Response to Referee 2 (RC2)

Manuscript: egusphere-2026-4222 **Title:** Evaluating the performance of a numerical weather prediction model for precipitation and temperature in Luxembourg and the Greater Region: insights from WRF and WRFDA 3D-VAR

We thank the referee for a specific and constructive review. The recommendation to run CV5 produced the clearest result in the revision and reframed the paper's conclusion; several sticky notes identified places where our text contradicted our own figures.

Referee comments are in *italics*; our responses follow.

Major points

RC2-1 — Clarify the simulation set-up before Figure 4

> *it is not clear, until the Conclusion section that the WRF model was operated in cycling mode for the whole month, with one simulation covering the whole period (if I correctly understood). I suggest to clarify the setting of the model run before, when discussing Figure 4. Also, Figure 4 must be improved to clarify how the model was run.*

The manuscript's description was **not accurate**.

The submitted text stated that "the assimilation increments were propagated forward continuously throughout the month, allowing the model to progressively integrate the information content of the observations". That is not what the runs do. The experiments use a **rapid-update (RUC) strategy in which every cycle is re-initialised from the concurrent GFS analysis**:

1. at analysis time T, the model is initialised from the GFS analysis valid at $T-6$ h;
2. a 6-hour WRF forecast provides the first guess;
3. WRFDA assimilates the observations to produce the analysis;
4. the analysis initialises the 6-hour forecast valid at $T+6$ h, which is what we verify;
5. the next cycle begins again at step 1 — **the analysis is not carried forward**.

We verified this three independent ways: the `fg` MD5 checksums that the run scripts log are byte-identical between the CV3 and CV5 experiments at every cycle sampled, while `wrfvar_output` differs; the O−B block of the WRFDA `statistics` file is byte-identical between the two configurations; and `qcstat_conv_01` matches in all 37 cycles of the chunk covering the flood. Crucially these remain identical at cycle 13 and cycle 27 after a cold start, which is only possible if the analysis is discarded each cycle.

The design is deliberate, not a defect. It prevents model drift accumulating over a month-long experiment, and it has an important benefit for this paper: **all experiments share an identical first guess at every analysis time**, so differences between them arise solely from the assimilation step. This is what makes the CV3-versus-CV5 comparison requested in RC2-2 a clean single-variable test.

Sect. 3 now describes this precisely, before the figure, and states explicitly that the analysis is not carried over between cycles. The Discussion notes that increments therefore do not compound — one reason the month-aggregated impacts are modest.

The cycle is now described step by step in the text, including forecast length (6 h), cycle frequency (6 h) and output frequency (see also sticky notes 13, 16, 21). The generic WRF flow chart of submitted Fig. 4 has been replaced by a study-specific schematic of the assimilation cycle, **revised Fig. 9**. It shows the GFS re-initialisation, the 6-hour first-guess forecast, the 3D-VAR step with its observations, and the verified 6-hour forecast; that the analysis is not carried over is stated in the caption.

RC2-2 — Use the CV5 option

> *I suggest that the authors perform an additional simulation using the CV5 option (and using the NMC method for computing the background error matrix) to improve the impact of data assimilation. This would also give an important sensitivity test.*

Done, and it changed the paper's central conclusion. Rather than adding a single CV5 run we repeated the *entire* experiment matrix with both covariances, so CV3 and CV5 can be compared for each observation configuration with nothing else varying: CONV, ZTD and CONV+ZTD, each with `cv_options=3` and with `cv_options=5` and an NMC-derived **B**, over the full month. With the no-assimilation control, that is seven analysed configurations.

