Response to Reviewers

We sincerely thank both reviewers for their constructive and insightful comments, which have greatly helped us improve the clarity, rigor, and presentation of the manuscript. Below, we provide a detailed, point-by-point response to each comment. For clarity, reviewer comments are shown in *italic*, and our responses follow each comment.

1 Response to Reviewer #1

1.1 Justification of the model's rationale

Reviewer comment: My concern regarding the rationale of the paper has been partially addressed, and I appreciate the analysis on the relative importance of the input features. However, the manuscript still lacks a concrete explanation or experimental support for the mechanism by which the GAN architecture improves performance over simpler or non-generative alternatives. I think my point here will become more clear with the explanation on the need of simpler methods for the comparison.

Response: We sincerely thank the reviewer for this insightful comment. We fully agree that meaningful justification for adopting a deep-learning GAN framework requires explicit comparison with simpler methods to clarify its added value. In response, we have made the following additions and clarifications in the revised manuscript:

1. Incorporation of MSEM as a baseline method. We introduced **MSEM (Multi-Source Ensemble Mean)**, a traditional **linear similarity-weighted fusion** method, as a strong and operationally relevant baseline. This method is widely used in multi-model ensemble forecasting systems due to its **simplicity, interpretability, and low computational cost**. Our analyses (Overall Evaluation, Spatial Analysis, Significance Testing, and Case Study) reveal that:

- *MSEM provides stable and cost-effective performance for light to moderate rainfall events*, showing that simple linear fusion approaches can deliver solid results in less complex scenarios.
- *MSEM struggles with heavy rainfall (20–40 mm) and extreme precipitation*, where TS and FSS scores drop noticeably, indicating that linear weighting cannot capture the spatial organization of intense rainfall.

2. Why GFRNet outperforms MSEM. While MSEM applies static weights and lacks the ability to model nonlinear relationships, **GFRNet leverages deep feature extraction and a generative framework** to overcome these limitations:

- *Dynamic nonlinear fusion:* GFRNet adaptively integrates multi-source NWP inputs across space and time, learning complex and evolving dependencies beyond what static weighting can achieve.
- *GAN-based structural constraints:* In addition to MSE/MAE content losses, GFRNet uses an adversarial loss to constrain the **precipitation structure distribution**, ensuring that the generated outputs preserve realistic spatial patterns.
- *Superior performance in challenging cases:* GFRNet more accurately reconstructs the core structure of heavy rainfall and avoids spurious "spill-over" precipitation, yielding significantly higher TS and FSS scores than MSEM for extreme rainfall scenarios.

In summary, by including MSEM as a strong and cost-effective operational baseline, we demonstrate that GFRNet not only matches the stability of simpler approaches for routine rainfall but also **clearly surpasses them for heavy rainfall and complex precipitation structures**. This comparison provides strong experimental evidence supporting the necessity of a GAN-based deep learning framework in this study.

1.2 Lack of comparison with simpler methods

Reviewer comment: This concern remains unaddressed. In your response, you acknowledge that such comparisons would add value but state that "primary focus of our study is to address the limitations of non-generative models and evaluate the advantages of GAN-based techniques". However, such advantages can only be meaningfully evaluated through comparison

with classical alternatives—especially those that might yield similar results with significantly lower complexity and higher interpretability.

If the primary contribution lies in deep learning architectures, and precipitation serves primarily as a benchmark, a journal more focused on artificial intelligence might be a more suitable venue.

The lack of comparison with simpler methods is not a minor point. For instance, the "blurriness" of NWP model outputs may result, at least in part, from the bilinear interpolation applied. What if elevation or forecast cycle were used to locally adjust the interpolation? What if a local regression approach was used to estimate rainfall as a function of elevation in each subregion? Such strategies might outperform bilinear interpolation without the need for complex machine learning models.

Similarly, forecasting methods such as the analogue method—a well-established baseline in meteorology—could be implemented via K-nearest neighbors, even avoiding interpolation altogether. Dimensionality reduction techniques could also be applied to simplify the input space before using such methods. Without at least exploring these simpler alternatives, the comparison is incomplete and the added value of GFRNet remains unclear. Deep learning and AI are undeniably powerful, but I believe they should be applied where their added accuracy justifies the associated cost in complexity and loss of interpretability. In summary, without a fair comparison with classical methods, the paper's core value proposition is not convincingly demonstrated.

