Response to Referee 1 (RC1)

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 detailed review. Addressing the comments changed several of the paper's conclusions, including withdrawing two claims that did not survive the diagnostics requested below.

Referee comments are in *italics*; our responses follow. Line and page numbers refer to the revised manuscript unless stated otherwise.

Specific comments

RC1-1 — Restructure the manuscript

> *Please restructure manuscript in given form Introduction → Data → Model and Experimental Design → Verification Methodology (current Sect. 2) → Results → Discussion (a genuine discussion section) → Conclusions*

Done, exactly as requested. The manuscript now reads:

§	Section
1	Introduction
2	Data
3	Model and Experimental Design
4	Verification Methodology
5	Results
6	Discussion
7	Conclusions and Outlook

Study area and event description (Tables 1–2, Fig. 2) → Data. WRF/WRFDA formulation and experimental design → Model and Experimental Design. Accuracy Assessment Metrics → Verification Methodology. Results and Discussions → split into Results and a new Discussion (RC1-6).

RC1-2 — POD/FAR cannot distinguish skill from a frequency increase

> *A uniform increase in forecast rain frequency will produce exactly this pattern with no change in discrimination. Please add, for the same 0.1 mm threshold and for at least two higher thresholds (e.g. 1, 5, 10 mm/6 h) e.g. Frequency bias, ETS/CSI, etc*

The referee's hypothesis is correct. We now report FBI, POD, FAR, CSI and ETS at six thresholds — 0.1, 1, 5, 10, 20 and 30 mm/6 h — against the rebuilt 264-station network (RC1-3), with confidence intervals from a bootstrap that resamples model grid cells.

Threshold 0.1 mm / 6 h

Experiment	FBI	POD	FAR	CSI	ETS	ΔETS vs no DA	95 % CI
No DA (CTRL)	0.96	0.679	0.294	0.529	**0.383**	—	—

Experiment	FBI	POD	FAR	CSI	ETS	ΔETS vs no DA	95 % CI
CONV CV3	1.37	0.813	0.408	0.521	0.337	**-0.047**	[-0.055, -0.038]
CONV CV5	1.25	0.794	0.364	0.546	0.377	-0.006	[-0.015, +0.003]
ZTD CV3	1.16	0.751	0.353	0.533	0.368	**-0.015**	[-0.025, -0.004]
ZTD CV5	1.09	0.754	0.306	0.566	0.416	**+0.032**	[+0.021, +0.044]
CONV+ZTD CV3	1.15	0.757	0.343	0.543	0.381	-0.002	[-0.013, +0.009]
CONV+ZTD CV5	1.11	0.762	0.312	0.566	0.414	**+0.031**	[+0.019, +0.042]

Threshold 1 mm / 6 h

Experiment	FBI	POD	FAR	CSI	ETS	ΔETS vs no DA	95 % CI
No DA (CTRL)	0.86	0.567	0.344	0.437	**0.351**	—	—
CONV CV3	1.46	0.755	0.482	0.443	0.329	**-0.022**	[-0.034, -0.010]
CONV CV5	1.25	0.724	0.422	0.473	0.371	**+0.020**	[+0.009, +0.031]
ZTD CV3	1.23	0.703	0.427	0.461	0.358	+0.007	[-0.005, +0.019]
ZTD CV5	1.07	0.680	0.364	0.490	0.398	**+0.047**	[+0.033, +0.060]
CONV+ZTD CV3	1.16	0.693	0.400	0.474	0.377	**+0.025**	[+0.013, +0.038]
CONV+ZTD CV5	1.09	0.695	0.361	0.500	0.408	**+0.057**	[+0.044, +0.069]

The referee's diagnosis is exactly right for the configuration used in the first submission. With the climatological **CV3** covariance the model simply rains more often: FBI rises from 0.96 to 1.37, FAR from 0.29 to 0.41, and ETS *falls* significantly. **CONV with CV3 is worse than no assimilation at every one of the six thresholds.** The other CV3 configurations are mixed rather than uniformly harmful: CONV+ZTD with CV3 is significantly *better* than the control at 1 mm/6 h (+0.025). We have narrowed the claim accordingly throughout the manuscript.

This is a property of the background error covariance, not the observations: with the NMC-derived **CV5** covariance the same observations improve the score while frequency bias stays near unity — CONV+ZTD gains +0.031 at 0.1 mm and +0.057 at 1 mm.

