Responses to Referee Comments

Manuscript: egusphere-2025-4590

RC1

Title: Comment on egusphere-2025-4590, Anonymous Referee #1
Date: 27 Jan 2026

The authors developed a framework to combine a radar-based rainfall error model with rainfall nowcasting. The authors demonstrate that applying the non-linear CSGD model to rainfall ensemble nowcasting generally yields the best performance. This research enhances the reliability of radar-based rainfall nowcasting. I would recommend a major revision given the comments below:

Question 1

Reviewer comment

CSGD has been applied to sub-daily or hourly time scale rainfall error modeling, such as:

- Peng, K., D.B. Wright, Y. Derin, S.H. Hartke, Z. Li, J. Tan, *A Novel Near-Realtime Quasi-global Satellite-Only Ensemble Precipitation Dataset*, Water Resources Research, 2025.
- Li, Z., D.B. Wright, S.H. Hartke, D.B. Kirschbaum, S. Khan, V. Maggioni, Pierre-Emmanuel Kirstetter, *Toward A Globally-Applicable Uncertainty Quantification Framework for Satellite Multisensor Precipitation Products based on GPM DPR*, IEEE Transactions on Geoscience Remote Sensing, 2023.

I would not recognize the "potential of extending the CSGD method–originally developed for daily satellite precipitation estimation–to hourly and sub-hourly timescales" as a major finding in this research.

Response

We appreciate the reviewer's comment and agree that CSGD-type error modeling has already been explored beyond daily scales in several recent studies (although the applications to sub-daily scales are rather limited). We therefore agree that we should not overstate the contribution of this study to temporal extension of CSGD from daily to hourly/sub-hourly scales. Instead, the key novelty lies in the proposed radar-nowcasting postprocessing framework: (i) calibrating gauge-referenced climatological and conditional CSGD models at station locations, (ii) interpolating the calibrated parameter sets to obtain pixel-wise conditional CSGD models over the radar domain, and (iii) embedding this probabilistic correction within an operational STEPS nowcasting workflow for both deterministic and ensemble nowcasts. Under this framework, radar-based nowcasts can be adjusted in real time at the native 5-min resolution and evaluated consistently at both hourly (MIDAS) and 5-min (EA-ST) scales.

In other words, the hourly/sub-hourly application should be viewed as an implementation setting required by radar nowcasting, rather than the principal scientific contribution. The main contribution is the integration of gauge-referenced CSGD calibration and spatial parameter interpolation

into a radar-nowcasting context, together with a systematic evaluation of corrected deterministic and ensemble nowcasts. We will revise the Abstract, Introduction, and Conclusions accordingly to avoid overstating the temporal-scale extension as a standalone major finding, and instead emphasize the nowcasting-context integration, pixel-wise probabilistic adjustment, and the demonstrated improvements in corrected radar nowcasts.

Question 2

Reviewer comment

Scheuerer et al. (2015) developed CSGD to be used in precipitation forecasting. Could the authors clarify the rationale for applying an error model calibrated using historical radar observation time series to rainfall nowcasts, rather than calibrating the CSGD model directly on the nowcasted rainfall fields? The latter approach appears more direct for addressing precipitation errors that originate from both radar measurements and the nowcasting model itself. Under the current framework, the capacity to mitigate errors specifically associated with rainfall nowcasting seems limited.

Response

Thank you for the insightful question. We agree that Scheuerer and Hamill (2015) introduced CSGD in a forecasting/postprocessing setting, and that calibrating a statistical model directly against the forecast fields can be a natural choice in many applications. In this study, however, we deliberately separate two conceptually distinct contributors to predictive uncertainty and error: (1) uncertainties intrinsic to the nowcasting algorithm (e.g., growth/decay, extrapolation limits, and scale-dependent predictability), and (2) uncertainties intrinsic to the radar/QPE chain, including radar measurement and processing effects (signal processing, Z–R conversion, and gauge calibration). Our objective is to quantify and mitigate the second component (i.e., the radar/QPE-related uncertainty and error structure) and then propagate this calibrated error model to rainfall nowcasts produced by the STEPS system.

Following Massari and Maggioni (2020), we use *error* to denote the deviation between an estimate and the (unknown) truth, and *uncertainty* to denote the expected range within which the true value is likely to lie at a given confidence level. Under this framing, calibrating the CSGD parameters directly on nowcasted rainfall fields would conflate the radar/QPE uncertainty and error (source 2) with the nowcasting-model uncertainty and error (source 1), making it difficult to attribute improvements to the radar/QPE uncertainty model itself. Moreover, direct calibration on nowcasts is intrinsically conditional on the particular nowcasting algorithm, configuration, and lead time; the calibrated parameters would change if the nowcaster (or its motion estimation, cascade decomposition, perturbation design, etc.) changes, and would typically require an archived set of forecasts for each lead time and regime to estimate stable CSGD parameters.

