram.

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

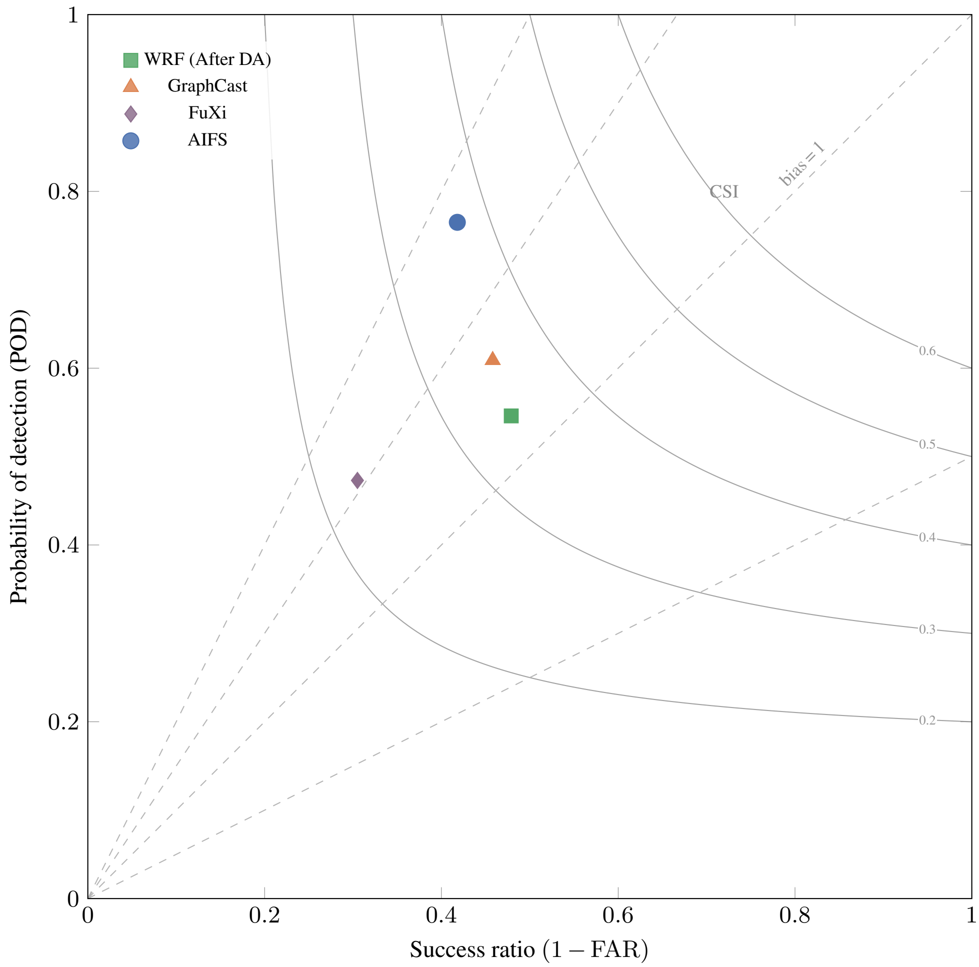

Figure 8. Performance diagram of categorical precipitation skill. Curved grey lines are constant CSI (0.2–0.6) and dashed grey lines constant frequency bias (0.5, 1, 1.5, 2). Points nearer the top-right corner are more skilful.

6.3 Event-scale evaluation: the July 2021 flood

The aggregated scores are complemented by a spatial evaluation against the Belgian weather-radar composite for the peak of the July 2021 flood. Hourly RADFLOOD21 radar accumulations (Goudenhoofdt et al., 2023) were aggregated to 6-hour windows. This evaluation is restricted to 2021 because RADFLOOD21 was released specifically for this event, whereas no equivalent open radar product covers 2016 and 2018. Figure 9 compares the 6-hour precipitation fields at three consecutive timestamps spanning the event peak (14 July 18:00 UTC to 15 July 06:00 UTC).