In observation space, with an identical background and identical observations:

	O−B RMSE	O−A RMSE	Reduction
GNSS ZTD, CV3	1.143 cm	0.686 cm	**39.9 %**
GNSS ZTD, CV5	1.143 cm	0.389 cm	**66.0 %**

In forecast space, verified against the 264 independent stations (RC1-3), the contrast is sharper than we expected and reverses the sign of the result:

Threshold	CONV CV3	CONV CV5	CONV+ZTD CV3	CONV+ZTD CV5
0.1 mm/6 h	**−0.047**	−0.006	−0.002	**+0.031**
1 mm/6 h	**−0.022**	**+0.020**	**+0.025**	**+0.057**
5 mm/6 h	**−0.031**	+0.001	−0.000	+0.008
10 mm/6 h	**−0.048**	**−0.029**	−0.018	−0.014

(ΔETS against the control; **bold** = 95 % interval excludes zero under the grid-cell bootstrap. Of these, only the 1 mm gains for ZTD with CV5 and CONV+ZTD with CV5 also survive the day-block bootstrap; the 0.1 mm gains do not.)

With CV3, CONV assimilation is worse than no assimilation at every threshold tested, from 0.1 to 30 mm/6 h; the other two CV3 configurations are mixed rather than uniformly harmful. The frequency bias explains it: CV3 inflates it from 0.96 to 1.37 at 0.1 mm, so the model rains more often without discriminating better. CV5 holds the bias near unity (1.11) and turns the same observations into a significant improvement.

So the referee's suspicion was right, and stronger than suggested: the covariance was not merely limiting the impact, it was reversing its sign. The paper's conclusion is now that **for this domain the representation of background error matters more than the choice between conventional and GNSS observations** — confirming the prediction made in the first submission's own outlook.

What CV5 is, precisely. CV5 here is *not* flow-dependent: like CV3 it is held fixed through the assimilation. What differs is its origin — generated for this domain by the NMC method from forecast differences rather than a global climatology, and horizontally homogeneous within each vertical level. The covariance was generated from forecasts over the same June–July 2021 window the paper evaluates, so CV5 is an **in-sample** sensitivity experiment: it establishes that the covariance formulation changes the result, not the skill an operationally independent covariance would deliver. Generating one from an independent period is the immediate next step.

Two further caveats: the CV5 gain is confined to low and moderate intensities (nothing improves significantly at or above 10 mm/6 h), and it is regionally uneven — +0.099 over Belgium, +0.089 over Germany, +0.029 over France, and −0.001 over Luxembourg itself.

RC2-3 — O−B and O−A statistics for humidity and temperature

> *it would be interesting to see statistics of humidity and temperature before and after DA for the different configurations (O-B and O-A analyses) in order to evaluate the impact of different observations in the DA process.*

Done (this also answers RC1-5). Month-aggregated, observation-count-weighted, over 122 analysis cycles:

Obs type	Variable	Obs/cycle	O−B RMSE	O−A CV3	O−A CV5	Red. CV3	Red. CV5
GNSS ZTD	ztd (cm)	213	1.143	0.686	0.389	39.9 %	66.0 %
SYNOP	T (K)	371	1.628	1.204	1.150	26.1 %	29.4 %
SYNOP	q (kg/kg)	356	0.00110	0.00076	0.00099	30.8 %	10.2 %
TEMP	T (K)	1391	1.041	0.749	0.728	28.1 %	30.1 %
TEMP	q (kg/kg)	774	0.00047	0.00032	0.00031	31.3 %	33.4 %
TAMDAR	T (K)	915	1.068	0.774	0.771	27.6 %	27.8 %
TAMDAR	q (kg/kg)	44	0.00060	0.00048	0.00043	19.5 %	27.6 %

Quality control (mean per cycle, accepted / rejected): GNSS ZTD 213.3 / 2.1; SYNOP T 370.6 / 0.1; SYNOP q 355.5 / 0.8; TEMP T 1391.2 / 8.6; TEMP q 774.4 / 136.3; TAMDAR T 915.4 / 6.5.

Humidity behaves differently from temperature between the two covariances: CV5 improves radiosonde and TAMDAR humidity but is *worse* than CV3 for SYNOP surface humidity (10.2 % versus 30.8 %), which is consistent with the vertical structure functions of the domain-specific CV5 covariance spreading surface moisture information through the column rather than holding it at the surface. This is now discussed.

RC2-4 — Figures

> *Figure 5 … is unnecessary. Figure 4 must be improved. Figure 12 … is redundant because the comparison with RADAR was already shown in Figure 11 … No comparison are shown for temperature, which is improved by DA. Why don't use Figure 12 to show results for temperature.*

Figure numbers below are those of the submitted manuscript; revised numbers are given where the figure survives.