In contrast, our framework is modular: we first calibrate a radar/QPE error model using historical radar–gauge information (thus targeting source 2), and then apply the resulting pixel-wise probabilistic adjustment to any nowcast produced by the nowcasting system, enabling clearer attribution of what component is being corrected. We acknowledge the reviewer's point that such a design does not aim to "correct the nowcasting algorithm's intrinsic limitations" (source 1). This is intentional: the intrinsic predictability limits of precipitation nowcasting are well recognized, and improving them generally requires advances in the nowcasting model itself or model-specific learn-

ing/postprocessing (e.g., deep generative nowcasting) rather than an observation-error model alone. Our goal is to quantify how much of the total nowcast error can be reduced by addressing radar/QPE-related uncertainties, such as those associated with Z–R conversion and gauge adjustment. We will revise the manuscript to clarify this scope explicitly and to avoid implying that the proposed approach fully mitigates nowcasting-model errors. We will also add a discussion noting that a complementary, model-specific calibration on nowcasts (i.e., joint correction of sources 1 and 2) is a promising extension, consistent with recent bias-correction/postprocessing work for rainfall nowcasts.

Question 3

Reviewer comment

Since CRPS is the objective function of CSGD and is also one of the widely used metrics to evaluate the accuracy of ensemble prediction, I would recommend the authors to report the CRPS value as well for the comparison between linear & non-linear CSGD. It can also be used to evaluate ensemble nowcasting accuracy.

Response

We thank the reviewer for this insightful suggestion. We agree that, since the Continuous Ranked Probability Score (CRPS) is the objective function used for CSGD calibration and is also a standard metric for ensemble verification, reporting CRPS provides a more complete evaluation of the probabilistic nowcasts than point-estimate metrics alone. Following the reviewer's recommendation, we computed CRPS for the original STEPS ensemble and for the CSGD-adjusted ensembles (Linear and Non-Linear) across all lead times, and we will include these results in the revised manuscript for both the MIDAS dataset (hourly scale) and the EA-ST dataset (5-min scale).

The CRPS results show different lead-time behaviours for the two datasets, and these differences are consistent with the error characteristics of each dataset and with the role of pixel-wise CSGD postprocessing. For the MIDAS dataset (hourly verification), the CRPS skill improvement generally increases with lead time. This behaviour is consistent with the fact that the original STEPS nowcasts exhibit a systematic intensity bias that becomes more pronounced at longer lead times (as shown in Table 4), while the CSGD adjustment effectively reduces this bias. Because CRPS is sensitive to both reliability and sharpness, the bias reduction contributes substantially to the overall CRPS improvement, and this contribution becomes more visible as lead time increases.

For the EA-ST dataset (5-min verification), the CRPS skill improvement decreases with lead time. This behaviour is also physically and statistically consistent with the verification results. In this case, the original ensemble is already close to unbiased (OB $\approx$ 1.0, Table A6), so the CSGD adjustment does not gain the same benefit from bias correction as in MIDAS. At the same time, the forecast–observation spatial correspondence decays rapidly with lead time (i.e., increasing decorrelation effects, Table A4), while the CSGD-adjusted ensembles remain relatively sharp. Under these conditions, CRPS can penalize a sharp but misplaced forecast more strongly than a broader ensemble, because the forecast is "confident" but less well aligned with the observed precipitation pattern. In other words, as lead time increases in the 5-min verification setting, the decorrelation error becomes more dominant than the intensity-bias correction benefit, which explains the downward trend in CRPS skill.

These additional CRPS results therefore complement the point-estimate evaluations and help clarify the strengths and limitations of the proposed framework: the CSGD adjustment is particularly effective when systematic intensity bias is a dominant error source, whereas its CRPS benefit is reduced when decorrelation errors dominate and the original ensemble is already nearly unbiased. We will revise the manuscript accordingly and add a brief discussion to explicitly explain this metric-dependent behaviour.