The radar shows an intense, elongated rain band stretching from Belgium across Luxembourg into western Germany. WRF (After DA) is the only system that reproduces this banded mesoscale structure and its sharp gradients. GraphCast captures the correct location of the rainfall envelope but is strongly smoothed and underestimates peak intensities. FuXi produces almost no

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

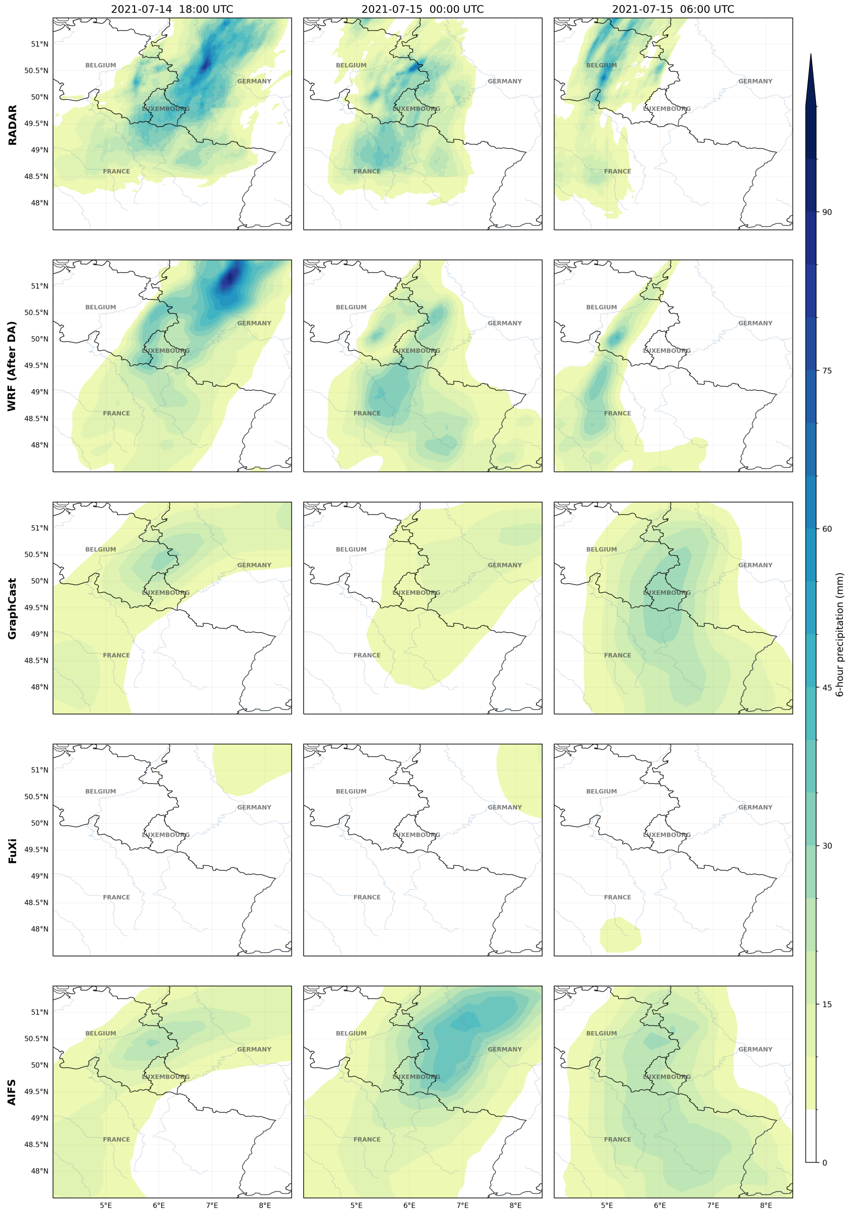

Figure 9. Six-hour precipitation fields over Luxembourg and the Greater Region at the peak of the July 2021 flood (columns: valid 2021-07-14 18:00 UTC, 2021-07-15 00:00 UTC, and 2021-07-15 06:00 UTC). Rows show the Belgian radar composite, WRF (After DA), GraphCast, FuXi, and AIFS. All model fields are 6-hour forecasts initialised 6 h before the valid time; the shared colour scale gives the 6-hour accumulation in mm.

precipitation signal during the peak, consistent with its poor categorical scores. AIFS produces the largest and most coherent precipitation envelope of the three AI systems—explaining its high detection rates—but its fields remain smooth and broad

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

relative to the radar. Thus, even where the aggregated skill of AIFS or GraphCast matches or exceeds WRF, the physics-based model retains a clear advantage in representing the fine-scale structure and amplitude of extreme convective rainfall.