Grid-cell resampling handles station dependence within a cell but treats the 121 verification times as fixed. Because precipitation persists between cycles, we repeated every comparison with whole calendar days as the resampling unit (31 blocks) — a more demanding test: 4 of 36 comparisons survive it, against 17 under cell resampling; we report both. At 1 mm/6 h the CV5 gains survive under both schemes (+0.047 for ZTD, day-block CI [+0.011, +0.087]; +0.057 for CONV+ZTD, [+0.020, +0.097]), as does the CONV-with-CV3 degradation at 0.1 mm and 30 mm. The gain at 0.1 mm/6 h does not survive (day-block interval includes zero). The positive conclusion now rests on the 1 mm threshold; the 0.1 mm gains are described as suggestive.

Two limits, stated explicitly in the paper:

1. **The gain is confined to low and moderate intensities.** At 10 mm/6 h and above, no configuration improves on the control. We make no claim of improved categorical skill at the intensities most relevant to flooding.

2. **The gain is regionally uneven.** For CONV+ZTD with CV5 at 1 mm/6 h it is +0.099 over Belgium, +0.089 over Germany, +0.029 over France and −0.001 over Luxembourg itself. Since this rests on four independent national networks it is a property of the forecast, not of the station sample.

Thresholds extend to 30 mm/6 h, following the referee's suggestion and the range used by Torcasio et al. (2023, NHESS) — only possible with the expanded network: 156 observed events across 122 stations at 30 mm/6 h, versus roughly a dozen on the 21-station network of the first submission.

RC1-3 — Independence of the validation observations

> *The validation stations must be shown to be independent of the assimilated observations.*

The verification network was not independent. We rebuilt it from scratch.

What we found about the original network. Matching every assimilated SYNOP report in all 124 `obs_gts` files against the 34 validation stations by coordinate, 13 were within one 12 km grid length of a station assimilated in essentially every cycle. The NOAA ISD subset is worse than that figure suggests: NOAA ISD is largely a re-archive of the same GTS synoptic feed, and **12 of the 15 ISD stations carry the WMO identifier of an assimilated report**, at distances of 0.00–5.17 km — Zeebrugge 0.00 km to WMO 06418, Beitem 0.21 km to 06414, Ernage 0.60 km to

1. These are the same physical stations reaching us through two archives.

The rebuilt network: 264 stations in 234 model grid cells, from four national providers:

Country	Provider	Stations	Grid cells
Luxembourg	AgriMeteo / ASTA	34	27
Belgium	SPW Hydrométrie Wallonie	33	33
Germany	DWD Climate Data Center	105	90
France	Météo-France	92	86
Total		**264**	**234**

The screening rule. A station is accepted only if it lies at least one 12 km grid length from *every* SYNOP report assimilated in *any* of the 124 observation-input files spanning the study window. The assimilated set varies between 1273 and 1351 stations per cycle and comprises 1361 distinct stations over the month, so the test uses the union across all cycles, not a single snapshot.

A second test the first version omitted. For the Belgian gauges we also screened against the 234 GNSS sites whose zenith total delays are assimilated, because a gauge co-located with an assimilated receiver is not independent for the ZTD and CONV+ZTD experiments. This is not redundant: it removes **28 of the 61** Belgian gauges that pass the SYNOP test, including one lying 121 m from the GNSS site of the same name (KAIN), and others at 2.0 km (WAREMME/WARE) and 3.3 km (ARLON/ARLO). Applying both tests leaves one station set that is independent for every experiment, so the configurations stay directly comparable.

Belgium required a different network entirely. All 14 RMI automatic weather stations match the identifier of an assimilated SYNOP station, so the Belgian AWS network *is* the Belgian SYNOP network. The SPW hydrological rain-gauge network is separate from the synoptic network and is the only route to independent Belgian verification. Its 33 accepted gauges each fall in a grid cell no other station uses, so they add 33 genuinely new cells.

Grid cells, not stations, are the effective sample size. Stations sharing a 12 km cell are handed an identical model value, so they are not independent samples of model error. The 264 stations occupy 234 distinct cells, against 21 in the first submission, and the bootstrap resamples cells.

Data integrity and time base were verified, not assumed. All model output is complete over the period with no missing or duplicated valid time and no invalid value. Because WRF output is UTC, each network's time base was checked by lag-correlating network-mean observed against network-mean modelled precipitation: all four peak at **zero lag** (r = 0.78–0.90), confirming the AgriMeteo local-time shift, the DWD and Météo-France UTC archives and the Belgian UNIX-millisecond timestamps are all correctly aligned.

