Responses to Referee Comments

Manuscript: `egusphere-2025-4590`

RC2

Title: Comment on egusphere-2025-4590, Anonymous Referee #2
Date: 15 Aug 2026

General statement

We sincerely thank the reviewer for the constructive and insightful comments. These suggestions prompted us to re-examine both the methodological design and the interpretation of our results, and helped us establish a more coherent scope and clearer take-home messages. For clarity, our responses below are organized according to the extent of the required revisions, beginning with the major methodological issues and followed by more specific points.

The revised study is centered on two primary research questions:

1. Whether CSGD parameters calibrated from historical radar–gauge observations can be transferred to radar-based nowcasts to reduce the radar/QPE-related component of forecast error and characterize its associated uncertainty.
2. Whether the calibrated CSGD parameters can be spatially interpolated to ungauged locations, thereby enabling domain-wide application of the CSGD framework.

To clarify the scope of the study, we distinguish two conceptually different components contributing to discrepancies between radar-based nowcasts and gauge observations: (1) limitations of the nowcasting model in representing rainfall-field motion and precipitation evolution, and (2) limitations associated with radar measurement and QPE processing. The present study specifically focuses on the second component and examines whether the radar/QPE-related error structure characterized from historical radar–gauge observations can be transferred to radar-based nowcasts through CSGD post-processing. Under this scope, we agree that additional robustness analyses are required. In particular, temporal cross-validation (CV) is required to assess whether the CSGD calibration remains robust across different time periods, while spatial CV is required to evaluate the reliability of interpolated CSGD parameters at ungauged locations.

We have also reconsidered the role of ensemble nowcasting in the manuscript and propose removing the ensemble-nowcasting branch from the main analysis for the following reasons:

1. The primary objective of this study is to reduce radar/QPE-related forecast error and characterize the associated uncertainty, whereas the stochastic STEPS ensemble is primarily intended to represent predictive uncertainty associated with rainfall-field evolution and motion through stochastic perturbations. Including both mechanisms in the primary analysis therefore makes attribution of the radar/QPE-related component targeted by CSGD less transparent.
2. Furthermore, both STEPS and CSGD provide probabilistic representations, but they characterize different sources of uncertainty. Applying a conditional CSGD distribution to individual members of an already probabilistic STEPS ensemble introduces two distinct layers of uncer-

tainty, whose respective contributions cannot be isolated straightforwardly without a dedicated joint-calibration framework.

3. The original manuscript already contains several complementary analyses, and the additional temporal and spatial robustness assessments requested during review further expand the scope. Removing the ensemble-nowcasting branch therefore allows the revised manuscript to remain focused on the two primary research questions above and provides a clearer assessment of the contribution of CSGD post-processing.

Accordingly, the revised manuscript focuses on deterministic radar nowcasts, for which the incremental effects of CSGD-L and CSGD-NL can be isolated more directly. The probabilistic information provided by the CSGD is nevertheless retained and evaluated using distribution-based verification metrics, such as the CRPS.

We also re-examined the deterministic verification procedure and identified an error in the original RMSE aggregation. Previously, RMSE was calculated separately for each station within each event and lead time, and the resulting station-specific values were then averaged. Although this provides a valid summary of local RMSEs, it is not equivalent to calculating RMSE directly from the pooled forecast–observation errors because of the nonlinear square-root operation. It also gives equal weight to station–event subsets containing different numbers of valid observations. In the revised analysis, RMSE is calculated separately for each event and lead time using all valid forecast–observation pairs across gauge locations and forecast times within that event. The resulting event-level RMSEs are then summarized across the validation events. We have recomputed the corresponding results using this revised procedure.

We further reconsidered the choice of a representative point estimate from the conditional CSGD. Wright et al. (2017) used the conditional median for deterministic evaluation, while noting that the conditional mean is always nonzero and greater than the median, whereas the median can become zero when the conditional probability of precipitation is sufficiently low. They therefore acknowledged that neither quantity is universally optimal as a measure of central tendency. The median is particularly useful when precipitation occurrence and the preservation of dry conditions are central to the application, because it can remain exactly zero when the predictive distribution assigns substantial probability to no rainfall.

The objective of the present verification is somewhat different. Our analysis is conditioned on preselected rainfall events and is primarily concerned with the magnitude of rainfall expected during these events, rather than with reproducing dry-period occurrence over an unrestricted time series. Under this setting, using the median may place disproportionate emphasis on the probability mass at zero and can yield a zero or relatively low representative value even when the conditional distribution retains a substantial positive-rainfall component. We therefore use the conditional mean, i.e. the expected value of the full CSGD predictive distribution, as the deterministic representative value. The revised experiments further show that the mean-based adjustment produces lower RMSE than the median-based adjustment (within the first 3 h), providing empirical support for this choice.

General remarks

1. Temporal cross-validation, MFB benchmarking, and gauge-overlap effects

Reviewer comment

1-a. Methodologically, there are several aspects which should be improved, first of all the validation strategy. The temporal validation approach seems correct, with a clearly separate calibration and verification period. However, repeating the validation over different subset (k-fold cross-validation, taking into account temporal correlations) is necessary to get information about the variance/significance of the obtained results.

1-b. Another, minor, methodological concern that I have, is that the authors might be applying a double correction: the Nimrod radar dataset already makes use of real-time rain gauge measurements to correct the QPE. Are these also in MIDAS? This should be clarified and the possible implications should be discussed. Moreover, it would be good to compare to a simple baseline (Mean field bias correction with MIDAS) to show that the proposed complex model really has added value.

Response

We thank the reviewer for these suggestions. To assess the temporal robustness of the CSGD calibration while accounting for temporal dependence, we adopted a year-blocked leave-one-year-out cross-validation strategy, following Valdez et al. (2022). Each year from 2016 to 2020 was withheld in turn, while the CSGD models were recalibrated using the remaining four years and evaluated against MIDAS observations in the withheld year. Using complete years as validation blocks avoids randomly splitting temporally correlated hourly observations between calibration and validation.

As shown in Figs. 1–2 and Table 1, CSGD-NL reduced RMSE in all five held-out years by 3.02–5.89%, with a mean reduction of 4.33 ± 1.28%, indicating robust performance across different calibration periods. In contrast, CSGD-L produced a mean RMSE change of −0.88 ± 0.79%. Following the reviewer's suggestion, we also included mean-field bias (MFB) correction as a simple benchmark. For each fold, a spatially uniform multiplicative bias factor was estimated from paired MIDAS–NIMROD observations in the four calibration years and applied to the withheld year. MFB reduced RMSE by only 0.31 ± 0.10% on average, whereas CSGD-NL outperformed MFB in all five folds, providing an additional mean RMSE reduction of approximately 4%. This indicates that the improvement from CSGD-NL cannot be explained solely by correction of a spatially uniform mean bias.