190 Figure 10 shows 6-hourly precipitation time series from 10 to 17 July 2021 at one representative station in each country (Ettelbruck, Spangdahlem, Liège, and Vatry).

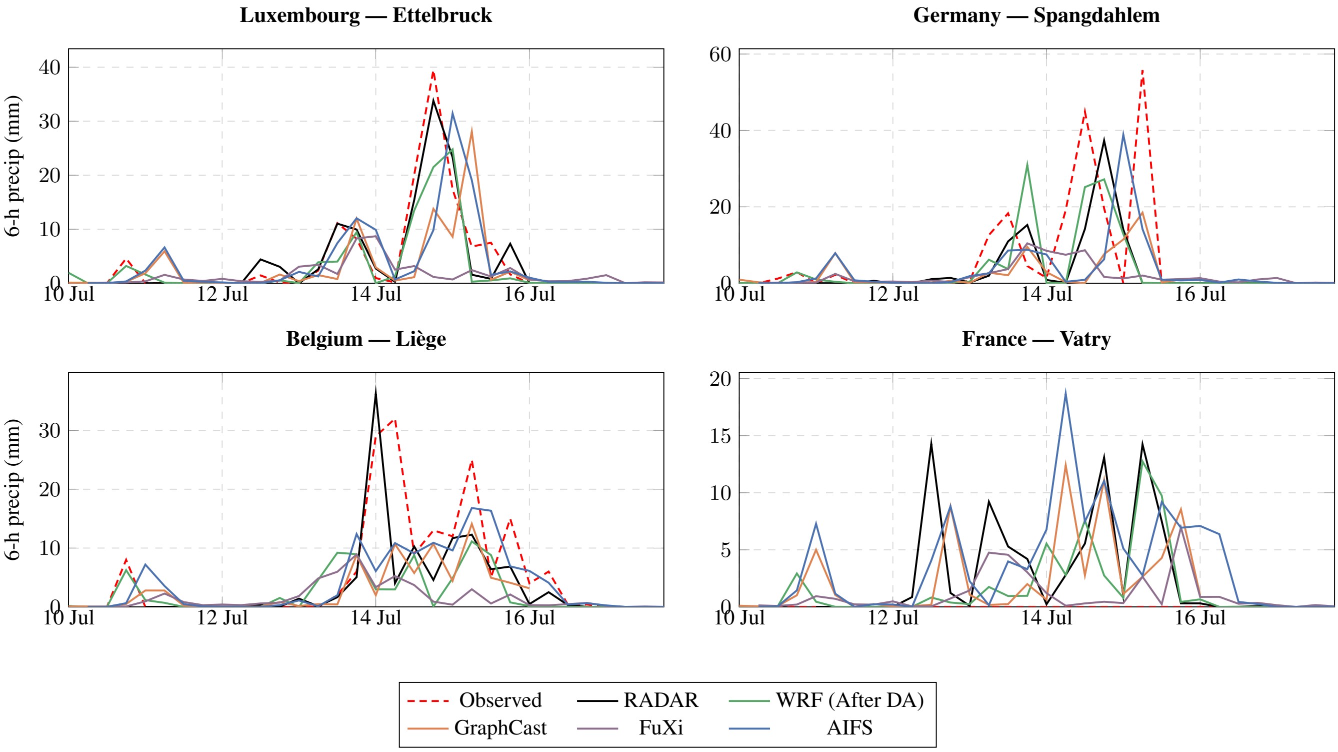

Figure 10. Six-hourly precipitation time series for 10–17 July 2021 at one validation station per country: gauge observations, the Belgian radar estimate, WRF (After DA), GraphCast, FuXi, and AIFS. Model values are 6-hour forecasts valid at each time, sampled at the nearest grid point to the station; gaps indicate periods without radar coverage.