RC1-4 — Lapse-rate correction for T2m

> *Please state explicitly whether a lapse-rate correction was applied to T2m before verification, apply one ... report the distribution of station-model height differences, and repeat the temperature statistics. Do the same for surface pressure if it is used quantitatively.*

No correction had been applied. We have now applied one, and it changed the paper's temperature conclusion.

The temperature network uses the same independence screening as RC1-3, extended to every station that reports air temperature: 34 in Luxembourg, 92 in France (Météo-France hourly files) and 22 of 105 in Germany (DWD's temperature archive; the other 83 report precipitation only). No Belgian station contributes — SPW is hydrometric. This gives **148 stations, 121 six-hourly times, 17,774 pairs per configuration**, all screened to at least 12.5 km from any assimilated SYNOP position.

Height differences. Model terrain is taken from the `HGT` field at the same grid point from which the station value is extracted; station heights come from the agrometeorological list in Luxembourg, the DWD station description file in Germany and the `ALTI` field of the hourly records in France. Across the 148 stations, model minus station height is:

- mean **+14.9 m**, median **+14.6 m**, standard deviation **63.3 m**
- range **−183.3 m** to **+238.3 m**
- 55 stations differ by more than 50 m, 17 by more than 100 m
- 91 of the 148 lie below the model surface

Corrected statistics. Applying $\Gamma = 6.5$ K km^{-1} (the ICAO value the referee suggested):

Experiment	Bias raw	Bias corrected	MAE corrected	RMSE corrected
No DA (CTRL)	−0.758	−0.663	1.584	2.031
CONV CV3	−0.557	**−0.462**	**1.496**	**1.991**
CONV CV5	−0.512	−0.417	1.508	2.002
ZTD CV3	+0.335	+0.430	1.759	2.269
ZTD CV5	−0.574	−0.479	1.618	2.122
CONV + ZTD CV3	−0.470	−0.375	1.532	2.054
CONV + ZTD CV5	−0.488	−0.393	1.547	2.062

This overturns a headline claim. The submitted abstract stated that ZTD was the most effective configuration for temperature bias (+97.4 %). On the independent network ZTD is not the best configuration for temperature by any measure: with CV3 it is the only configuration that overshoots the bias, to +0.430 K, and with either covariance it is the only one that improves neither the mean absolute nor the root-mean-square error. We have withdrawn that claim.

The correction shifts every experiment by the same +0.10 K (a property of the stations, not the experiment), so it does not reorder configurations — only the magnitude of the reported bias and any percentage computed from it. The ordering is unchanged at 5.5 and 6.0 K km^{-1}.

Configurations separate clearly: conventional observations improve all three metrics with either covariance; the combined configuration improves bias and MAE while leaving RMSE slightly worse than control; ZTD alone improves bias but neither error magnitude — consistent with SYNOP reporting surface temperature directly and ZTD only column-integrated moisture. Tested with the same two resampling schemes as the categorical scores (134 grid cells; 31 calendar days): the change in **bias** is significant under both schemes for every configuration. The change in **error magnitude** mostly is not — only CONV with CV5 shows an MAE reduction surviving both schemes (−0.077 K, day-block interval [−0.148, −0.003]); the CONV-with-CV3 reduction is significant for cells but not days ([−0.189, +0.014]); and **no configuration significantly reduces the RMSE**. The three significant degradations all come from ZTD alone: with CV3 it raises MAE by +0.175 K and RMSE by +0.238 K, and with CV5 it raises RMSE by +0.092 K.

The statement now in the paper: every configuration except ZTD with CV3 moves the bias towards zero; the only significant improvement in error magnitude comes from a configuration containing conventional observations; the only significant degradations come from ZTD alone. No best configuration is claimed for temperature. The intervals are archived in `t2m_bootstrap.txt` with the analysis files.

Surface pressure is used only qualitatively, in the time series of revised Fig. 6. No correction has been applied and the caption now states this explicitly.