Response: We sincerely appreciate the reviewer's thoughtful comment regarding the need for broader comparisons with simpler methods. We fully agree that evaluating the advantages of deep learning models requires comparison with widely used, classical ensemble post-processing techniques. In the revised manuscript, we have clarified the following:

1. Inclusion of MSEM as a core baseline. We introduced *MSEM (Multi-Source Ensemble Mean)*, a traditional **linear similarity-weighted fusion** approach, as a strong and representative baseline. MSEM has been widely adopted in operational multi-model forecasting systems because of its simplicity, interpretability, and low computational cost. Our analyses (Overall Evaluation, Spatial Analysis, Significance Testing, and Case Study) show that:

- MSEM performs **consistently well for light and moderate rainfall events**, demonstrating that simple linear fusion can still deliver stable performance in straightforward scenarios;
- However, for **heavy rainfall (20–40 mm) and extreme precipitation**, MSEM's performance drops significantly (TS and FSS decline), and it fails to capture the spatial organization of intense rainfall—highlighting the inherent limitations of static weighting.

2. Consideration of PMM and why the manuscript focuses on MSEM. We also tested *Probability Matching Mean (PMM)*, another classical ensemble post-processing method that aligns the cumulative distribution functions (CDFs) of model forecasts before averaging to better preserve precipitation intensity distributions. Tests on the 2024 flood season showed that:

- **PMM and MSEM behave almost identically**: both methods are stable for light and moderate rainfall but show clear degradation for heavy and extreme rainfall(Figure 1), falling short of the deep learning models in restoring precipitation structure.
- Because PMM and MSEM led to the same conclusions, we chose to present only MSEM in the main text to maintain narrative focus and avoid redundancy. We note here that PMM was evaluated, and it does not change the key findings.

3. Why GFRNet outperforms linear methods like MSEM/PMM. Unlike MSEM and PMM, which are **static linear methods**, GFRNet leverages deep learning and a generative framework to achieve several key advantages:

- *Dynamic nonlinear fusion:* GFRNet adaptively integrates multi-source NWP inputs across space and time, learning complex, context-dependent relationships beyond static weighting;
- *GAN-based structural constraints:* Beyond content losses (MSE/MAE), GFRNet incorporates an adversarial loss that constrains the precipitation structure distribution, leading to more realistic and coherent rainfall fields;
- *Superior performance for heavy rainfall:* GFRNet reconstructs the core structure of heavy rainfall more accurately, reducing spurious "spill-over" precipitation and achieving significantly higher TS and FSS than MSEM and PMM;

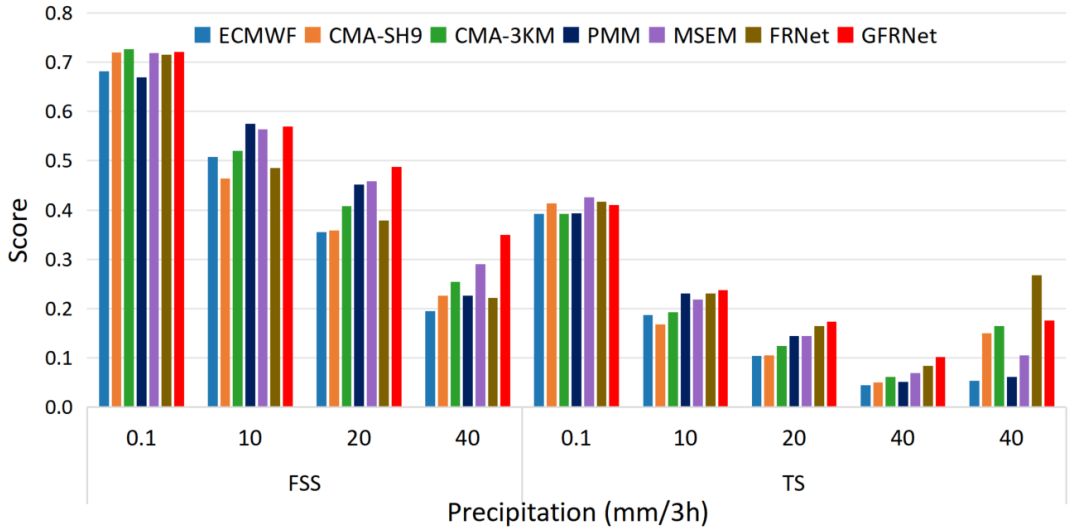

Figure 1. Comparison of forecast performance among multiple methods — ECMWF, CMA-SH9, CMA-3KM, PMM, MSEM, FRNet, and GFRNet — for the 2024 flood season, evaluated using Fractions Skill Score (FSS) and Threat Score (TS) at precipitation thresholds of 0.1, 10, 20, and 40 mm/3 h. The results illustrate that while traditional ensemble methods (MSEM, PMM) maintain stable performance for light and moderate rainfall, their skill declines sharply for heavy and extreme rainfall, whereas GFRNet delivers consistently higher FSS and TS scores at higher thresholds, demonstrating its superior capability in reconstructing precipitation structure and intensity.