The time series corroborate both the aggregated metrics and the spatial analysis. At Ettelbruck and Liège, AIFS and WRF capture the occurrence and approximate timing of the heaviest 6-hour totals, while GraphCast responds with attenuated amplitudes and FuXi remains near zero throughout the peak. Vatry presents an instructive contrast: its gauge recorded no rainfall, yet
195 every forecast system—and the radar estimate—indicates substantial precipitation, suggesting a localised miss or unreliable gauge rather than a simultaneous false alarm.

7 Conclusions

This study benchmarked an observation-corrected WRF configuration against three state-of-the-art AI weather models (GraphCast, AIFS, and FuXi) across three high-impact flood events (2016, 2018, and 2021) over Luxembourg and the Greater Region,

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

verifying precipitation and 2 m temperature over 90 simulation days per system. Several main conclusions emerge. First, 6-hourly 3D-VAR assimilation of conventional and GNSS ZTD observations successfully raises WRF's categorical precipitation skill. Second, AIFS detects rain events far better than all other models (POD 0.765) and achieves the highest overall net skill (CSI 0.370), while WRF maintains the lowest false-alarm ratio. Third, GraphCast and AIFS unequivocally surpass WRF and FuXi for temperature forecasting. Fourth, the per-event breakdown (Table 3) shows no skill degradation for GraphCast or FuXi on the out-of-sample 2021 event, indicating that the meteorological character of each episode outweighs training-data overlap.

Several caveats temper the apparent superiority of AIFS. All three case-study events lie inside its fine-tuning period, and AIFS inherits information from ECMWF's full operational data-assimilation system—vastly more observations than the conventional and GNSS ZTD network assimilated into WRF. Furthermore, AIFS operates on a coarser grid ($\sim$28 km), and its high detection rates partly reflect spatial smoothness that inflates POD at the expense of FAR. Finally, all AI models systematically underestimated the amplitude of extreme rainfall—a property of models trained with pixel-wise losses—so their categorical skill at 1 mm should not be extrapolated to the higher thresholds most relevant to flash-flood warning.

The AI models produce these forecasts on a single GPU, whereas WRF requires 52 CPU cores, underscoring the remarkable efficiency of data-driven inference. For operational flood forecasting in fast-responding catchments, these results suggest a complementary strategy: deploying AI models for rapid and highly skilful temperature fields and high-probability precipitation detection, while relying on observation-corrected WRF where low false alarms and faithful rainfall amounts are decisive. Future work will extend this comparison to further events and couple both forecasting systems directly to a hydrological model.

Code and data availability. All models and primary datasets used in this study are publicly available. The WRF and WRFDA source code (version 4.5) is distributed by NCAR and accessible from the WRF model users page (https://www2.mmm.ucar.edu/wrf/users/). ERA5 reanalysis fields are accessible from the Copernicus Climate Change Service Climate Data Store (https://cds.climate.copernicus.eu). The GraphCast code and pre-trained weights are available from Google DeepMind (Lam et al., 2023); the AIFS Single 1.0 checkpoint is published by ECMWF under a CC-BY-4.0 licence (https://huggingface.co/ecmwf/aifs-single-1.0); and the FuXi inference code and weights are available via the ECMWF `ai-models` package (https://pypi.org/project/ai-models-fuxi/) (Chen et al., 2023). GNSS ZTD estimates were obtained from the Nevada Geodetic Laboratory (http://geodesy.unr.edu), conventional precipitation observations from the NOAA Integrated Surface Database (https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database) and the AgriMeteo network (https://www.agrimeteo.lu), and the RADFLOOD21 composite weather-radar rainfall product from RMI Belgium (https://doi.org/10.5281/zenodo.7740059). The Python scripts used for ERA5 downloading, GNSS ZTD pre-processing, WRF Rapid Update Cycle execution, AI model variable extraction, radar aggregation, verification metric computation, and figure generation are archived on Zenodo at https://doi.org/10.5281/zenodo.20794937 (Rehman, 2026).