RC1-5 — The paper never shows what ZTD assimilation did to the model

> *(1) OMB and OMA statistics … for ZTD, and separately for SYNOP, TEMP and TAMDAR … (2) At least one figure of the analysis increment … for a representative cycle during 14-15 July*

(1) Done. We extracted the WRFDA `statistics` and `qcstat_conv` diagnostics from all 122 analysis cycles. Month-aggregated, observation-count-weighted results:

Obs type	Variable	Obs/cycle	O−B RMSE	O−A RMSE (CV3)	Reduction
GNSS ZTD	ztd (cm)	213	**1.143**	**0.686**	**39.9 %**
SYNOP	T (K)	371	1.628	1.204	26.1 %
SYNOP	q (kg/kg)	356	0.0011	0.0008	30.8 %
SYNOP	p (Pa)	369	76.40	65.71	14.0 %
TEMP (sound)	T (K)	1391	1.041	0.749	28.1 %
TEMP (sound)	q (kg/kg)	774	0.0005	0.0003	31.3 %
TEMP (sound)	u (m/s)	1462	2.715	2.373	12.6 %
TAMDAR	T (K)	915	1.068	0.774	27.6 %
TAMDAR	u (m/s)	901	2.602	2.373	8.8 %

Quality-control counts (mean per cycle, after QC / rejected): GNSS ZTD 213.3 / 2.1 (1.0 % rejected); SYNOP T 370.6 / 0.1; TEMP T 1391.2 / 8.6; TAMDAR T 915.4 / 6.5. Full table and per-cycle series in the revised manuscript.

The ZTD row shows the delays influencing the analysis: a mean background departure of −0.378 cm (background overestimates the delay) is reduced to +0.014 cm with CV3 and +0.033 cm with CV5; RMSE falls from 1.143 cm to 0.686 cm (39.9 %) with CV3 and 0.389 cm (66.0 %) with CV5. O−B is identical across configurations, as expected since every cycle starts from the same GFS background — confirming the experiments differ only in the assimilation step.

Repeating the same experiment with CV5 (RC2-2) reduces the ZTD O−A RMSE to **0.389 cm, a 66 % reduction**, from an identical background and identical observations, so the difference is attributable to **B** alone. A closer fit is not by itself evidence of a better analysis: the consistency check in the revised Sect. 5 shows CV5 fitting the delays to about half their assigned error, which must be read alongside the correlated-error and bias-correction limitations discussed in Sect. 6.

(2) Done. Figure 15 shows the analysis increment (analysis minus its own background) for the cycle valid at 12:00 UTC on 14 July 2021 — column IWV, 700 hPa specific humidity, and lowest-model-level temperature.

A second figure shows the separate CONV, ZTD and CONV + ZTD increments, computed from the archived first guess and analysis of each configuration at the same cycle. Column water vapour increments are 0.52 mm RMS for CONV, 0.45 mm for ZTD alone and 0.68 mm for the combination — ZTD alone produces a moisture adjustment comparable to the entire conventional network. File paths and md5 sums are archived in `increments_cv5_provenance.txt`. We additionally show CONV + ZTD with **CV3 beside CV5**: this pairing is more informative, since the two share a bit-identical background and identical observations, so the difference between columns is produced by **B** alone.

With CV3 the increments are broad and smooth, spanning several hundred kilometres with little relation to where the observations are (column-IWV increment RMS 1.63 mm); with CV5 the same information is deposited in compact, localised structures near the observations (RMS 0.67 mm). A homogeneous isotropic covariance spreads each observation over its prescribed correlation length regardless of the meteorological situation — diluting information and imposing increments where nothing supports them. This is the mechanism behind the 40 % vs 66 % difference in analysis fit.

A further result, relevant to RC1-4 and RC1-6. Examining the increments field by field shows the 3D-VAR analysis leaves several fields *exactly* untouched:

Field	**Max |increment|**
T2, Q2, U10, V10 (2-m / 10-m diagnostics)	**0**
SMOIS, TSLB (soil moisture, soil temperature)	**0**
T, QVAPOR, U, V, PH, MU (3-D atmospheric state)	changed

The analysis does not directly update T2, soil moisture or soil temperature, so its influence on 2-m temperature develops indirectly during the subsequent 6-hour forecast, through humidity, cloud, radiation, boundary-layer mixing and the surface fluxes these produce. This confirms the premise of RC1-6 ("given that 3D-Var with CV3 does not update soil moisture"). It is also a reason to treat the sub-0.4 K differences between experiments in the temperature bias with caution: the increment in the lowest *analysed* model level has an RMS of 0.58 K, comparable to or larger than those differences.