- **Submitted Fig. 5 (physics interactions): removed.** RC1 made the same recommendation, and we agree it added nothing specific to this study. **Submitted Fig. 4 (generic WRF flow chart) has also been removed** at RC1's suggestion, and replaced by the study-specific assimilation-cycle schematic described under RC2-1, which is **revised Fig. 9**.
- **Submitted Fig. 11 (radar comparison)** is **revised Fig. 10**; **submitted Fig. 12 (GPM IMERG)** is **revised Fig. 11**. The referee is right that the GPM figure duplicates the radar figure at the same times, and right that radar is the better reference over land.

We have retained it only as a brief independent check on the radar product and reduced the accompanying discussion accordingly.

- **All three spatial figures regenerated from the final experiment matrix.** The radar, GPM IMERG and spatial-temperature figures were produced from superseded runs in the first submission; they are now generated from the archived final forecasts, matching the same experiment set as the statistics.

- **Temperature comparison: added.** On the final runs the impact is 0.86 K RMS for CONV, 0.84 K for ZTD and 0.85 K for CONV + ZTD, with slight net cooling in all three (−0.15, −0.22, −0.18 K) — very similar magnitudes. The submitted abstract's claim that ZTD produced the largest, most coherent signal (RMS 1.49 K, mean −0.50 K) is withdrawn. This is the spatial counterpart of the station result — after height correction the configurations cannot be separated — so the figure is presented as the signature of assimilation, not evidence for a preferred experiment.

RC2-5 — Compare the conclusions with published work

> *no comparison is done with the results of other published papers. I suggest to do some comparison with published works to put the results of this paper in a wider context.*

Done. The Discussion now includes a dedicated subsection, "Comparison with published GNSS assimilation studies", in addition to the existing Discussion material.

The closest comparator is **Torcasio et al. (2023, NHESS)** — the same journal, the same model and assimilation system, and GNSS ZTD over Italy. Three points of contact matter:

1. **They used CV5.** Our result explains why that choice was consequential rather than incidental: with CV3, CONV assimilation degrades categorical skill at every threshold, and the other CV3 configurations are mixed rather than uniformly harmful. Our work supplies the controlled CV3-versus-CV5 contrast their configuration did not need to make.

2. **They verified against more than 3000 rain gauges.** That is two orders of magnitude beyond the 21 independent stations of our first submission, and it is the practical reason we rebuilt the network to 264 stations before the differences between configurations could be resolved at all.

3. **They report significant gains at 25–30 mm $(3\ \mathrm{h})^{-1}$.** Following that precedent, and the referee's own suggestion, we extended our thresholds from 10 to **30 mm $(6\ \mathrm{h})^{-1}$**. This is only meaningful with the expanded network: 30 mm/6 h has 156 observed events across 122 stations, whereas the original 21-station network would have had roughly a dozen.

Our ETS improvements with CV5 — between +0.03 and +0.06 at 0.1 and 1 mm/6 h — are of the same order as the few-percent gains reported in this literature, so the results are neither anomalously large nor negligible. **Giannaros et al. (2020)** report larger gains over Greece than ours, and gains that extend to intense rainfall: for accumulations above 20 mm, POD higher by 10.6 % (dry season) and 3.9 % (wet), FAR *lower* by 7.7 % and 16.9 %, and ETS higher by 11.4 % and 8.1 %, all statistically significant. Our earlier characterisation of that study — detection improving more than intensity — is corrected: it shows both improving. The comparison is instructive in two ways: their false alarm ratio falls where ours rises with CV3, and they derive their background error covariance for their domain by the NMC method from 24 h minus 12 h forecast differences over five control variables — the same construction we test as CV5. Both of the closest published comparisons therefore used a domain-specific covariance and obtained clear improvements, which is consistent with our finding that the covariance, not the observation type, governs the outcome.