Regarding the potential double correction, operational NIMROD already incorporates rain-gauge information through a spatially uniform multiplicative gauge-adjustment factor (Harrison et al., 2000), which is conceptually similar to MFB correction. Because MIDAS Open includes Met Office-operated gauges, some overlap with gauges used in operational NIMROD processing is possible, although the exact station-level overlap during 2016–2020 cannot be determined from publicly available documentation. The small additional improvement from MFB is therefore consistent with much of the mean bias having already been corrected in NIMROD. Nevertheless, despite this potential overlap, CSGD-NL consistently reduced RMSE by 4.33 ± 1.28% across the five held-out years, indicating that substantial residual radar–gauge error structure remains after the operational gauge adjustment. This supports the potential of CSGD-NL to improve NIMROD-based nowcasts by correcting residual conditional error characteristics beyond a simple mean-field adjustment.

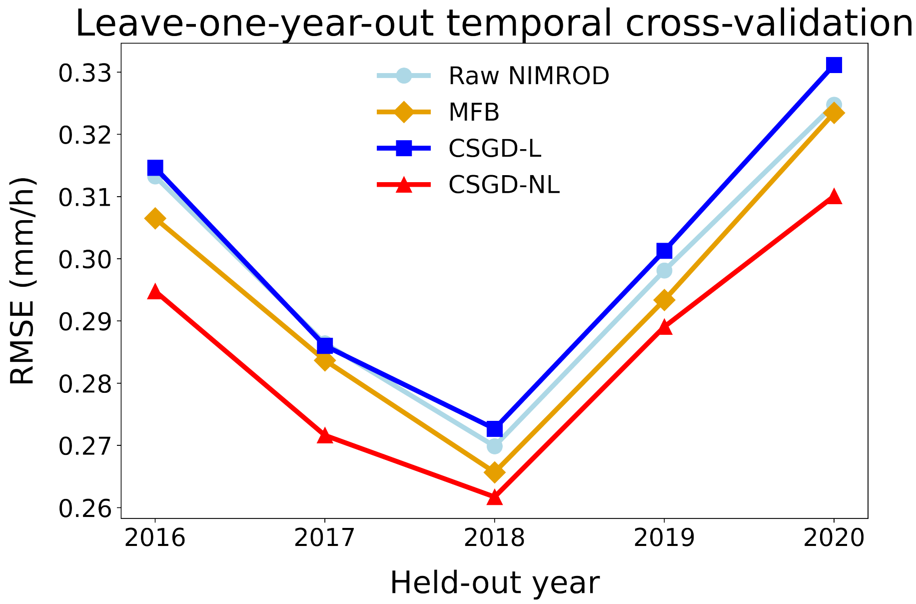

Figure 1: Leave-one-year-out temporal cross-validation of raw NIMROD, MFB, CSGD-L, and CSGD-NL. Each year from 2016 to 2020 was withheld in turn, with the remaining four years used for calibration.

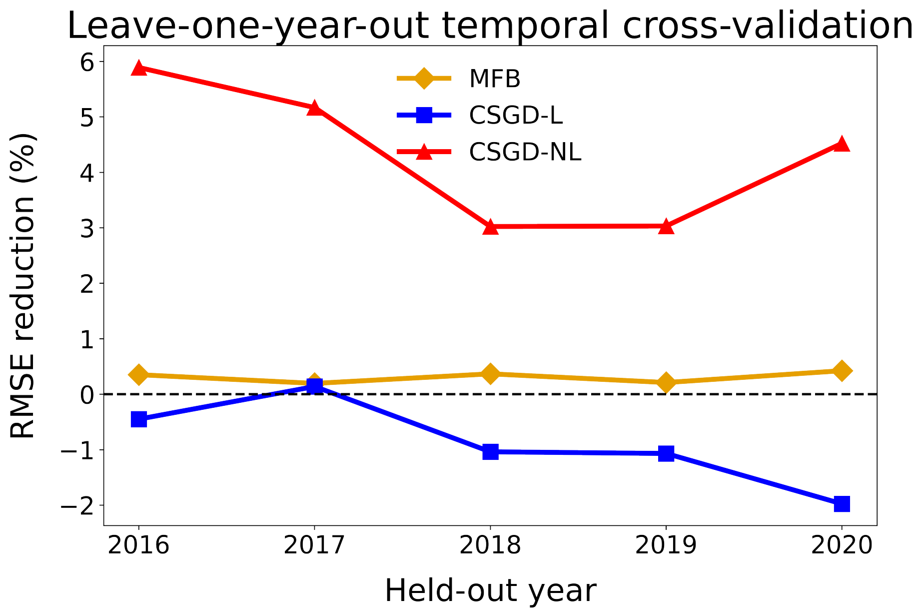

Figure 2: RMSE reduction relative to raw NIMROD for each held-out year. Positive values indicate improvement.

Table 1: Leave-one-year-out temporal cross-validation results. RMSE is in $\mathrm{mm\,h^{-1}}$; positive reductions indicate improvement relative to raw NIMROD.

Held-out year	Raw RMSE	MFB RMSE	CSGD-L RMSE	CSGD-NL RMSE	MFB reduction (%)	CSGD-L reduction (%)	CSGD-NL reduction (%)
2016	0.3132	0.3121	0.3146	0.2948	0.35	-0.45	5.89
2017	0.2864	0.2859	0.2860	0.2716	0.20	0.14	5.17
2018	0.2699	0.2689	0.2727	0.2617	0.37	-1.04	3.02
2019	0.2981	0.2975	0.3013	0.2891	0.21	-1.07	3.03
2020	0.3248	0.3234	0.3312	0.3101	0.42	-1.97	4.52
Mean ± SD	0.2985 ± 0.0216	0.2975 ± 0.0214	0.3012 ± 0.0230	0.2855 ± 0.0191	0.31 ± 0.10	−0.88 ± 0.79	4.33 ± 1.28

2. Spatial cross-validation of CSGD parameter interpolation

Reviewer comment

As for the spatial interpolation verification, the set of rain gauges that was used to estimate the errors in the spatial interpolation of the CSGD parameters was not picked randomly, but chosen to be centrally located, so the results might be too optimistic. Like for the temporal validation, the authors should apply k-fold cross-validation to get at least some information about the significance of these results.