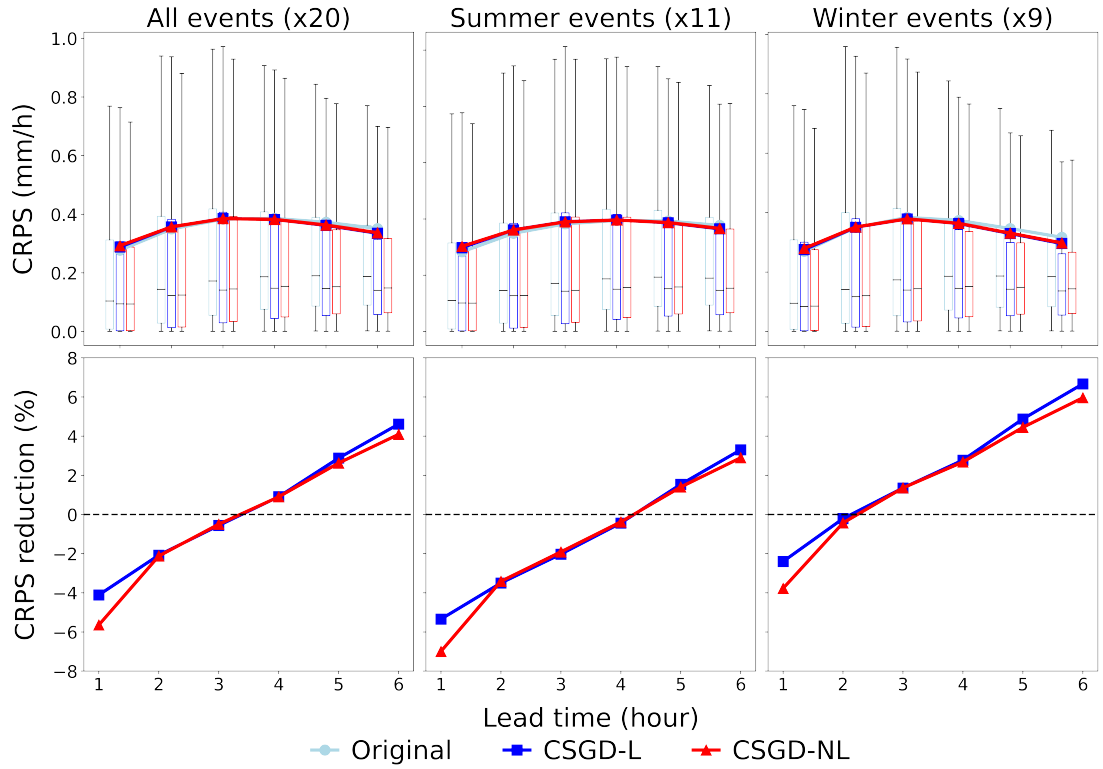

Figure 1: CRPS verification for the MIDAS dataset (hourly scale), comparing the original STEPS ensemble and the CSGD-adjusted ensembles (Linear and Non-Linear) across lead times.

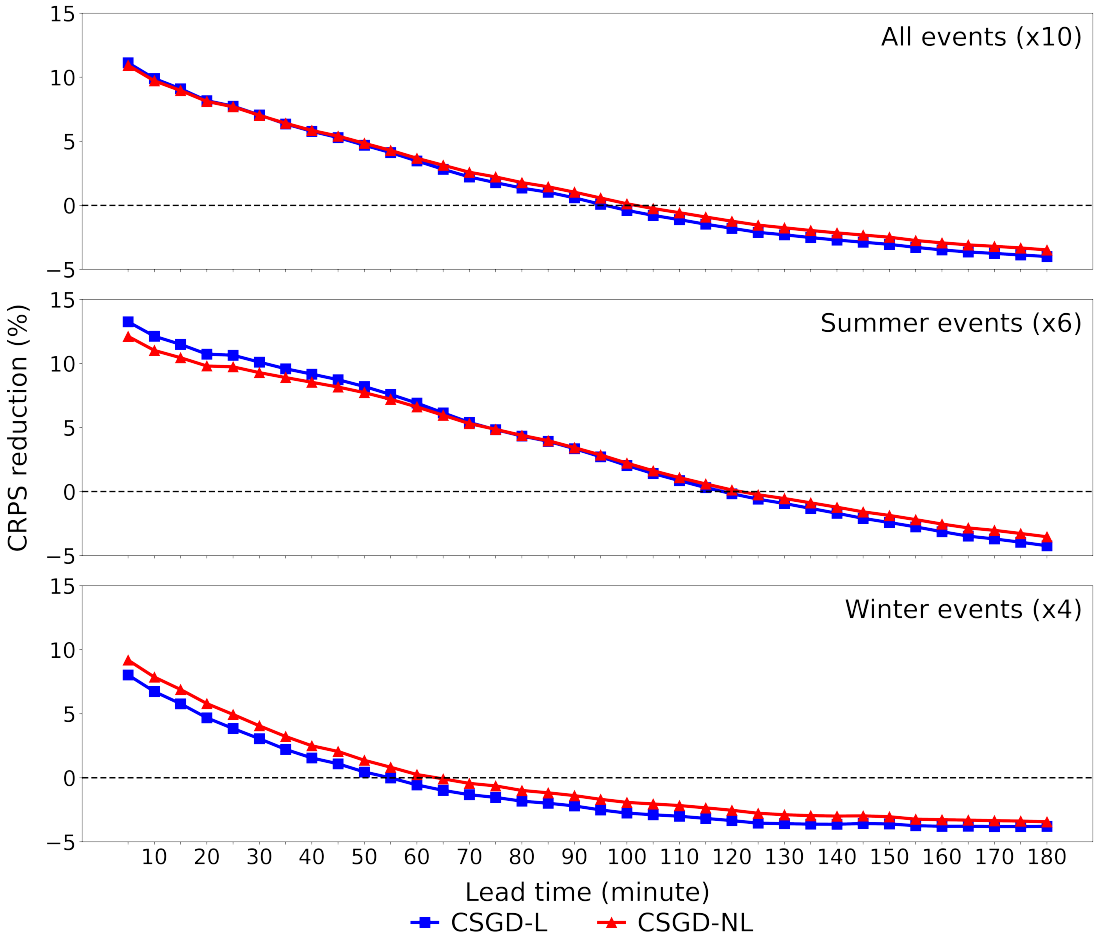

Figure 2: CRPS skill reduction/improvement for the EA-ST dataset (5-min scale), comparing the original STEPS ensemble and the CSGD-adjusted ensembles (Linear and Non-Linear) across lead times.