– *Greater extensibility and future potential:* Deep learning models like GFRNet offer a fundamentally higher ceiling. As more physical variables (e.g., wind, temperature, humidity, geopotential height) are incorporated in future work, traditional statistical methods would struggle to effectively utilize these complex features. GFRNet, by contrast, can learn the nonlinear relationships between these variables and precipitation generation in an end-to-end fashion—paving the way for further improvements in physical consistency and generalization.

In summary, we have clarified that PMM was also tested and yielded the same conclusions as MSEM, which is why only MSEM is presented in the main text for conciseness. The comparisons confirm that GFRNet is **at least as stable as linear methods for light-to-moderate rainfall, clearly superior for heavy rainfall and structure restoration, and uniquely positioned to benefit from additional physical variables in the future**—justifying the adoption of a GAN-based deep learning framework in this study.

1.3 Experimental design and statistical significance

Reviewer comment:I remain concerned about the experimental design. Most critically, the manuscript does not include any statistical hypothesis tests to determine whether observed differences are significant or due to random variability. Table 4, for instance, is difficult to interpret without such tests—particularly considering the limited dataset. This should be a central concern for the authors.

This issue is exacerbated by the heavy filtering applied to the original dataset. I still do not understand the motivation behind defining "valid samples," nor do I fully understand Figure 2. If samples are filtered based on precipitation criteria, how is it possible that the resulting dataset still contains a mix of valid and invalid samples? Could the authors clarify this?

Moreover, I question how Figure 2b demonstrates a fat-tailed distribution. What is the metric or evidence supporting this claim? Based on this assumption, a custom weighted loss function is designed. But this chain of reasoning—and the corresponding design choices—seems to hinge on multiple small decisions that are not fully justified. I am concerned that these

steps may unintentionally lead to overfitting—not necessarily through the model itself, but through tuning the data processing to the specific characteristics of the dataset.

In this context, I find particularly problematic the statement in line 136: "This increase in the proportion of valid samples improved the stability and efficiency of the model training." This suggests that the validity criteria were defined a posteriori to improve performance, which undermines the generalizability of the results. All these comments should also be understood in the context of a somehow limited dataset.

Response: We sincerely thank the reviewer for these detailed and constructive comments. We fully understand the importance of rigorous experimental design and transparent reporting, and we have addressed these concerns through multiple revisions in the manuscript, as detailed below:

1. Addition of statistical significance testing. In Section 3.4 of the revised manuscript, we added a comprehensive **statistical significance analysis**. Using data from the **2022–2024 rainy seasons**, we performed paired t-tests on both the **full-sample set** and the **Top 10% subset**, and marked the statistical levels ($p < 0.05$, $p < 0.01$, $p < 0.001$) directly in the boxplots. These results confirm that GFRNet's performance advantages in most precipitation scenarios are **statistically significant rather than random fluctuations**, while also transparently reporting thresholds (e.g., 0.1 mm) where differences are not significant. This provides a clear statistical foundation for our conclusions.

2. Clarification and revision of the "valid sample" concept. We acknowledge that the previous description of "valid samples" caused confusion, and we have **removed the problematic phrasing from the manuscript**. In the revised version, we clarified the sampling logic:

- *Image-level sampling* is applied **only to the training set**, to reduce the large proportion of "no-rain" samples and improve learning efficiency for rainy scenes.
- **validation and test sets remain entirely unsampled and use all original data**, ensuring unbiased evaluation results and avoiding any "a posteriori tuning."
- Even after image-level sampling, rainfall at the **pixel level** remains highly imbalanced, which directly motivated the weighted loss design.