Response

We thank the reviewer for this suggestion. We agree that the original validation based on 15 centrally located gauges could provide an overly favourable estimate of interpolation performance. We therefore replaced it with a 10-fold random spatial cross-validation using all 156 MIDAS gauges. In each outer fold, 15–16 gauges were withheld and the remaining 140–141 gauges were used for spatial interpolation, such that every gauge served as an outer holdout exactly once.

To avoid using the outer validation gauges for interpolation-method selection, we additionally performed a 5-fold inner cross-validation within the outer-training gauges. The candidate ordinary Kriging and RBF methods considered in the original analysis were evaluated using inner-CV MAE, and the best-performing method for each CSGD parameter was then fitted to all outer-training gauges and evaluated at the outer-held-out gauges. The directly calibrated parameter values at the held-out gauges were used only as validation references and did not enter either method selection or interpolation.

As summarized in Table 2, the revised 10-fold cross-validation provides both the mean interpolation error and its variability across spatial subsets, substantially broadening the assessment compared with the original centrally selected validation set. The resulting MAEs were generally comparable to those obtained from the original validation, indicating that the interpolation performance was not dependent on the centrally located gauge subset. Method selection was also stable across folds: seven of the ten CSGD parameters selected the same interpolation method in all ten outer folds, while the dominant method was selected in at least eight folds for the remaining parameters.

Importantly, the spatially withheld interpolated parameter values generated in the outer cross-validation are subsequently used in the end-to-end nowcast validation at the corresponding MIDAS gauges, rather than their directly calibrated station-specific parameters. Thus, each gauge is spatially excluded from the parameter interpolation used for its subsequent validation.

Table 2: Comparison between the original interpolation validation using 15 centrally selected gauges and the revised 10-fold spatial cross-validation. Cross-validation values are reported as the mean ± standard deviation of MAE across the ten outer folds.

CSGD model	Parameter	Original MAE	10-fold CV MAE
Climatological	μ_c	0.03	0.0275 ± 0.0101
	σ_c	0.05	0.0563 ± 0.0162
	δ_c	0.02	0.0221 ± 0.0068
Conditional-L	α_2	0.06	0.0608 ± 0.0145
	α_3	0.04	0.0447 ± 0.0110
	α_4	0.06	0.0595 ± 0.0135
Conditional-NL	α_1	0.02	0.0152 ± 0.0054
	α_2	0.04	0.0429 ± 0.0090
	α_3	0.08	0.0648 ± 0.0156
	α_4	0.06	0.0610 ± 0.0134

3. Spatio-temporal validation at ungauged locations

Reviewer comment

A major methodological flaw is that, in the nowcast validation, all gauges are used for the calibration. Therefore, the validation does not represent the error for any random pixel, only for gauged pixels. A proper validation of the whole process, that represents skill at any location (including where there's no gauges), should therefore use data that was both temporally and spatially withheld from the calibration procedure. This must be addressed in the revised manuscript.

Response

We thank the reviewer for identifying this important methodological issue. We revised the validation framework to ensure both temporal and spatial independence. CSGD parameters are calibrated using radar–gauge observations from 2016–2020, whereas nowcast verification uses independent rainfall events from 2021–2022. Spatial independence is achieved through the 10-fold spatial cross-validation described above: for each outer fold, CSGD parameters at the held-out gauges are obtained exclusively by interpolation from the remaining gauges. Thus, each validation gauge is treated as an ungauged location for CSGD parameter estimation.

Under this spatially withheld setting, we evaluated deterministic performance using both the conditional mean and median (Figures 3 and 4). CSGD-NL consistently outperformed CSGD-L, and the mean-based CSGD-NL achieved larger RMSE reductions within the first 3 h, the most relevant lead-time range for nowcasting (Figure 5). Together with its interpretation as the expected rainfall amount during the selected rainfall events, this supports using the conditional mean as the deterministic representative value in the revised analysis.

We further evaluated the full predictive distribution using CRPS. As shown in Figure 6, CSGD-NL reduced CRPS by more than 20% relative to the original nowcast across all lead times, with consistent improvements for both summer and winter events. These results demonstrate that the benefit of CSGD-NL persists when validation locations are spatially withheld from calibration.

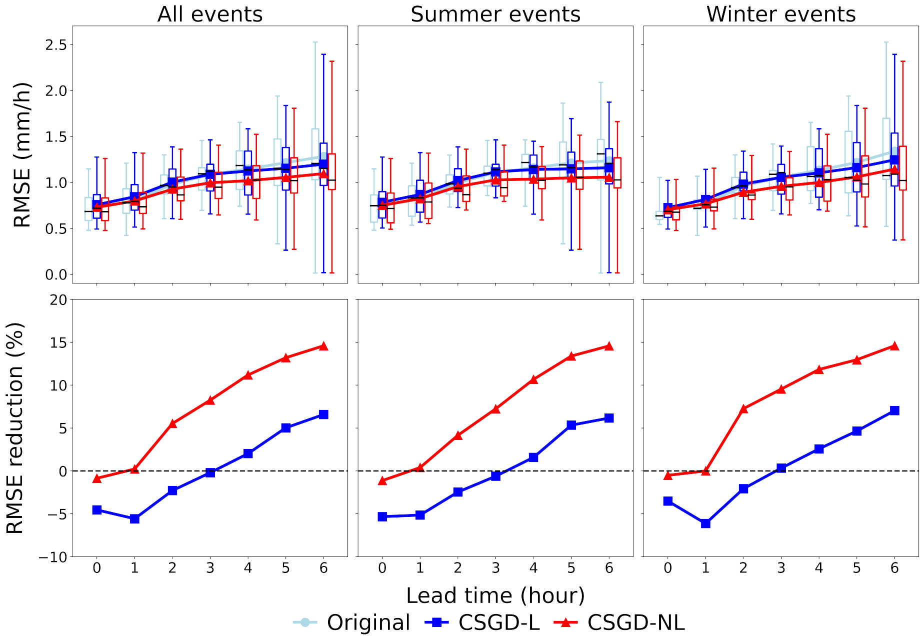

Figure 3: Spatially withheld deterministic verification using the conditional CSGD mean. The upper panels show event-level RMSE for the original nowcast, CSGD-L, and CSGD-NL, while the lower panels show RMSE reduction relative to the original nowcast for all, summer, and winter events.