RC1-6 — A genuine Discussion

> *Section 3 is titled "Results and Discussions" but does not compare the present results with the studies cited in the Introduction ... Please add a genuine Discussion*

Done. A new Sect. 6 (Discussion) addresses each of the five points raised:

1. **Comparison with Torcasio et al. (2023), Giannaros et al. (2020), Poletti et al. (2019) and Vladimirov et al. (2020).** Our ZTD impact is smaller, and we identify four reasons: several of them evaluate selected episodes, in which a favourable increment is not diluted by a month of quiescent cycles, whereas Torcasio et al. is the closest month-long comparison; they use case-specific, ensemble-based or, in the case of Torcasio et al., domain-specific NMC-derived background errors, whereas half of our configurations use climatological CV3 (the CV5 experiments quantify exactly what this costs — 40 % versus 66 % reduction in ZTD analysis-fit RMSE); our design re-anchors the background to GFS every cycle so increments do not compound; and our quantitative verification rests on point gauges rather than spatial rainfall fields.

2. **12 km grid spacing with parameterised convection, and the double-penalty problem.** Discussed explicitly: the Grell–Freitas scheme cannot represent the convective cells that produced the extreme local rainfall, so the model produces approximately the right areal extent but not the intensity peaks. When a parameterized cell is displaced by one or two grid lengths, point-gauge verification counts both a miss and a false alarm, so broad weak rain can score better than a well-placed feature. This is exactly the FBI-up/ETS-down signature at the 0.1 mm threshold. It does not, however, mean the signal is clearer at the heavier thresholds: at 5 mm/6 h and above no configuration improves significantly on the control, and at 10 mm every configuration scores below it.

3. **GFS already contains most of the assimilated conventional information.** Discussed: the CONV increment is largely a re-assimilation of information the background already holds, consistent with the modest CONV impact and with CONV mainly increasing rain frequency rather than improving placement. We note that GNSS ZTD is the one observation type here not already in GFS at comparable density — an argument *for* the ZTD component.

4. **The physical pathway from a ZTD moisture increment to a T2m bias reduction.** Discussed: 3D-VAR with CV3 updates the atmospheric state only; soil moisture and soil temperature are not analysed. A moisture increment therefore reaches 2-m temperature only indirectly during the forecast — through column humidity, cloud radiative effects, boundary-layer mixing and the surface fluxes these in turn modify — rather than through any direct adjustment of the land surface, which dominates the diurnal cycle. This is a further reason for treating the small temperature differences with caution, and motivates the soil-moisture recommendation in the outlook.

5. **The limitations of CV3.** Now quantified rather than merely acknowledged, via the CV5 comparison, and collected in a Limitations subsection.

Minor comments

Comment	Response
L1–L12 Abstract reports percentages without units or baselines	Abstract rewritten with absolute values throughout, on the independent network, e.g. the mean precipitation bias moves from −0.48 mm/6 h in the control to −0.09 mm/6 h for ZTD with CV5.