3. Justification for the long-tail distribution and weighted loss.

- The **long-tail distribution of rainfall at the pixel scale** is a widely acknowledged fact in meteorological deep learning literature; nearly all relevant studies (Ayzel et al., 2020; Tan et al., 2024; Shi et al., 2015) face this issue and address it with loss-weighting strategies.
- Figure 2b illustrates this imbalance, showing that even after image-level sampling, extreme-rainfall pixels account for less than 0.3% of all pixels.
- Based on this observation, we designed an exponential weighted loss (MSE/MAE + weighted loss). The ablation study in Table B1 validates this design: removing the weighting caused a clear drop in TS for precipitation $\geq$10 mm.
- Furthermore, the loss weighting **did not lead to overfitting**—GFRNet's performance remained stable on the independent 2023–2024 test sets, confirming that the model learned robust and generalizable patterns.

4. Addressing generalization concerns and dataset size. We share the reviewer's concern about potential overfitting. To assess generalization, we added **independent flood season data from 2023 and 2024** (not used for training) for evaluation.

- GFRNet demonstrated **stable and consistent performance across 2022–2024**, with TS and FSS scores showing no signs of instability or over-specialization.
- We acknowledge that **further data expansion remains desirable**. Incorporating additional years and broader regional coverage will likely further improve the model's performance and generalization in future work.

In summary, the revised manuscript now includes **statistical significance testing**, removes confusing terminology around "valid samples," clarifies the rationale for sampling and weighted loss design, and presents new evidence of GFRNet's robust generalization using independent test sets. These changes ensure that the study's design, methodology, and conclusions are more transparent, rigorous, and reproducible.

1.4 MS-SSIM vs. RMSE

Reviewer comment: Another methodological point that requires further justification is the choice of MS-SSIM over RMSE as a primary evaluation metric. First, given its central role in the analysis, the MS-SSIM metric should be defined mathematically so that readers can judge its appropriateness. Second, the manuscript should offer a clear argument as to why spatial structure should be prioritized over magnitude in evaluating model performance. As shown in Table 5, ECMWF achieves the best RMSE. Without knowing how ECMWF would perform spatially using a better interpolation method (e.g., structure-aware or topography-guided), it is difficult to accept that GFRNet clearly outperforms it overall. The manuscript would benefit from a more explicit discussion of the tradeoffs involved.

Response: We sincerely thank the reviewer for these valuable comments. In the revised manuscript, we have expanded and refined the relevant sections to provide a more comprehensive and transparent explanation, summarized as follows:

1.Rationale for a multi-metric evaluation framework

- *Necessity of multiple metrics:* Precipitation forecast evaluation is inherently complex, and no single metric can fully represent model performance. We therefore adopted four complementary categories of metrics: (i) Binary metrics (TS, POD, FAR, BIAS), (ii) Neighborhood metric (FSS), (iii) Continuous metric (RMSE), and (iv) Structural metric (MS-SSIM). This combination enables cross-validation and a more nuanced understanding of model capabilities.
- *Clarification:* We explicitly stated that MS-SSIM is not the "core" evaluation metric of this study but rather a supplementary one, offering a structural perspective alongside TS, FSS, and RMSE.

2.Clarification of MS-SSIM's role and added value

- *Established use:* MS-SSIM has been widely adopted in short-range and nowcasting studies (Yin et al., 2021; Tan et al., 2024). We added references to support its relevance.
- *Mathematical definition:* In Section 2.3, we now provide the full mathematical formulation of MS-SSIM, ensuring transparency and reproducibility.
- *Additional insights:* MS-SSIM captures **spatial structure coherence** (e.g., rainfall band continuity, sharpness of edges) that binary scores or RMSE alone cannot. For example, FRNet showed higher TS but lower MS-SSIM, highlighting its blurriness in rainfall structure — an issue GFRNet mitigates effectively.

3.ECMWF's performance, contribution, and relation to GFRNet

- *Strengths of ECMWF:* ECMWF remains an internationally trusted NWP system. Our results reaffirm its strong performance for **light and moderate rainfall (0.1–10 mm)** and its skill in predicting synoptic-scale rainfall distribution.
- *Contribution to GFRNet:* Ablation experiments (Appendix B) confirm that ECMWF consistently improves GFRNet's forecasts across all rainfall intensities, proving its foundational importance in our fusion approach.
- *Acknowledging limitations:* Metrics including TS, FSS, case analyses, and significance testing show that ECMWF underestimates **heavy rainfall ($\geq$20 mm)**, where GFRNet adds value.
- *Positioning GFRNet:* GFRNet builds upon ECMWF's solid base, leveraging GAN-based fusion and structural constraints to enhance heavy-rainfall prediction and spatial fidelity.