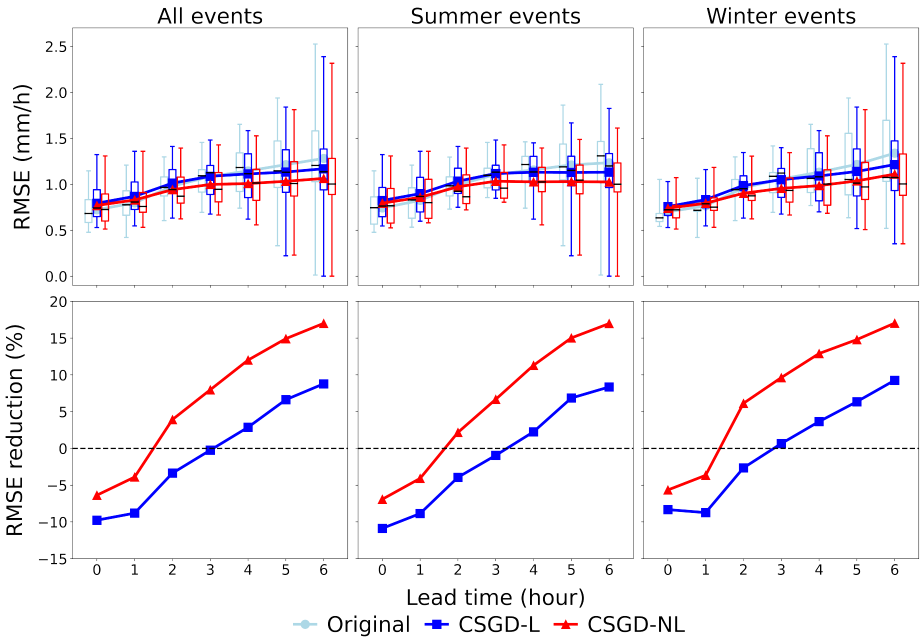

Figure 4: Spatially withheld deterministic verification using the conditional CSGD median. The upper panels show event-level RMSE, while the lower panels show RMSE reduction relative to the original nowcast for all, summer, and winter events.

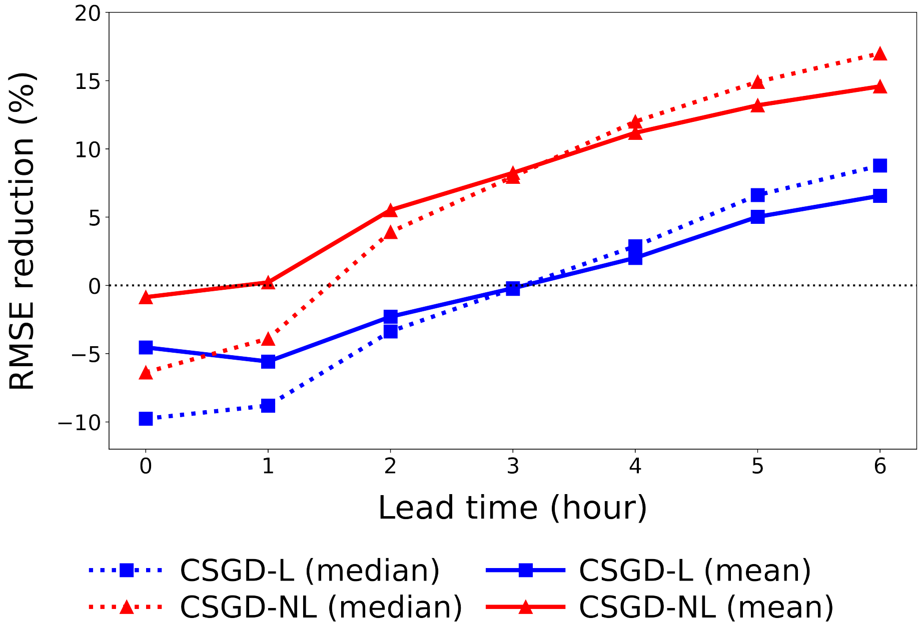

Figure 5: Comparison of RMSE reduction obtained using the conditional mean and median for CSGD-L and CSGD-NL under spatially withheld validation. Positive values indicate improvement relative to the original nowcast.

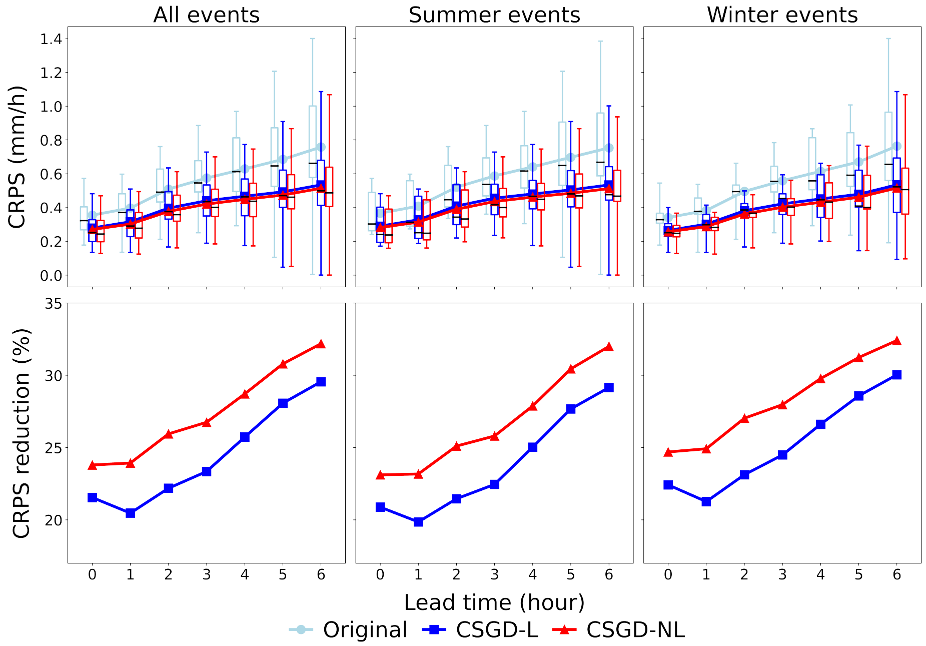

Figure 6: Spatially withheld probabilistic verification using CRPS. The upper panels show event-level CRPS for the original nowcast, CSGD-L, and CSGD-NL, while the lower panels show CRPS reduction relative to the original nowcast for all, summer, and winter events.