Comment	Response
L6–L7 "largest gains in bias (+37.4 %)" reads as an increase	Rephrased as a reduction. The numbers themselves also changed: recomputed on the rebuilt network the bias moves from −0.48 mm/6 h in the control to +0.51 for CONV with CV3 and +0.18 with CV5, and CONV+ZTD with CV3 gives the smallest residual bias (−0.03 mm/6 h).
L9–L10 "+97.4 %" needs absolute values	This claim has been **withdrawn** following RC1-4; the temperature result is now stated with absolute values and no winner is declared.
L21 missing space in "time(Hapuarachchi et al., 2011)"; missing full stop	Fixed.
L23–L24 "Moreover, The likelihood"; missing full stop after "(Tom et al., 2022)"	Fixed.
L31 missing full stop after "(Ferreira et al., 2022)"	Fixed.
L34 missing space in "et al., 2023).The"	Fixed (and the same pattern in the Fig. 11 caption).
L40–L41 missing space in "infrastructure(Koks et al., 2022)"	Fixed.
L44 "nearly 253 fatalities"	Changed to "253 fatalities".
L48–L49 missing space in "Services(Luxembourg Institute. . . "	Fixed.
L119–L121 Figures 4 and 5 are generic; recommend removing	**Both removed.** Figure 7 (pre-processing) is retained, as the referee allowed.
L121 "Table 4 and Table 3" out of order	Fixed; tables are now cited in the order they appear.
L140 Eq. (1) factor ½; use **R** not **O**; consistent notation block	Fixed: the ½ is now on both terms, **O** → **R** throughout, and all symbols are defined in one notation block. We also now state that **R** is diagonal and give the ZTD observation error (0.8 cm), per RC2.
L205–L208 observation counts given as totals	Replaced with mean number assimilated per cycle for each type, plus numbers passing QC (table + per-cycle time series).
L221–L222 "anomalously large" uncertainty — specify threshold	We are unable to give the threshold. The pre-processing predates this revision and its log records only that epochs with anomalously large formal uncertainty were removed, not the rule applied; we have said exactly that in Sect. 2.5 rather than reconstruct a criterion after the fact. Separately and verifiably, WRFDA's own quality control at the assimilation step rejects 2.1 of 215.4 ZTD reports per cycle (1.0 %), leaving 213.3 — those are counts from a different operation at a different stage, and the manuscript now says so.
L321–L325 omits that CONV+ZTD degrades RMSE	Now stated explicitly in the text, together with the corresponding CV5 degradation.
L356–L361 outlook	Extended with GNSS slant delays and tomography, hourly cycling with analyses propagated between cycles, and soil-moisture initialisation, as suggested.
Tables 1 and 2 never cited, appear in reverse order	Both now cited in the event-description paragraph, in order.
Table 2 return periods confusingly differ from the text	Both captions and the text now state the accumulation period explicitly (daily vs seven-day) and note that the two are not directly comparable.
Figure 2 never cited	Now cited in the event description.

Comment	Response
Figure 3 caption lists figures out of order; should distinguish assimilated from validation stations	Cross-references reordered, and a new station map has been produced: circles mark the stations verifying precipitation (n = 264) and triangles the subset also verifying 2 m temperature (n = 148), over the model's own terrain; per-country counts are given in the caption. Candidate stations that failed the independence screening are not shown.

Summary of changes that alter the paper's conclusions

Answering RC1-2 and RC1-3 together changed the central result:

1. **The verification network was not independent (RC1-3).** 12 of the 15 NOAA ISD stations carry the WMO identifier of an assimilated report. The network has been rebuilt to 264 stations in 234 grid cells across four countries, screened against every assimilated SYNOP position and, for Belgium, every assimilated GNSS site.

2. **The background error covariance, not the observation type, governs the result (RC1-2).** With CV3 — the configuration of the first submission — CONV assimilation is *worse than no assimilation* at all six thresholds; ETS falls from 0.383 to 0.337 at 0.1 mm/6 h, CI [−0.055, −0.038]. With CV5 the same observations improve it: CONV+ZTD +0.031 at 0.1 mm and +0.057 at 1 mm, the latter significant under both the cell and the day-block bootstrap — confirming the prediction made in the first submission's own outlook.

3. **The improvement is bounded in intensity and in space.** Nothing improves at or above 10 mm/6 h. The gain is largest over Belgium (+0.099) and Germany (+0.089), smaller over France (+0.029) and absent over Luxembourg (−0.001).

4. **The "+97.4 % temperature bias" claim is withdrawn** (RC1-4). Verified on an independent network and corrected to station height, ZTD is not the best configuration for temperature bias: with CV3 it is the only configuration that overshoots, to +0.430 K. Only CONV with CV5 shows a reduction in error magnitude that survives both resampling schemes, no configuration significantly reduces the root-mean-square error, and the three significant degradations all come from ZTD assimilated alone, so no best configuration is identified for temperature.

5. **The ZTD experiment definition was corrected.** `use_tamdarobs` defaults to true in WRFDA and was never set in any version of the run script, including those used for the first submission, so the original "ZTD" experiment also assimilated TAMDAR aircraft data, which outnumbered the ZTD observations by roughly 18:1. The revised runs disable it, so ZTD now means ZTD.

6. **The resampling unit was corrected** from station to model grid cell, since stations sharing a 12 km cell are handed the same model value.

What survives from the first submission: the value of the GNSS observations — adding ZTD to CONV gives a robust improvement at 1 mm/6 h (+0.037, significant under both resampling schemes), whereas adding CONV to ZTD does not produce a difference significant under both; the finding that assimilation improves detection more than intensity; and the observation-space result, which is independent of the verification network entirely — the ZTD analysis fit improves from an O−B RMS of 1.14 cm to 0.69 cm with CV3 and 0.39 cm with CV5.