On the background error covariance specifically, the literature leaves the question open. The WRFDA guidance states that it is difficult to say in advance which option performs better and that the impact varies case by case, and published comparisons are mostly CV5 against CV6 rather than against CV3. Our result is therefore a contribution to an open question rather than a confirmation of settled practice.

Added, previously missing: two co-author works — **Thundathil et al. (2024, GMD)** extended WRFDA with a tropospheric-gradient observation operator, and **Thundathil et al. (2025, AMT)** showed that adding gradients to zenith delays from a sparse network gives an improvement comparable to zenith delays alone from a network of twice the density. Since we assimilate zenith delays only, that points to a concrete route to further improvement over this domain without additional receivers, and it is now named as the next step in the Conclusions.

Poletti et al. (2019) combine radar nowcasting with numerical forecasts, and Vladimirov et al. (2020) assimilate conventional observations, satellite atmospheric-motion vectors and GNSS *radio-occultation refractivity* rather than ground-based zenith delays. Both are now described accurately and cited for the broader context of moisture-related assimilation rather than as directly comparable experiments.

The Discussion also explains why our ZTD impact is smaller than some published single-event studies: month-aggregated rather than single-event verification; a GFS-anchored design in which increments do not compound; a sparser GNSS network; point-gauge rather than spatial-field verification; the 12 km parameterized-convection limitation and the double-penalty problem; and the redundancy between GFS and the assimilated conventional observations.

Sticky-note comments

#	p	Comment	Response
1	1	Missing "." after Kuksina	Fixed.
2	1	Missing "." after Hapuarachchi	Fixed, together with the missing space.
3	1	Add more ZTD/CONV DA papers to the introduction	Added: Bennitt and Jupp (2012) on operational GPS ZTD assimilation at the Met Office, Lindskog et al. (2017) on ZTD assimilation in HARMONIE-AROME, and Guerova et al. (2016) reviewing ground-based GNSS meteorology in Europe.
4	2	"Moreover, The" → "the"	Fixed.
5	2	Missing "." after Tom; explain why	Full stop fixed, and the explanation added: rapid urban expansion replaces permeable ground with impervious surfaces, raising runoff volumes and shortening time to peak discharge, while settlement extends onto floodplains faster than drainage and warning capacity can follow.
6	4	"juillet" = July	The LIST figure is reproduced unaltered, so its French labels are retained; the caption now glosses *Débit*, *Pluie* and *14 & 15 juillet 2021*.
7	4	Use English labels/axes	As above — we cannot relabel a reproduced figure, so we gloss it in the caption. If the editor prefers, we can redraw it from the underlying series.
8	5	"Laura Poletti" → "Poletti"; fix in references	Fixed. The `.bib` author field read `Laura Poletti, Maria`, which BibTeX parsed as the surname "Laura Poletti"; it now reads `Poletti, Maria Laura` and the citation renders as "Poletti et al. (2019)".