4. Lead-time dependence of CSGD performance

Reviewer comment

4-a. However, the method seems to perform adequately at (very) long lead times, but performs worse at short (i.e. skillful) lead times, making it hard to justify why one would use this method for most nowcasting applications.

4-b. As for improving the nowcasting, the manuscript makes some claims which are very selective and not well-supported. The authors report skill improvement for nowcasts at 6-hour lead times evaluated on MIDAS. My first concern is that these are very long lead times for purely extrapolation-based nowcasting, which typically loses most skill after 1-2 hours unless combined with NWP forecast data. Indeed correlations with gauges are very low for 6h lead times, cfr table 3. How does the method do for shorter lead times, where nowcasts are supposed to have more skill? It seems to degrade the nowcasts there, which a serious disadvantage. For the average user, this would make their model less useful than the original nowcast. For example, in Table 4 it is clear that after 1h, the CSGD methods will underestimates rainfall wrt the original, which is a major issue.

Response

We thank the reviewer for raising this important concern regarding the short leadtime applicability of the proposed approach. To further assess performance at short lead times, we complemented the hourly MIDAS verification with the 5-min EA-ST dataset. For MIDAS, the RMSE reduction of CSGD-NL increased from 0.22% at 1 h to 5.52% at 2 h and 8.24% at 3 h, as shown in Figure 3. The corresponding CRPS reduction was already 23.93% at 1 h and increased from 23.80% at 0 h to 32.19% at 6 h (see Figure 6).

The higher-resolution EA-ST verification provides a more direct assessment of short-term performance. CSGD-NL produced RMSE reductions at all 5-min lead times, increasing from 4.25% at 0 min to 7.39% at 30 min and remaining approximately 6–10% from 30 to 180 min (Figure 7). More importantly, CRPS was reduced consistently at every 5-min lead time, ranging from 24.68% to 27.70% over 0–180 min (Figure 8). These results show that the benefit of CSGD-NL is not confined to long lead times.

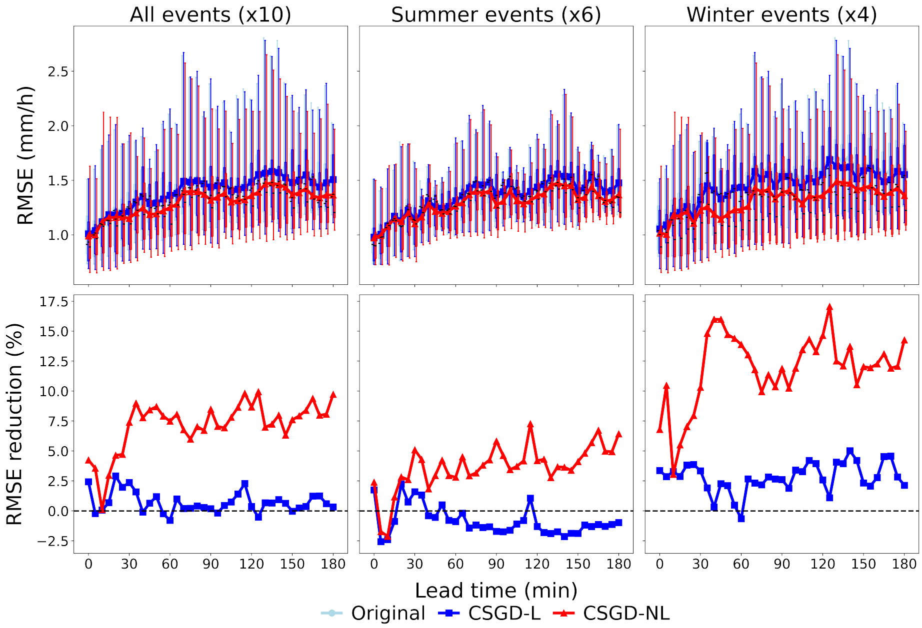

Figure 7: Event-level RMSE verification using the 5-min EA-ST observations and the conditional CSGD mean. The upper panels show RMSE for the original nowcast, CSGD-L, and CSGD-NL, while the lower panels show the corresponding RMSE reductions relative to the original nowcast. Positive values indicate improvement.

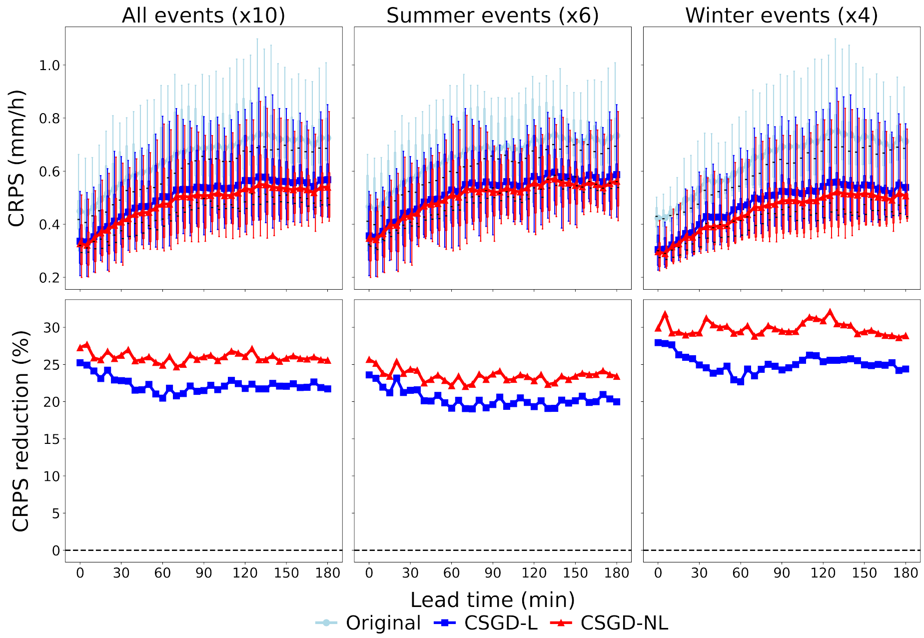

Figure 8: Event-level CRPS verification using the 5-min EA-ST observations. The upper panels show CRPS for the original nowcast, CSGD-L, and CSGD-NL, while the lower panels show the corresponding CRPS reductions relative to the original nowcast. Positive values indicate improvement.