Video supplement. An animated version of the radar–model comparison covering the full 10–17 July 2021 period is available at https://doi.org/10.5446/73607 (Rehman and Teferle, 2026).

Author contributions. H. Rehman: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft. F. N. R. Teferle: Supervision, Funding acquisition, Writing – review and editing.

Competing interests. The authors declare that they have no conflict of interest.

Financial support. This work was supported by the Fonds National de la Recherche Luxembourg (FNR) under the Industrial Fellowship (IF) grant (Project No. 17130773).

Acknowledgements. The authors gratefully acknowledge the University of Luxembourg for computational resources (ULHPC) and institutional support. During the preparation of this work the authors used Google DeepMind's Gemini AI assistant in order to improve language readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Administration des Services Techniques de l'Agriculture (ASTA): AgriMeteo Luxembourg: Agro-meteorological Station Network, https://www.agrimeteo.lu, accessed: 2025-12-12, 2025.

Barker, D. M., Huang, W., Guo, Y.-R., Bourgeois, A. J., and Xiao, Q. N.: A Three-Dimensional Variational Data Assimilation System for MM5: Implementation and Initial Results, Monthly Weather Review, 132, 897–914, https://doi.org/10.1175/1520-0493(2004)132<0897:ATVDAS>2.0.CO;2, 2004.

Benjamin, S. G., Grell, G. A., Brown, J. M., Smirnova, T. G., and Bleck, R.: Mesoscale Weather Prediction with the RUC Hybrid Isentropic-Terrain-Following Coordinate Model, Monthly Weather Review, 132, 473–494, https://doi.org/10.1175/1520-0493(2004)132<0473:MWPWTR>2.0.CO;2, 2004.

Bi, K., Xie, L., Zhang, H., Chen, X., and Gu, X.: Accurate Medium-Range Global Weather Forecasting with 3D Neural Networks, Nature, 619, 533–538, https://doi.org/10.1038/s41586-023-06185-3, 2023.

Blewitt, G., Hammond, W. C., and Kreemer, C.: Harnessing the GPS Data Explosion for Interdisciplinary Science, Eos, 99, https://doi.org/10.1029/2018EO104623, 2018.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., Gupta, J. K., Thambiratnam, K., Archibald, A. T., Wu, C., Heider, E., Welling, M., Turner, R. E., and Perdikaris, P.: A foundation model for the Earth system, Nature, 641, 1180–1187, https://doi.org/10.1038/s41586-025-09005-y, 2025.

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., and Li, H.: FuXi: a cascade machine learning forecasting system for 15-day global weather forecast, npj Climate and Atmospheric Science, 6, 190, https://doi.org/10.1038/s41612-023-00512-1, 2023.

Daniels, T. S., Tseng, L. W., Moninger, W. R., Koch, S. E., Petty, G. W., Szoke, E. J., Birkenheuer, D., and Smith, T. L.: The Tropospheric Airborne Meteorological Data Reporting (TAMDAR) Sensor Campaign, in: Preprints, 10th Symposium on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface, American Meteorological Society, Atlanta, GA, 2006.

Dudhia, J.: Numerical Study of Convection Observed during the Winter Monsoon Experiment Using a Mesoscale Two-Dimensional Model, Journal of the Atmospheric Sciences, 46, 3077–3107, https://doi.org/10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2, 1989.

Fosser, G., Gaetani, M., Kendon, E. J., Adinolfi, M., Ban, N., Belušić, D., Caillaud, C., Careto, J. A. M., Coppola, E., Demory, M.-E., de Vries, H., Dobler, A., Feldmann, H., Goergen, K., Lenderink, G., Pichelli, E., Schär, C., Soares, P. M. M., Somot, S., and Tölle, M. H.: Convection-permitting climate models offer more certain extreme rainfall projections, npj Climate and Atmospheric Science, 7, 51, https://doi.org/10.1038/s41612-024-00600-w, 2024.

GDAL/OGR contributors: GDAL/OGR Geospatial Data Abstraction Software Library, 2023.