Question 4

Reviewer comment

Can the authors explain why, in Figs 8 & 9, the RMSE of CSGD is higher at a lead time of 1–3 hours? I would recommend the authors to report the evaluation for CSGD at the initial time step (i.e., lead time = 0 min, no nowcasting applied), so that we can investigate whether the error was propagated from the initial.

Response

We appreciate the reviewer's query regarding the RMSE behaviour at short lead times and the suggestion to evaluate the initial time step (lead time = 0 min). Following this advice, we added a lead-0 diagnostic (no nowcasting applied) and examined how the choice of a deterministic point estimator used to summarize the CSGD-adjusted predictive distribution (mean vs. median) affects RMSE-based comparisons.

Our framework produces a probabilistic adjustment (a full conditional distribution at each location

and time), while deterministic metrics such as RMSE require a single representative value. Two standard choices are the conditional mean and conditional median. The mean is the Bayes estimator under squared-error loss (L2), whereas the median is the Bayes estimator under absolute-error loss (L1). In precipitation applications, these summaries have different practical implications: for highly skewed distributions with substantial probability mass near zero, the conditional mean is always strictly positive and typically exceeds the median, while the median can be exactly zero when the probability of precipitation is below 0.5. Consequently, mean- and median-based point summaries can lead to different bias characteristics in low-rain regimes, and neither is universally optimal for all diagnostic objectives (Wright et al., 2017).

Consistent with this distinction, in the lead-0 diagnostic (no nowcasting), using the CSGD mean yields lower RMSE than the median-based point summary, as expected under an L2 metric when the uncertainty is comparatively smaller and the distribution is less dominated by heavy tails. However, for lead times > 0, the nowcasting distribution becomes increasingly skewed and intermittent, and using the mean as a point summary can introduce a drizzle-type positive bias; therefore, we retain the CSGD median as the primary point summary for lead times > 0 to ensure physically realistic dry-condition representation and robustness under skewness. To be fully transparent, we report both lead-0 mean and lead-0 median diagnostics (deterministic and ensemble cases) and keep the lead-time > 0 evaluation consistent with the median-based summary used in our main analysis.

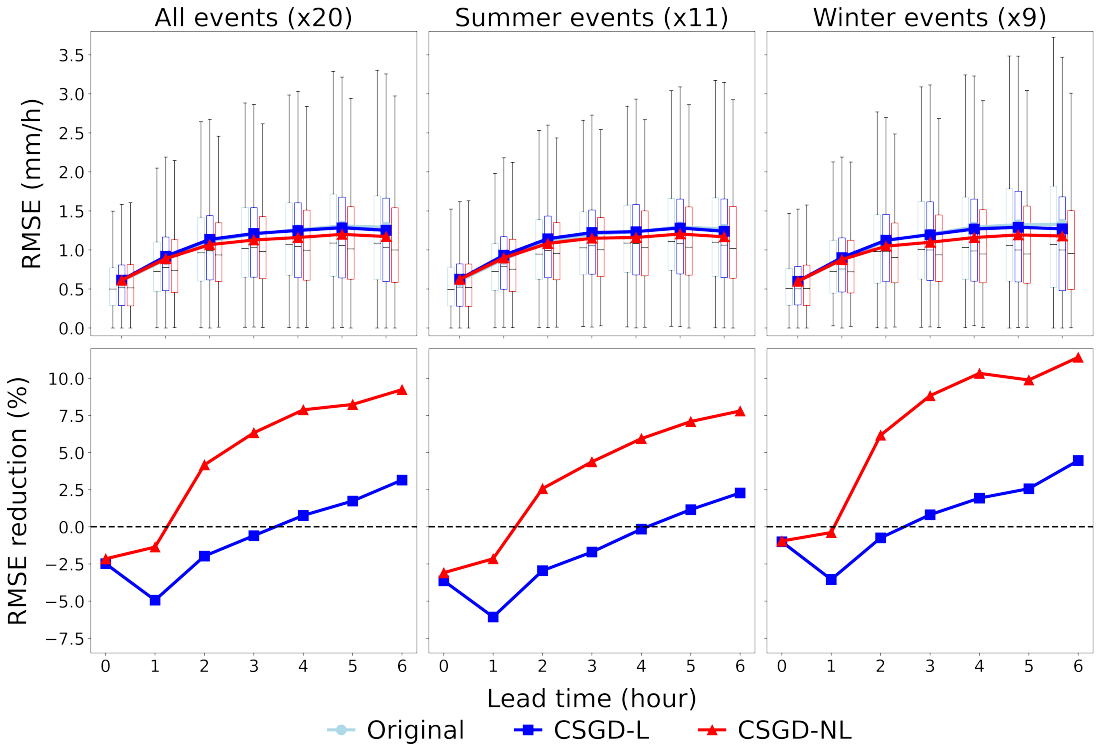

Figure 3: Deterministic nowcasting RMSE verification (including the added lead-0 diagnostic), using the CSGD median as the deterministic summary for all lead times.