#	p	Comment	Response
9	5	Move highlighted sentence to discussion	Done. The results-preview sentence has been removed from the Introduction; the corrected finding is stated in the Discussion and Conclusions instead. It also no longer held as written: both of its claims were withdrawn by the categorical and height-corrected analyses.
10	6	Duplicate of lines 98–99	Duplicated climate description removed; only the quantitative detail retained.
11	6	Why "1–12 km" resolution range?	Claim removed. The text now simply states the resolution used (12 km) and justifies it.
12	6	Cite tables in order	Fixed.
13	6	Figure 4: how is the DA cycle configured? Forecast length?	Answered in the text (see RC2-1): 6-hour cycle, 6-hour forecast, GFS re-initialisation each cycle, analysis not carried forward.
14	7	Is the physics from previous optimisation studies?	Stated explicitly: the physics was **not** tuned for this study, follows earlier WRF applications over the Greater Region, and is held fixed across all experiments so differences are attributable to assimilation. A physics sensitivity study is noted as outstanding.
15	9	Delete Figure 5	Deleted.
16	9	Figure 4 unreadable; describe the DA cycle in text; is it continuous over 1 month?	Figure redrawn and the cycle described in text. To answer directly: it is a continuous month-long *sequence* of 6-hourly analysis/forecast cycles, but each cycle is re-initialised from GFS — it is not a single continuous simulation.
17	10	H = forward operator	Fixed.
18	10	Cost function should carry ½	Fixed — ½ now on both terms.
19	10	x is the state vector, not the analysis	Fixed; the analysis is now defined as the state vector minimising J.
20	10	(A) Is $\mathbf{R}$ diagonal? What errors? (B) CV3 is not optimal	(A) Stated: $\mathbf{R}$ is diagonal, variances from the WRFDA observation-error table, GNSS ZTD error 0.8 cm, and the seven `max_error` gross-error factors are now listed in Sect. 3.2 with their values (t 4.0, uv 3.0, tb 3.0, rf 3.0, rv 2.5, q 1.5, pw 5.0, the last governing the GNSS delays). (B) Addressed by the CV5 experiment (RC2-2).
21	11	Which GFS cycle? Forecast length? Output frequency?	All four daily GFS cycles are used, each matched to the corresponding analysis time. Now stated in Sect. 2.
22	11	ERA5 is a reanalysis — did you use raw observations?	Important clarification, now made explicit: we assimilated the **raw** BUFR observation reports from the ECMWF archive, converted to LITTLE_R. No gridded ERA5 field was ingested.
23	12	"observed stations" → surface stations?	Fixed.
24	12	Add "inside Luxembourg"	Fixed.
25	14	Cite figures in order	Fixed (Figure 3 caption).
26	16	Lower-case "before"/"after"	Fixed.
27	17	Put the significance statistics in Results	Done — the method stays in Verification Methodology, the outcome moved to Results.

#	p	Comment	Response
28	20	Text contradicts Fig. 11: CONV appears to *increase* rainfall	**The referee is correct.** Our text said the opposite of our figure. The paragraph has been rewritten: CONV increases both the area and intensity of forecast rainfall. This is confirmed quantitatively — frequency bias rises, for CONV with CV3 at 0.1 mm/6 h, from 0.96 in the control to 1.37 (RC1-2). That figure applies to CONV with CV3 specifically; the other CV3 configurations reach 1.16 and 1.15 at the same threshold, so we do not describe all CV3 experiments as behaving alike.
29	20	"before DA" is misleading → use CTRL	Adopted throughout: the run without assimilation is now CTRL, defined where the experiments are introduced.
30	20	Again: CONV gives the largest rainfall, ZTD reduces it	Confirmed and corrected. The revised text states that CONV produces the most widespread rainfall, ZTD the least, and CONV + ZTD in between — exactly as the referee describes.
31	21	Figure 12 redundant; why GPM over land when radar exists?	The GPM figure was retained rather than replaced, as an independent spatial cross-check on the radar product, and a separate temperature figure was added; we have not substituted one for the other, and say so plainly. Both captions now state that each field is drawn on its own native grid with no regridding, that all panels are 6-hour accumulations in mm, and that the two figures do not share a colour scale because each is scaled to its own observation maximum. On why the model fields differ between the two figures: they contain the same native-grid model output but use different observation-based colour scales, which changes their visual appearance; this is now stated in the caption.
32	22	The colour threshold varies with time (70 mm at 18 UTC)	Correct. The caption now states that the colour maximum is set per time step from the reference field, with examples (35 mm at 00:00, 70 mm at 18:00).
33	22	CONV DA: rainfall more dispersed but not less intense	Confirmed and corrected in the rewritten paragraph.
34	22	The large false alarm over Germany at 12 UTC is never mentioned	Now stated explicitly in the text as a spurious rainfall maximum with no counterpart in the radar field, and linked to the frequency-bias increase.
35	23	Rising FAR suggests assimilation increases precipitation, contrary to the authors' comment	Correct, and now the paper's stated interpretation. Quantified with frequency bias and ETS at six thresholds, 0.1 to 30 mm/6 h (RC1-2).
36	24	The trade-off can be quantified by the Equitable Threat Score	Done — ETS (and CSI, FBI) added at 0.1, 1, 5, 10, 20 and 30 mm/6 h. ETS falls at 0.1 mm with CV3 and rises with CV5; at 5 mm and above no configuration improves significantly on the control.