In summary, the revised manuscript clarifies the rationale for a multi-metric framework, properly defines MS-SSIM, and clearly frames ECMWF's indispensable role, while explaining how GFRNet enhances and extends its strengths.

1.5 Additional minor comments

Reviewer comment:I believe my original comment on the case studies still holds: their selection and interpretation could be more clearly justified. Figures 4, 5, and 6 remain difficult to interpret. Improvements in visual clarity and labeling would greatly enhance their value.

Response: We sincerely thank the reviewer for these constructive comments. In the revised manuscript, we have carefully addressed these issues as follows:

1. **Case selection and representativeness**
   - The original submission included three cases from the 2022 rainy season. Following your feedback and after reconsidering the overall balance of the manuscript, we refined this section to focus on **two representative cases**: one from the 2022 rainy season and one from the 2024 rainy season.
   - Both cases are large-scale, high-impact precipitation events, capturing different synoptic settings (e.g., frontal rainfall and precipitation influenced by the subtropical high). This selection provides a more concise yet comprehensive basis for understanding model behavior.
2. **Figures and caption improvements**
   - All case study figures (Figures 4–6) have been fully redesigned: color schemes were refined, legends and axes were clarified, and titles were standardized for consistency and better visual interpretation.
   - **Every figure caption was rewritten and expanded** to provide clearer guidance for interpretation, explicitly highlighting key regions and differences among models, making the figures far easier to read and understand.
3. **Key insights from the two cases**
   - *Distinct NWP characteristics:* The two cases highlight that each NWP model contributes differently: **ECMWF** provides strong skill in large-scale precipitation placement but tends to underestimate heavy rainfall; **CMA-SH9** better captures some mesoscale organized rainfall; and **CMA-3KM** is more sensitive to local convective rainfall but introduces more noise and false alarms.
   - *Role of MSEM:* The MSEM method linearly fuses NWPs through similarity-based weighting, producing a stable and cost-effective baseline. It performs reliably for light and moderate rainfall, but its ability to recover fine details of heavy rainfall remains limited.
   - *GFRNet's strengths and limitations:* GFRNet leverages a GAN-based framework to nonlinearly integrate multi-source NWP information. This allows the model to capture complementary strengths of different NWPs and restore spatial structure for heavy rainfall more effectively. However, GFRNet's strategy is deliberately slightly conservative to avoid overprediction, which can occasionally lead to mild underestimation of extreme rainfall, and its performance still depends on the quality of the NWP inputs.

In summary, the revised case study section now presents a more concise and representative selection of events, integrates clearer figures and captions, and delivers deeper insights into the behavior of NWPs, the blending logic of MSEM, and the strengths and limitations of GFRNet, thereby enhancing the overall clarity and value of this section.

2 Response to Reviewer #2

2.1 Model Explainability and Physical Reasoning

Reviewer comment: Although the author has verified the model's performance through experiments, the physical mechanism explaining how the model improves precipitation forecasts via the GAN strategy is insufficient. The reviewers pointed out that

the impact of terrain and meteorological features on precipitation needs more in-depth discussion, particularly on how the model captures these physical relationships.

Response: We thank the reviewer for highlighting this important point. In the revised manuscript, we made several additions and modifications to address this concern:

1. **Ablation study (Appendix B2 Input Source Contribution Analysis)**
   - We conducted a more systematic set of ablation experiments, assessing not only the removal of different NWP data sources (ECMWF, CMA-SH9, CMA-3KM) but also the effects of removing topographic information and time-encoding features.
   - Based on these experiments, we calculated the **Relative Importance Score (RIS)** for each input feature across different precipitation intensity ranges, quantitatively revealing the contribution of each feature to model predictions.
   - These results provide indirect yet quantitative insights into what GFRNet is "learning" during the training process.
2. **Spatial analysis (Section 3.2)**
   - We expanded the spatial performance analysis, showing that **GFRNet delivers more pronounced improvements in regions with complex terrain**.
   - This indicates that the model effectively leverages terrain-related signals and mitigates systematic biases present in NWP inputs, suggesting the model may be learning terrain-related information explicitly or implicitly.
3. **Physical mechanism discussion (Discussion section)**
   - We elaborated on GFRNet's generative mechanism, explaining how the **adversarial framework drives the model** to generate precipitation fields that better capture realistic rainfall structures—particularly in retaining fine-scale features of intense rainfall.
   - We also emphasized that **GFRNet integrates multi-source NWP inputs** to constrain the generation process, striking a balance between structural realism and physical consistency.