Goudenhoofdt, E., Journée, M., and Delobbe, L.: Observational rainfall data of the 2021 mid-July flood event in Belgium – Part 2. Radar product RADFLOOD21, https://doi.org/10.5281/zenodo.7740059, 2023.

Grell, G. A. and Freitas, S. R.: A Scale and Aerosol Aware Stochastic Convective Parameterization for Weather and Air Quality Modeling, Atmospheric Chemistry and Physics, 14, 5233–5250, https://doi.org/10.5194/acp-14-5233-2014, 2014.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., de Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Vil-

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

laume, S., and Thépaut, J.-N.: The ERA5 Global Reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 Hourly Data on Pressure Levels from 1940 to Present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.bd0915c6, 2023a.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 Hourly Data on Single Levels from 1940 to Present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.adbb2d47, 2023b.

Hong, S.-Y., Noh, Y., and Dudhia, J.: A New Vertical Diffusion Package with an Explicit Treatment of Entrainment Processes, Monthly Weather Review, 134, 2318–2341, https://doi.org/10.1175/MWR3199.1, 2006.

Jiménez, P. A., Dudhia, J., González-Rouco, J. F., Navarro, J., Montávez, J. P., and García-Bustamante, E.: A Revised Scheme for the WRF Surface Layer Formulation, Monthly Weather Review, 140, 898–918, https://doi.org/10.1175/MWR-D-11-00056.1, 2012.

Johny, K., Manoj, M., Fathima C, R., Shaji, A., and Sharma, O.: Bridging physics and machine learning: AI-enhanced WRF ensemble approach to extreme rainfall prediction, Applied Soft Computing, 197, 115 212, https://doi.org/10.1016/j.asoc.2026.115212, 2026.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., Hassanzadeh, P., and Prabhat: Physics-informed machine learning: case studies for weather and climate modelling, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, 20200 093, https://doi.org/10.1098/rsta.2020.0093, 2021.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., and Hu, W.: Learning Skillful Medium-Range Global Weather Forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Bouallègue, Z. B., Quinting, J., Rabier, F., Richardson, D., Prudhomme, C., Dopazo-Bueno, C., Hahner, T., Itterly, K., Manzanas, R., Pappenberger, F., Simmons, A., and Wedi, N.: AIFS – ECMWF's data-driven forecasting system, arXiv preprint arXiv:2406.01465, 2024.

Lorenc, A. C.: The Potential of the Ensemble Kalman Filter for NWP — A Comparison with 4D-Var, Quarterly Journal of the Royal Meteorological Society, 129, 3183–3203, https://doi.org/10.1256/qj.02.132, 2003.

Lucius, M.: Geologie von Luxemburg, Service Géologique du Luxembourg, Luxembourg, 1948.

Luxembourg Institute of Science and Technology: Hydrological Analysis of the July 2021 Flood Event in Luxembourg, Tech. rep., Luxembourg Institute of Science and Technology (LIST), Belvaux, Luxembourg, 2021.

Mathias, L., Ermert, V., Kelemen, F. D., Ludwig, P., and Pinto, J. G.: Synoptic Analysis and Dynamic Mechanisms of a Severe Convective Derecho Event over Western Europe, Weather and Forecasting, 32, 1589–1605, https://doi.org/10.1175/WAF-D-16-0194.1, 2017.

Mekawy, M., Saber, M., Mekhaimar, S. A., Zakey, A. S., Robaa, S. M., and Abdel Wahab, M.: Evaluation of WRF Microphysics Schemes Performance Forced by Reanalysis and Satellite-Based Precipitation Datasets for Early Warning System of Extreme Storms in Hyper Arid Environment, Climate, 11, https://doi.org/10.3390/cli11010008, 2023.

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., and Clough, S. A.: Radiative Transfer for Inhomogeneous Atmospheres: RRTM, a Validated Correlated-k Model for the Longwave, Journal of Geophysical Research: Atmospheres, 102, 16 663–16 682, https://doi.org/10.1029/97JD00237, 1997.

Müller, C.: Géographie physique du Luxembourg, Imprimerie Saint-Paul, Luxembourg, 1980.