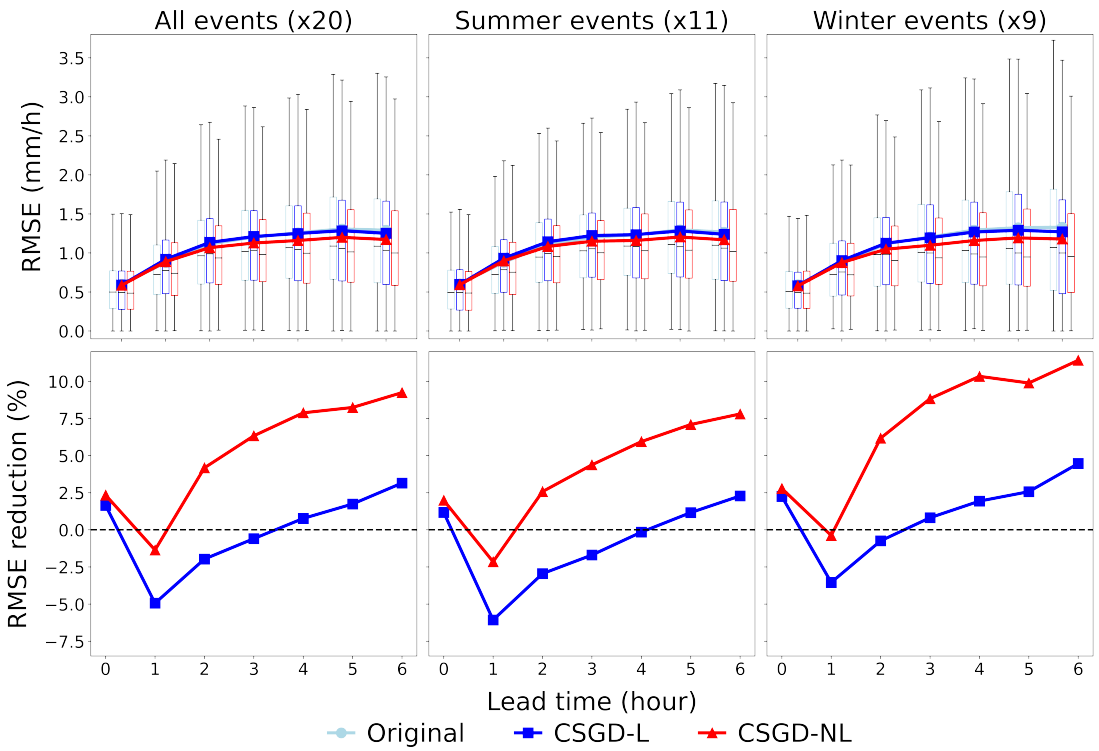

Figure 4: Deterministic nowcasting RMSE verification (sensitivity analysis). To highlight the initial error characteristics, the deterministic summary uses the CSGD mean at lead time 0, while remaining consistent with the main analysis by using the CSGD median for all lead times > 0.