5. Probabilistic, ensemble, and threshold-based verification

Reviewer comment

Another issue that I see, is that of the whole obtained CDF, only the median is used. From the description of the method, I expected that the whole distribution would be used, but only deterministic scores for the median are reported. Why is this distribution not used? Moreover, why are there no probabilistic scores CRPS, rank histograms, reliability diagrams etc. reported? And what about looking at the performance of the method (false positives/negatives, CSI etc) for different thesholds? These are standard evaluation methods for probabilistic nowcasts, and are readily available in python packages. I also wonder how the RMSE of an ensemble computed? Does it compare to the ensemble mean or add the errors of each member separately (not correct)?

Response

We thank the reviewer for these important comments on our verification. We agree that the RMSE calculation used for the ensemble results in the original manuscript was inappropriate, because errors were evaluated across individual ensemble members rather than against a representative ensemble forecast. Reconsideration of this issue also led us to remove the ensemble-nowcasting analysis from the revised manuscript, for the reasons outlined in the general statement.

We also agree that the full CSGD distribution should be evaluated rather than relying only on a deterministic point estimate. Accordingly, CRPS was added as the primary probabilistic verification metric. For MIDAS, CSGD-NL reduced CRPS by approximately 24–32% over 0–6 h (Figure 6). For the 5-min EA-ST data, the corresponding reduction remained approximately 25–28% over 0–180 min (Figure 8). These results demonstrate consistent improvement of the full predictive distribution at both hourly and 5-min resolutions.

Following the reviewer's suggestion, we further evaluated threshold-dependent performance using CSI and FAR (Figures 9–12). At the 0.1 and 1 mm h^{-1} thresholds, CSGD adjustment produced only small changes in CSI and FAR relative to the original nowcast for both MIDAS and EA-ST. Larger differences occurred at the 5 mm h^{-1} threshold, particularly at short lead times: CSGD-NL generally reduced FAR, but this was accompanied by a reduction in CSI, indicating a more conservative prediction of high-intensity rainfall. The differences progressively diminished with increasing lead time. Thus, CSGD-NL substantially improves the overall probabilistic score while largely preserving categorical performance at low-to-moderate thresholds.

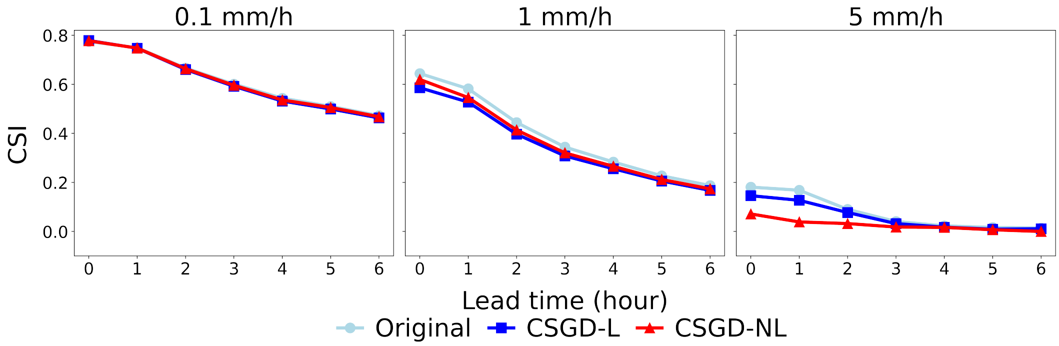

Figure 9: CSI for the original nowcast, CSGD-L, and CSGD-NL using hourly MIDAS observations across the evaluated rainfall thresholds and lead times.

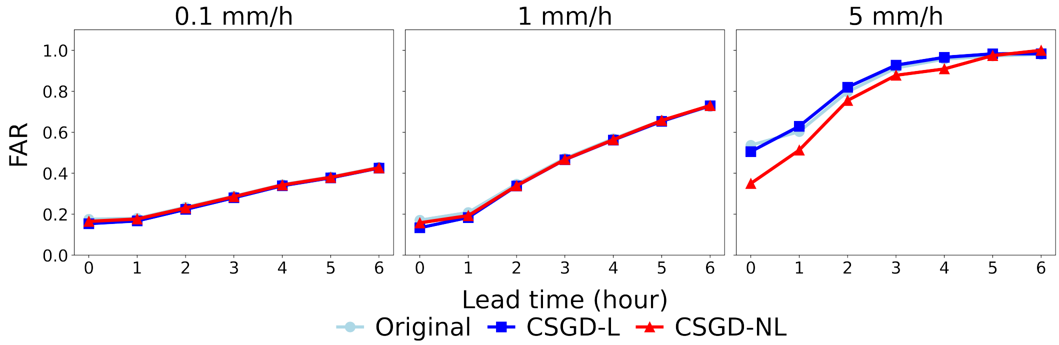

Figure 10: FAR for the original nowcast, CSGD-L, and CSGD-NL using hourly MIDAS observations across the evaluated rainfall thresholds and lead times.

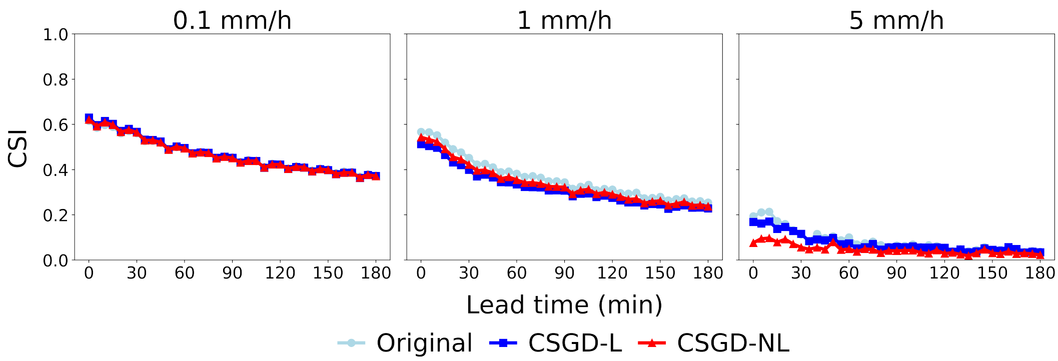

Figure 11: CSI for the original nowcast, CSGD-L, and CSGD-NL using 5-min EA-ST observations across the evaluated rainfall thresholds and lead times.