In addition, we candidly acknowledge that despite these efforts, it remains challenging to conclusively identify the exact physical laws learned by deep learning models. This limitation is common across AI applications in meteorology, not unique to this study. Enhancing the physical interpretability of such models will be a key focus of our future work.

The relevant additions can be found in **Section 3.2 (Spatial Performance Analysis)**, **Appendix B**, and the **Discussion section**.

2.2 Comprehensiveness of Experimental Design

Reviewer comment: While the author has conducted ablation experiments to verify the role of key components (such as the SE module and weighted loss function), comparisons with other simple models (such as Random Forest or SVM) are lacking, making it difficult to fully prove whether the complexity of the GAN is necessary.

Response: We thank the reviewer for this insightful comment regarding the comprehensiveness of the experimental design. In response, we have made the following revisions and clarifications:

1. **Introduction of MSEM as a baseline model** We incorporated **MSEM**, a linear similarity-based ensemble method that combines multi-source NWP results, as an additional baseline. MSEM is simple, interpretable, and computationally efficient, and does not involve deep learning or adversarial training. This provides a clear reference point for assessing whether the complexity of the GAN framework is warranted.
2. **Inclusion of MSEM comparisons across sections** We systematically added comparisons between GFRNet and MSEM throughout the manuscript, including the Overall Evaluation, Spatial Analysis, Significance Testing, and Case Study sections, to more comprehensively illustrate how GFRNet compares to a straightforward linear fusion approach under various precipitation scenarios.

3. **Findings highlight the added value of the GAN framework** Our analysis shows that MSEM performs robustly in light rainfall and some moderate rainfall cases, but its performance drops notably at the **20–40 mm** threshold and for heavy rainfall, where it cannot capture complex precipitation structures. In contrast, GFRNet's generative adversarial mechanism and nonlinear fusion capability provide clear improvements in moderate-to-heavy rainfall and high-impact precipitation events.

4. **Cost–benefit discussion of the GAN framework** We added discussion on the tradeoff between complexity and benefit. While GFRNet uses adversarial training, the GAN architecture is deliberately simple compared with more complex approaches (e.g., diffusion models or Transformer-based architectures) and is relatively lightweight. Given the clear gains in structural realism and heavy rainfall forecasting skill, we believe this modest complexity is justified and cost-effective.

In summary, by adding MSEM as a clear linear baseline, systematically comparing it with GFRNet, and discussing the cost–benefit tradeoff, we demonstrate that the GAN framework is both necessary and practical in the context of this study.

2.3 Data Size and Generalization Ability

Reviewer comment: Using only 4 years of data may limit the model's generalization ability, and the model's performance in other regions or over longer time series has not been verified.

Response: We sincerely thank the reviewer for raising this valuable point regarding the data size and generalization of GFRNet. We agree that generalization is a common challenge faced by deep learning models in meteorological applications.

1. First, to evaluate the generalization ability of GFRNet, we expanded the test period in the revised manuscript by including the 2023 and 2024 rainy seasons as independent tests, forming a continuous three-year test set (2022–2024, with approximately 3,500 samples). Results show that GFRNet's TS, FSS, and MS-SSIM scores remained consistent and steadily superior across the additional two years, particularly in systematic heavy rainfall events, where spatial structure reconstruction and intensity control remained stable. This provides evidence of the model's robustness and, to some extent, demonstrates that GFRNet's methodology has the potential for sustainable and stable application in operational contexts.

2. Second, regarding the scale of the training data, we acknowledge that the model was primarily trained on historical data from 2019–2021, with 2023–2024 data used only for independent testing. Incorporating a longer training record would likely further improve performance. This has been included in our future work plan, and we will progressively expand the dataset to include more years and regions for training and validation, aiming to further enhance the model's generalization and applicability.

3. Finally, while GFRNet demonstrated stable performance across the 2022–2024 rainy seasons and showed encouraging interannual generalization, the training and evaluation data are still largely limited to the most recent five rainy seasons. Compared to longer time spans and broader climatological contexts, this relatively narrow data range may impose some constraints on the model's generalization. In future work, we plan to expand the training and validation datasets to cover longer historical periods and diverse climatic regions, and to explore cross-regional and cross-temporal validation to more systematically assess the model's applicability and robustness.