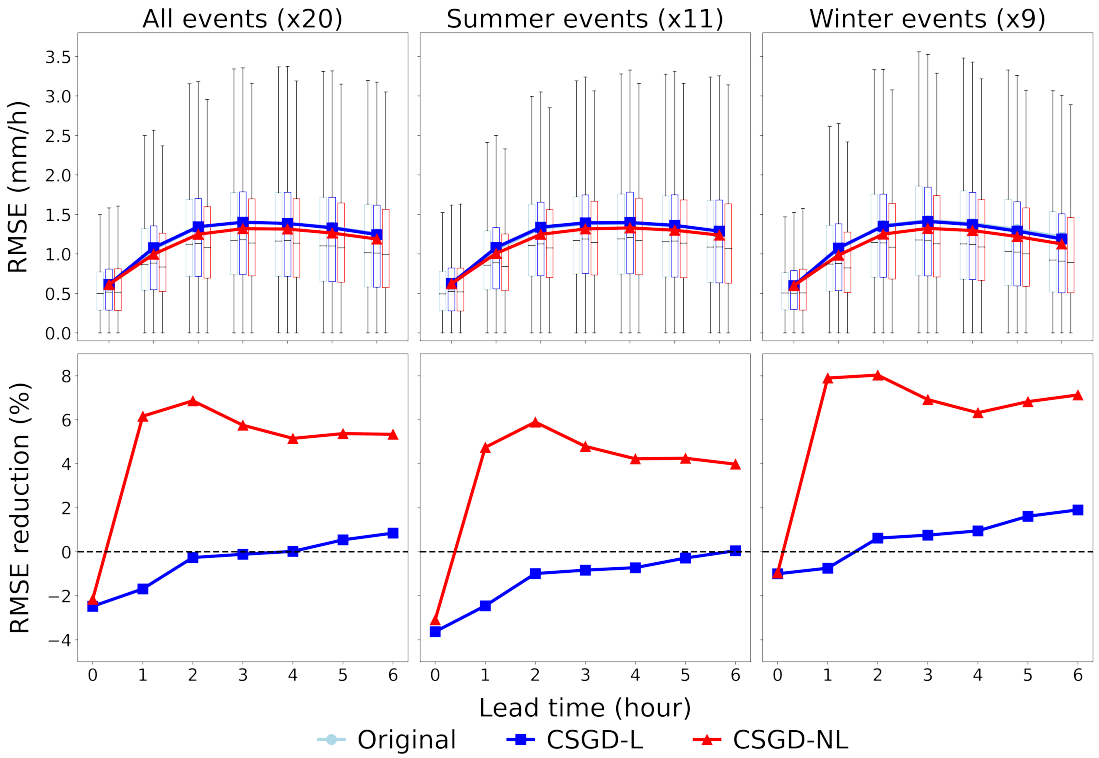

Figure 5: Ensemble nowcasting RMSE verification (including the added lead-0 diagnostic), using the CSGD median as the deterministic summary for all lead times.

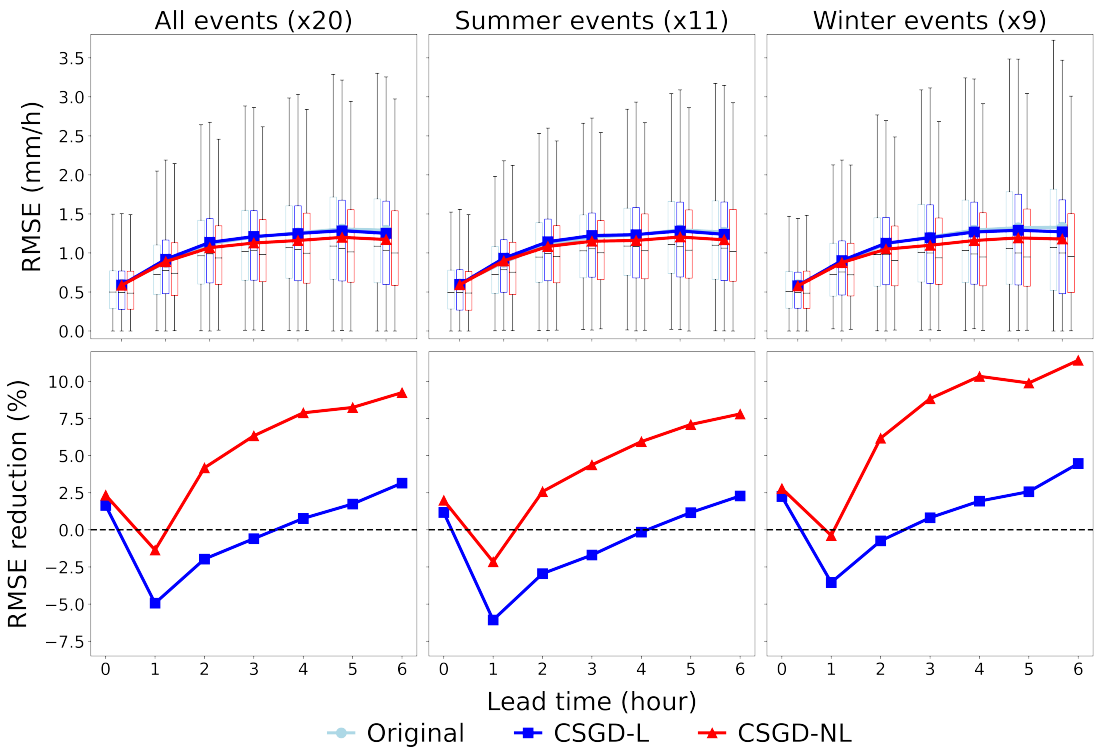

Figure 6: Ensemble nowcasting RMSE verification (sensitivity analysis). To highlight the initial error characteristics, the deterministic summary uses the CSGD mean at lead time 0, while remaining consistent with the main analysis by using the CSGD median for all lead times > 0.

Regarding the non-monotonic short-lead behaviour (higher RMSE at 1–3 h but not necessarily increasing thereafter), we find that this lead-time pattern is already present in the corresponding raw STEPS ensemble baseline for our event set. The CSGD adjustment does not introduce a new temporal shape; instead, it primarily modifies the amplitude calibration (bias/dispersion and intermittency) while largely preserving the lead-time dependence imposed by the nowcasting system. This behaviour is expected because the dominant contributors to lead-time growth in nowcasting error (e.g., displacement and growth/decay misspecification) are intrinsic to the nowcasting dynamics and are not directly corrected by an observation-error model. In the deterministic case, where the baseline does not exhibit the same non-monotonic pattern, the CSGD-adjusted RMSE likewise follows the deterministic baseline's lead-time evolution, supporting the interpretation that the observed short-lead shape is mainly driven by the underlying nowcasting behaviour rather than by the CSGD postprocessing itself.