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

National Centers for Environmental Information: Integrated Surface Database (ISD), NOAA National Centers for Environmental Information, 2021.

Nguyễn, T., Velásquez, P., Dietzel, A., and Cook, L.: Bias correction of precipitation from convection-permitting models at the point scale: a case study in Switzerland, Climatic Change, 178, https://doi.org/10.1007/s10584-025-04079-z, 2025.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., and Mardani, M.: FourCastNet: A Global Data-Driven High-Resolution Weather Model Using Adaptive Fourier Neural Operators, Proceedings of the National Academy of Sciences, 119, e2203585 119, https://doi.org/10.48550/arXiv.2202.11214, 2022.

Prein, A., Rasmussen, R., and Stephens, G.: Challenges and Advances in Convection-Permitting Climate Modeling, Bulletin of the American Meteorological Society, 98, 1027–1030, https://doi.org/10.1175/BAMS-D-16-0263.1, 2017.

Rehman, H. u.: Python scripts for WRF vs. AI weather model evaluation over Luxembourg flood events, https://doi.org/10.5281/zenodo.20794937, 2026.

Rehman, H. u. and Teferle, F. N. R.: Radar vs. WRF, GraphCast, FuXi, and AIFS precipitation comparison animation, 10–17 July 2021, https://doi.org/10.5446/73607, 2026.

Sham, F. A. F., El-Shafie, A., Jaafar, W. Z. B. W., Adarsh, S., Sherif, M., and Ahmed, A. N.: Improving rainfall forecasting using deep learning data fusing model approach for observed and climate change data, Scientific Reports, 15, 27 872, https://doi.org/10.1038/s41598-025-13567-2, 2025.

Shirali, E., Shahbazi, A. N., Fathian, H., Zohrabi, N., and Hassan, E. M.: Evaluation of WRF and artificial intelligence models in short-term rainfall, temperature and flood forecast (case study), Journal of Earth System Science, 129, 188, https://doi.org/10.1007/s12040-020-01450-9, 2020.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., Wang, W., Powers, J. G., Duda, M. G., Barker, D. M., and Huang, X.-Y.: A Description of the Advanced Research WRF Version 4, NCAR Technical Note NCAR/TN-556+STR, National Center for Atmospheric Research, Boulder, Colorado, https://doi.org/10.5065/1dfh-6p97, 2019.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., Wang, W., Powers, J. G., Duda, M. G., and Barker, D. M.: A Description of the Advanced Research WRF Model Version 4, Bulletin of the American Meteorological Society, 102, E112–E132, https://doi.org/10.1175/BAMS-D-20-0169.1, 2021.

Tewari, M., Chen, F., Wang, W., Dudhia, J., LeMone, M. A., Mitchell, K., Ek, M., Gayno, G., Wegiel, J., and Cuenca, R. H.: Implementation and Verification of the Unified Noah Land Surface Model in the WRF Model, in: Proceedings of the 20th Conference on Weather Analysis and Forecasting / 16th Conference on Numerical Weather Prediction, pp. 11–15, American Meteorological Society, Seattle, WA, 2004.

Thompson, G., Field, P. R., Rasmussen, R. M., and Hall, W. D.: Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization, Monthly Weather Review, 136, 5095–5115, https://doi.org/10.1175/2008MWR2387.1, 2008.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Waterlot, G. and Beugnies, A.: Ardenne, Guides géologiques régionaux, Masson, Paris, ISBN 9782225367847, 1973.

https://doi.org/10.5194/egusphere-2026-3655
Preprint. Discussion started: 20 July 2026

Weyn, J. A., Durran, D. R., Caruana, R., and Cresswell-Clay, N.: Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models, Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2021ms002502, 2021.

Wu, Y. and Xue, W.: Data-Driven Weather Forecasting and Climate Modeling from the Perspective of Development, Atmosphere, 15, https://doi.org/10.3390/atmos15060689, 2024.

Zhang, Z., Fischer, E., Zscheischler, J., and Engelke, S.: Physics-based models outperform AI weather forecasts of record-breaking extremes, Science Advances, 12, eaec1433, https://doi.org/10.1126/sciadv.aec1433, 2026.