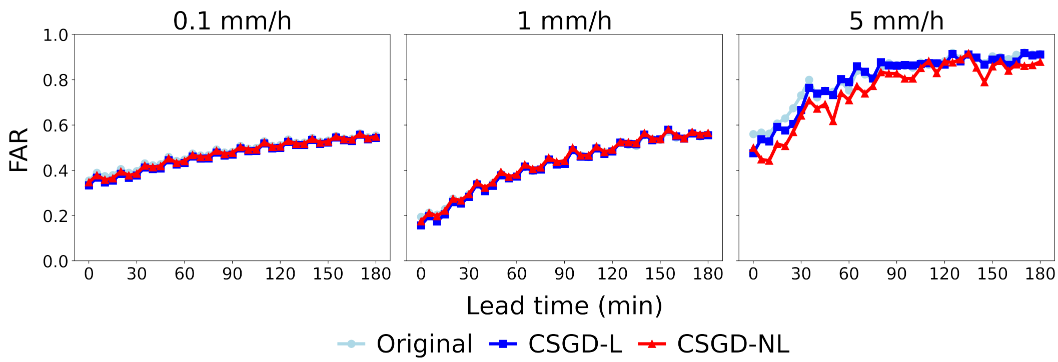

Figure 12: FAR for the original nowcast, CSGD-L, and CSGD-NL using 5-min EA-ST observations across the evaluated rainfall thresholds and lead times.

6. Precision of Reported Performance Improvements

Reviewer comment

As mentioned before, the significance/variability of the reported reductions should be discussed more in depth since this is an evaluation on a limited number of events (20 events in the validation set). The vocabulary "approaching x% or y%" is also a bit strange, just state the percentage (ideally with an error bar) instead of rounding up.

Response

We thank the reviewer for this important comment. To quantify the variability and statistical significance of the reported improvements, we supplemented the revised verification with event-based variability and significance analyses. For each lead time, 95% confidence intervals were estimated using paired bootstrap resampling, and differences between the original and CSGD-adjusted nowcasts were additionally assessed using paired Wilcoxon signed-rank tests with Holm correction for multiple comparisons.

As shown in Table 3, for the 20 MIDAS events, CSGD-NL produced statistically significant RMSE reductions from 2 h onward. The RMSE reduction increased from 5.52% at 2 h (95% CI: 2.82–8.39%; Holm-adjusted $p = 0.0021$) to 8.24% at 3 h (6.26–10.53%; $p < 0.001$) and 14.58% at 6 h

(11.94–17.20%; $p < 0.001$). CRPS was reduced significantly at every evaluated lead time, from 23.80% at 0 h (21.42–26.60%) to 32.19% at 6 h (29.53–34.97%), with improvements in all 20 events and Holm-adjusted $p < 0.001$ throughout.

The 5-min EA-ST verification showed a consistent pattern at the sub-hourly scale. CSGD-NL reduced RMSE by 7.39% at 30 min (95% CI: 1.31–15.79%), 7.49% at 60 min (2.98–14.22%), and 9.72% at 180 min (5.33–14.92%). CRPS reductions were consistently larger, ranging from 24.68% to 27.70% over 0–180 min, with all 10 EA-ST events showing improvement at every lead time and all bootstrap confidence intervals remaining above zero. Because the EA-ST analysis contains only 10 events and 37 closely spaced lead times, the conservative Holm correction resulted in adjusted p-values of 0.0723 for the strongest comparisons; we therefore interpret the EA-ST results primarily through their effect sizes, confidence intervals, and consistency across events rather than claiming formal significance after multiple-testing correction.

We would accordingly replace approximate expressions such as "approaching x%" with the corresponding percentage reductions and 95% confidence intervals. The revised figures additionally show event-to-event variability using boxplots.

Table 3: Event-level CSGD-NL performance for the MIDAS validation. Values in parentheses denote 95% paired-bootstrap confidence intervals. p_{Holm} denotes the Holm-adjusted paired Wilcoxon p-value.

Lead time	RMSE reduction (%)	p_{Holm}	CRPS reduction (%)	p_{Holm}
0 h	−0.86 (−5.00, 4.20)	0.286	23.80 (21.42, 26.60)	< 0.001
1 h	0.22 (−4.35, 5.69)	0.312	23.93 (21.65, 26.54)	< 0.001
2 h	5.52 (2.82, 8.39)	0.002	25.95 (24.01, 28.17)	< 0.001
3 h	8.24 (6.26, 10.53)	< 0.001	26.75 (25.01, 28.60)	< 0.001
4 h	11.18 (8.61, 13.98)	< 0.001	28.71 (26.86, 30.68)	< 0.001
5 h	13.19 (10.59, 15.86)	< 0.001	30.79 (28.63, 33.11)	< 0.001
6 h	14.58 (11.94, 17.20)	< 0.001	32.19 (29.53, 34.97)	< 0.001

7. Clarification of error, uncertainty, and the bias–variance trade-off

Reviewer comment

7-a. The clarity of the writing could be improved; for example, the introduction jumps back and forth between subjects and the goal of the manuscript is not immediately clear. The notions of error and uncertainty are used interchangably (see also my specific remarks below), for example Fig. 10 shows error variances (in $(mm/h)^2$) but then this is referred to as "uncertainty reduction of 20%". Some language is also imprecise, e.g. "Generally better" should be specified (in how many percent of cases, and is it significantly better?)

7-b. The variance reduction might be due to bias increase, which is not ideal either.

Response

We thank the reviewer for this clarification. We have revised the manuscript to distinguish consistently between forecast error and predictive uncertainty. The variance previously shown in Fig. 10 is now referred to explicitly as "error variance", and its reduction is interpreted together with the other deterministic verification metrics to account for possible changes in systematic bias.

Predictive uncertainty is represented by the full conditional CSGD distribution, and the quality of this probabilistic representation is assessed using CRPS. We have also reorganized the Introduction and replaced qualitative expressions such as “generally better” with explicit quantitative statements and, where appropriate, confidence intervals and significance tests.

8. Scope of the Claimed Sub-Hourly Applicability

Reviewer comment

The authors mention that they apply to sub-hourly resolutions, but the calibration only seems to be done with hourly data (5min data are also aggregated to 1h) so this claim should be weakened.

Response

We thank the reviewer for this clarification. We agree that the original wording overstated the temporal resolution of the CSGD calibration. The CSGD parameters are calibrated using hourly radar–gauge data and are subsequently applied to 5-min radar nowcasts. In the revised manuscript, we therefore distinguish explicitly between the hourly calibration resolution and the 5-min application and verification resolution, and have weakened the corresponding claim accordingly. The additional EA-ST analysis nevertheless demonstrates that the hourly-calibrated CSGD can provide measurable improvements when applied to 5-min nowcasts.