We will revise the manuscript to explicitly distinguish (i) the lead-0 diagnostic (no nowcasting) from (ii) lead-time > 0 nowcast verification, to clarify the role of the point-summary choice (mean vs. median) under RMSE, and to emphasize that the proposed framework targets radar/QPE-related amplitude uncertainty/error and its propagation to nowcasts, rather than the intrinsic predictability limits of the nowcasting algorithm.

Question 5

Reviewer comment

I would recommend that the author provide more clarification in the CSGD and the nowcasting

model's performance in Tables A1–A6. As for different nowcasting methods, different CSGD models, and different metrics, the performance varies. This may imply some shortcomings in the current model that can be further improved in further study.

Response

We thank the reviewer for requesting additional clarification on Tables A1–A6. In the revised manuscript we will make the reading of these appendix tables explicit by linking each table to the error component it diagnoses and by summarizing the robust patterns that are already evident in the results.

For the EA-ST 5-min verification, Tables A1–A2 report the dispersion of RMSE ($\mathrm{RMSE}_{I_{90}}$) across lead times. Similar to the MIDAS results, $\mathrm{RMSE}_{I_{90}}$ increases with lead time (Table 2), reflecting growing forecast uncertainty, and the non-linear CSGD adjustment reduces $\mathrm{RMSE}_{I_{90}}$ across nearly all lead times and storm subsets, indicating a consistent reduction in error variability (i.e., uncertainty dispersion) under both deterministic and ensemble settings. We will explicitly state this in Appendix A and cross-reference the corresponding discussion in the main text.

Tables A3–A4 report Pearson correlation for deterministic and ensemble nowcasts, respectively. Here, the behaviour is metric- and scenario-dependent: for the 5-min deterministic case (Table A3), correlations do not decrease monotonically with lead time due to the chaotic evolution of unperturbed small-scale convective features, which leads to non-uniform responses to postprocessing and occasional correlation deterioration. By contrast, in the ensemble case (Table A4), where small-scale unpredictability is naturally smoothed out across members, correlations decay steadily, and the non-linear CSGD tends to improve correlation more consistently than the linear model. This difference perfectly aligns with the role of CSGD in our framework: the adjustment mainly targets radar/QPE-related amplitude errors (conditional bias/dispersion). At 5-min resolution, correlation is strongly influenced by spatial displacement/decorrelation errors intrinsic to the nowcasting dynamics, which cannot be directly resolved by a pixel-wise observation-error model. We will revise the appendix to highlight this deterministic-vs-ensemble contrast.

Finally, Tables A5–A6 report overall bias (OB). These tables show that the original EA-ST deterministic and ensemble nowcasts are close to unbiased during the first 3 hours, whereas the CSGD-adjusted nowcasts exhibit systematic underestimation over this early period. This behaviour is expected given our use of the predictive median as the deterministic point summary: for right-skewed precipitation distributions with substantial probability mass near zero, the median is conservative relative to the mean and can reduce light-precipitation contributions. Importantly, the tables also indicate that the original nowcasts progressively shift toward overestimation at longer lead times, and the CSGD adjustment mitigates this late-lead overestimation. We will clarify in the revised manuscript that (i) the early-lead OB degradation in EA-ST is mainly driven by the point-summary choice (median) in a regime where the baseline is already nearly unbiased, while (ii) the late-lead bias correction remains beneficial when the baseline bias grows.

Overall, Tables A1–A6 do not suggest numerical instability, but rather highlight the inherent metric-specific trade-offs when a probabilistic distributional postprocessor is collapsed into a single deterministic value, especially evaluated at a fine spatio-temporal scale where displacement errors dominate. We fully agree with the reviewer that these metric-specific variations highlight limitations in the current framework that warrant further study. For instance, if one prioritizes mass conservation (OB) for accumulation-focused diagnostics, exploring a bias-conserving point summary (e.g., the

mean, as discussed for t=0) or developing a joint postprocessing objective that explicitly constrains spatial accumulation alongside pixel-wise distributions would be highly promising extensions. We will make these interpretations explicit and add a short "how to read Tables A1–A6" guide in Appendix A to better navigate these nuances.

Question 6

Reviewer comment

The authors selected the median as the adjusted rainfall intensity for comparison. The median is likely to smooth out the extreme values. I would recommend use ensemble-based metrics to evaluate the ensemble accuracy. Both CSGD and ensemble nowcasting are good tools for ensemble-based decision making. Only focusing on median accuracy may not be a comprehensive evaluation.

Response

Thank you for this important comment. We agree that evaluating an ensemble solely via a single deterministic summary (e.g., a median) is not fully comprehensive, and we will clarify both (i) what "median" means in our framework and (ii) which ensemble-based verification will be reported.