In summary, by expanding the testing period, acknowledging current limitations, and outlining future dataset growth, we present a clearer picture of GFRNet's generalization performance and a concrete plan for strengthening it further.

2.4 Consistency Analysis of Results

Reviewer comment: Some results are contradictory (e.g., GFRNet has a lower TS than FRNet at the 40mm threshold), and the author attributes this to the conservativeness of the GAN but lacks statistical significance analysis.

Response: We sincerely thank the reviewer for this valuable comment on the consistency analysis of results. The reviewer noted that some findings appear contradictory (for example, GFRNet shows a slightly lower TS than FRNet at the 40 mm

threshold) and suggested adding statistical significance tests, as well as discussing the pros and cons of GFRNet's "conservative" strategy.

In response, we have made the following clarifications and additions in the revised manuscript:

1. **Statistical significance analysis.** In Section 3.4, we conducted paired *t*-tests for all precipitation thresholds (0.1, 10, 20, and 40 mm) across both the *all-sample set* and the *Top 10% coverage subset*. Significance levels ($p < 0.05$, 0.01, 0.001) are clearly annotated in Figure 12, ensuring that every reported difference between models is supported by rigorous statistical evidence.

2. **Clarification of TS results at the 40 mm threshold.** In the *all-sample set*, GFRNet's TS is slightly lower than FRNet's, but paired *t*-tests indicate that this difference is **not statistically significant**. In the *Top 10% subset* (representing more organized extreme rainfall events), FRNet's TS is marginally higher than GFRNet's (also without statistical significance). However, this "advantage" comes with a tradeoff: **FRNet is noticeably more aggressive in high-rainfall forecasts, which results in significantly lower FSS scores compared to GFRNet**. This indicates that FRNet has shortcomings in restoring spatial precipitation structures, while GFRNet places greater emphasis on maintaining spatial consistency and coherence.

3. **Rationale and implications of GFRNet's "conservative" strategy.** GFRNet adopts a more balanced and moderately conservative approach to heavy rainfall forecasting. This is not a performance weakness but a deliberate design choice: by avoiding overprediction, GFRNet substantially alleviates FRNet's issue of high BIAS (e.g., BIAS > 1.8 at the 40 mm threshold). This "conservativeness" helps reduce false alarms of extreme rainfall and supports the delivery of forecasts that are more stable and reliable for operational use.

Overall, the revised statistical tests confirm that the reported model differences are robust and statistically well-supported. Moreover, GFRNet strikes a more prudent balance between TS and FSS, ensuring strong predictive accuracy while preserving spatial structure fidelity and reliability for real-world applications.

2.5 Transparency of Technical Details

Reviewer comment: The selection of hyperparameters (e.g., gradient penalty coefficient loss weights a/b) lacks a description of the systematic optimization process, with only grid search mentioned but no results provided.

Response: We thank the reviewer for this valuable comment on the transparency of technical details. The reviewer noted that our description of hyperparameters (e.g., λ, a/b) mentioned grid search but did not provide sufficient explanation of the optimization process or the selection criteria.

In response, we have made the following clarifications and revisions in the manuscript:

1. **Workflow and logic of hyperparameter optimization** For the loss function hyperparameters a/b, our goal was to balance model performance for moderate-to-heavy and extreme rainfall. We first performed a limited-range grid search on **FRNet** (the version of the model without GAN) and selected the parameter combination that achieved the best performance on moderate-to-heavy rainfall cases. Once a/b was established, we used FRNet as the **generator** structure, introduced the discriminator component, and began training GFRNet. Because GAN training is inherently less stable, we then tuned the gradient penalty coefficient λ by observing the convergence behavior of the model's loss curves and identifying values that ensured a **stable downward trend in the loss function**.

2. **Explanation of the selection criteria in the Methods section** We explicitly state that the choice of a/b was guided by the model's **TS and FSS performance on moderate-to-extreme rainfall thresholds**, while the choice of λ was driven by **the stability of GAN training**, ensuring that the generator–discriminator training converged reliably.

3. **Reason for not retaining detailed grid search records** These tuning experiments were designed to quickly narrow down reasonable parameter ranges, so we did not maintain exhaustive records of every parameter combination. However, we list the **final adopted values of a/b and λ** in the manuscript and explain the logic behind their selection.