Specific remarks

1. Abstract, L9-10

Reviewer comment

"In this process, predicted rainfall intensities are transformed into cumulative distribution functions (CDFs), enabling probabilistic nowcasting." This sentence seems to suggest that the CSGD is needed to provide probabilistic forecasts, while in fact this can be also derived from STEPS ensemble nowcasts alone.

Response

We thank the reviewer for this clarification. We agree that the original wording was misleading and have revised it to state that CSGD provides a conditional predictive distribution for deterministic nowcasts.

2. p2 L27

Reviewer comment

what are "the inherent errors"?

Response

We thank the reviewer for pointing this out. The term “inherent errors” was unnecessarily vague and has been replaced with “forecast errors” in the revised manuscript.

3. p2 L31-33

Reviewer comment

The authors write about the concepts of uncertainty and error: "These two concepts primarily originate from two sources: the difficulty in accounting for uncertainty in nowcasting models and the differences between radar rainfall (RR) estimates and gauge rainfall (GR) measurements (Dai et al., 2015, 2013)." This explanation seems circular, please fix this sentence. What is perhaps meant is that error comes from both extrapolation errors and measurement errors? In the following paragraphs, the words uncertainty and error still seem to be used interchangeably.

Response

We thank the reviewer for pointing this out. We agree that the original explanation was circular. We have revised the sentence to read:

“Both forecast errors and the associated uncertainty can arise from two broad sources: limitations in representing precipitation advection and evolution in the nowcasting model, and limitations in radar measurement and QPE processing relative to gauge observations.”

We have also revised the following paragraphs to distinguish consistently between realized errors and the associated predictive uncertainty.

4. p2: In L41-47

Reviewer comment

the authors repeat themselves somewhat about DL/AI based nowcasting (already discussed in L22-26). Discuss whether this method could also be applied to DL/AI based nowcasting.

Response

We thank the reviewer for this suggestion. We have removed the repetitive discussion of DL/AI-based nowcasting and added the following discussion on its potential applicability:

“Because the CSGD adjustment operates on predicted rainfall intensity at each grid cell, it is not restricted to STEPS. Radar-based DL/AI nowcasting models generally produce gridded future radar or precipitation fields. When these outputs are expressed or converted to rainfall rates ($\mathrm{mm\,h^{-1}}$), they are compatible with the conditional CSGD framework used here.”

5. p2 L54

Reviewer comment

and further: please quote the seminal works on dual pol radar e.g. Bringi, Chandrasekar, Rhyzhkov, Zrnić

Response

We thank the reviewer for this suggestion. We have added the seminal works of Bringi and Chandrasekar (2001) and Ryzhkov and Zrnić (1995, 1996) to the discussion of dual-polarization radar.

6. p4 L95

Reviewer comment

explain NLDAS-2. is it really radar reference data?

Response

We thank the reviewer for pointing this out. NLDAS-2 was used as the reference precipitation dataset in Wright et al. (2017), but it should not have been described as radar reference data. Its precipitation forcing is primarily based on gauge observations, with radar information used for temporal disaggregation where available. We have revised the corresponding text accordingly.

7. p4 L102

Reviewer comment

the methodology can be a described more accurately, e.g. "treating RR data as the satellite dataset" I see what you mean, but this phrasing is a bit confusing.

Response

We thank the reviewer for pointing this out. We agree that the original phrasing was confusing. What we intended was that RR replaces the satellite precipitation estimates as the conditioning predictor in the original CSGD framework, while GR observations serve as the reference. We have therefore revised the sentence to:

“In this study, the CSGD framework is adapted by using radar rainfall estimates as the conditioning predictor, analogous to the role of satellite precipitation estimates in the original framework, while rain-gauge observations are used as the reference rainfall.”

8. p4 L105

Reviewer comment

"Restricted to grids containing...." - do you mean grid points?

Response

We thank the reviewer for pointing this out. We intended to refer to locations where rain-gauge observations are available. We have revised the sentence to:

"Given the point-based nature of GR observations, CSGD parameters can be directly calibrated only at rain-gauge locations."

9. P6 L150-152

Reviewer comment

can you explain the low fraction of stations with sufficiently high correlations with the Midas dataset? Are the other stations wrong or does it simply reflect the high spatial variability of rain?

Response

We thank the reviewer for this important clarification. The stations with correlations below 0.7 should not necessarily be interpreted as erroneous. Because EA-ST stations were compared with the nearest MIDAS stations rather than co-located gauges, lower correlations may reflect spatial separation between the two networks and the high spatial variability of rainfall, in addition to possible station-specific data inconsistencies. We have therefore revised the description from "quality control" to "cross-network consistency screening" and clarified that the threshold was used conservatively to retain stations showing strong agreement with the independent MIDAS observations.

10. P16 L379

Reviewer comment

please specify what the RMSE is calculated over in Eq. (13).

Response

We thank the reviewer for pointing this out. We have clarified that, at each lead time, RMSE is calculated over all valid nowcast–gauge observation pairs across the validation events, gauge locations, and forecast times.

11.

Reviewer comment

Some residual errors are clearly not spatially smooth (e.g. due to the bright band, as visible in Fig. 2b, or geographically fixed sources of non-meteorological echoes) and the parameters may not interpolate trivially to neighbouring pixels. Please discuss

Response

We thank the reviewer for this important point. We agree that instantaneous radar residual errors may contain spatially localized and non-smooth structures, such as bright-band effects or non-meteorological echoes. However, the quantities interpolated in this study are CSGD parameters calibrated from multi-year radar–gauge records rather than instantaneous residual-error fields. Consequently, the influence of transient, event-specific spatial irregularities is expected to be reduced in the calibrated parameters.

To accommodate differences in the spatial structure of individual CSGD parameters, we compared RBF and Kriging interpolation, including multiple variogram models for Kriging, and selected the best-performing approach for each parameter. The revised spatial cross-validation further evaluates the ability of these interpolated parameters to predict withheld gauge locations. Nevertheless, persistent localized radar artefacts may still not be fully represented by spatial interpolation.

12. P29 L509-510

Reviewer comment

please compute scores such as CSI

Response

We thank the reviewer for this suggestion. As described in our response to General Remark 5, we have added CSI and FAR at rainfall thresholds of 0.1, 1, and $5\,\mathrm{mm\,h^{-1}}$ for both the MIDAS and EA-ST verification datasets.