First, in our implementation the "adjusted rainfall intensity" is not obtained by taking the median across STEPS ensemble members. Rather, for each grid point (and for each member in the ensemble case), we map the nowcasted rain rate to a conditional CSGD predictive distribution, and then extract the median of that conditional CSGD as a point estimate for computing conventional deterministic metrics (RMSE, bias, etc.). This choice follows the practice that error metrics can be computed against the conditional CSGD median, and is motivated by the fact that the conditional CSGD mean is always nonzero and can exceed the median in low-rainfall regimes, potentially introducing drizzle-like bias; the median is more robust for highly skewed precipitation distributions.

Second, we fully agree that, because both STEPS and CSGD are probabilistic tools, the revised manuscript should include proper ensemble-based scores in addition to median-based (deterministic) accuracy. Importantly, CRPS is already the objective function used in our CSGD parameter estimation (i.e., parameters are calibrated by minimizing CRPS between empirical and theoretical CDFs). Therefore, in the revision we will explicitly report CRPS for the ensemble nowcasts before/after CSGD adjustment (as noted in Response 3).

Minor revision

Response

We thank the reviewer for pointing out these details. We have addressed both minor points in the revised manuscript as follows:

Point 1: We completely agree that aligning the flowchart with the text structure improves readability. This new layout better illustrates the sequential connections between the CSGD modelling (Section 3.2), the spatial structure analysis and parameter interpolation (Section 3.3), and the real-time STEPS nowcasting and adjustment (Section 3.4).

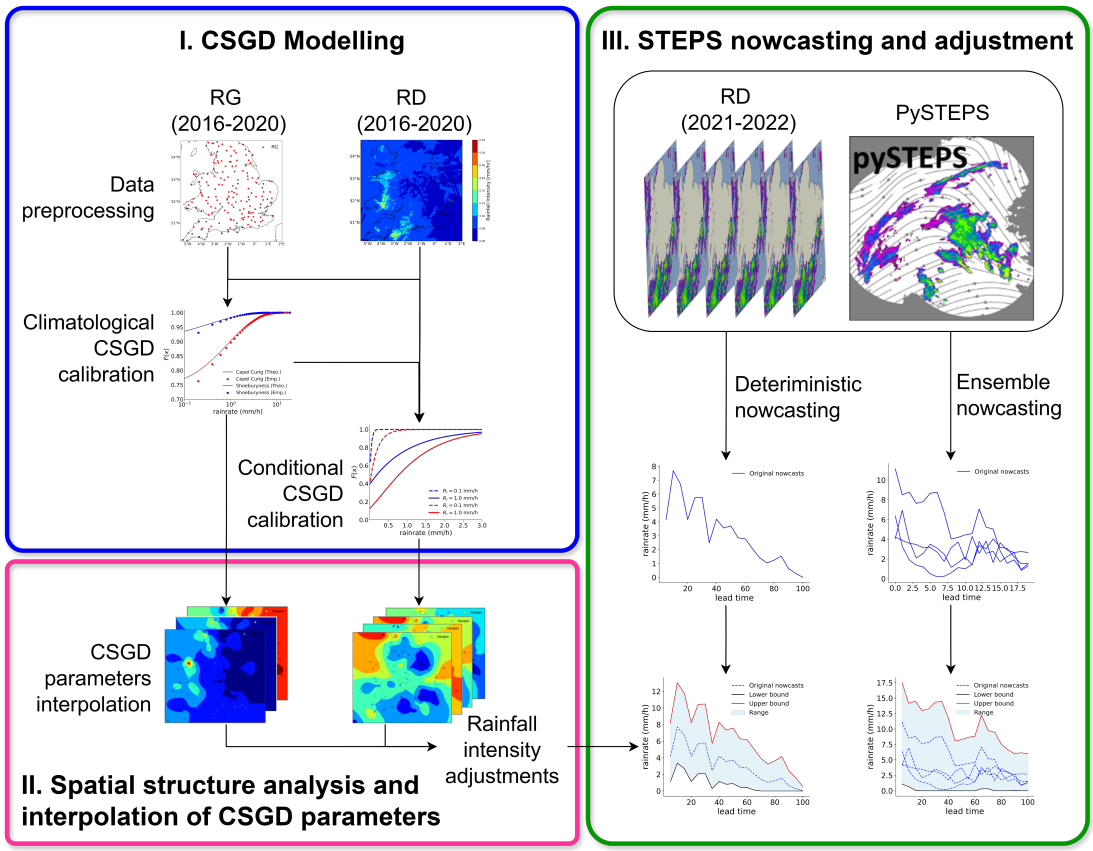

Figure 7: The revised workflow (updating Figure 3 in the manuscript), now restructured into three main boxes (I, II, and III) to explicitly align with the methodological steps detailed in Sections 3.2, 3.3, and 3.4, respectively.

Point 2: We apologize for the typo. We have corrected the notation to $\text{RMSE}_{I_{90}}$ on line 442 and have thoroughly checked the rest of the manuscript to ensure consistent formatting.