4. **Plans for future improvement** In future work, we plan to adopt more systematic hyperparameter optimization approaches (e.g., Bayesian optimization or evolutionary algorithms) and document the tuning process more comprehensively to further enhance transparency and reproducibility.

Overall, we have **added clarifications in the main Methods section** describing the workflow and decision logic for selecting a/b and λ, acknowledged the current limitations, and committed to improving hyperparameter tuning documentation in future research.

2.6 Language and Expression Standardization

Reviewer comment: Some language expressions are not professional enough, and there is inconsistency in terminology and citation format, while the readability of charts and tables could be improved.

Response: We sincerely thank the reviewer for this careful observation and constructive suggestion. In the revised manuscript, we have conducted a thorough language review to ensure consistency and improve academic tone. All figure and table captions have been refined for clarity, terminology has been standardized, and several references have been updated and reformatted to align with the journal's style requirements. These revisions have enhanced the overall readability and professionalism of the manuscript.

2.7 Specificity of Future Work

Reviewer comment: The discussion mentions directions like "integration of physical constraints" and "higher resolution," but lacks specific implementation plans.

Response: We thank the reviewer for pointing out that the description of future work in the earlier version was not sufficiently specific. We fully agree that having a clear and actionable research plan is critical for the value of this study and for guiding subsequent work. In response to this comment, we have expanded and refined the *Discussion* section of the revised manuscript to include the following more targeted and technically feasible directions:

1. **Refining the evaluation framework and performance metrics.** We plan to develop a more comprehensive evaluation framework that not only relies on conventional metrics such as TS and FSS but also incorporates measures of spatial structure consistency, rainfall intensity distribution, and application-oriented indices. This will allow us to systematically assess model performance under different precipitation scenarios. Notably, our current study indicates that GFRNet performs better in *systematic rainfall events* (e.g., frontal rainbands, typhoon outer rain) but still has room for improvement in more localized and scattered precipitation. Future metric design will more clearly distinguish between these scenarios, helping to identify the model's strengths and weaknesses and provide feedback for further improvements.

2. **Addressing learning for scattered light rainfall events.** We observed that GANs and other deep generative models tend to exhibit *mode collapse* or neglect when handling *scattered, isolated light rainfall events*. We plan to design targeted data sampling strategies and loss function adjustments, as well as explore data augmentation techniques, to enhance the model's sensitivity and learning stability for these events.

3. **Enhancing the understanding of rainfall generation through physical variables.** Currently, GFRNet still relies to some extent on the precipitation forecasts provided by NWPs. Moving forward, we will introduce additional dynamic and thermodynamic variables (e.g., temperature, humidity, wind fields, geopotential height) as model inputs to help the model directly learn the *physical processes of rainfall generation*. This approach will reduce the "black-box" reliance on NWP rainfall fields and improve the model's *physical consistency* and its ability to generalize to complex meteorological conditions.

4. **Exploring more stable generative model architectures.** We will investigate *Diffusion Models*, *Conditioned Diffusion*, and hybrid *GAN–Diffusion* frameworks. By combining the efficient generation capabilities of GANs with the distribution-learning stability of Diffusion models, we aim to significantly improve the reconstruction of fine-scale rainfall structures and the robustness of predictions for extreme precipitation events.

5. **Improving training stability and data strategies.** GAN training can be sensitive to hyperparameters and data distribution, sometimes leading to instability. We will optimize *sampling strategies*, develop more effective learning schemes for long-tail distributions, and refine *loss function design* while improving training schedules and regularization. These efforts aim to ensure more stable learning across the full spectrum of rainfall intensities and deliver more consistent predictions.

We sincerely appreciate the reviewer's insightful suggestions, which have helped us refine our research roadmap and articulate clearer, more concrete directions for future work.

References

Ayzel, G., Heistermann, M., and Winterrath, T.: RainNet v1.0: a convolutional neural network for radar-based precipitation nowcasting, Geoscientific Model Development, 13, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020, 2020.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., kin Wong, W., and chun Woo, W.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, 2015.

Tan, J., Huang, Q., and Chen, S.: Deep learning model based on multi-scale feature fusion for precipitation nowcasting, Geoscientific Model Development, 17, 53–69, https://doi.org/10.5194/gmd-17-53-2024, 2024.

Yin, J., Gao, Z., and Han, W.: Application of a Radar Echo Extrapolation-Based Deep Learning Method in Strong Convection Nowcasting, Earth and Space Science, 8, e2020EA001 621, https://doi.org/10.1029/2020EA001621, 2021.