1 Comment and Response 1

The paper presents an application of deep learning techniques, specifically U-Net and GAN-based models, to enhance short-term precipitation forecasting, with a focus on the fine-scale structure of intense rainfall events. The authors compare the accuracy of three numerical weather prediction (NWP) models against two deep learning techniques: a U-Net-based model (FRNet) and a GAN-based model (GFRNet). They use multiple evaluation metrics to assess the relative performances of these approaches. The goal is to evaluate the accuracy of predicting 3-hourly accumulated precipitation over the next 24 hours for a region in North China. The paper shows many metrics and concludes that GFRNet demonstrates significant operational value.

1.1 Question 1

The paper deals with an interesting topic, related to improving the forecasting of the fine structure of intense rainfall. Moreover, it makes use of some current tools in deep learning, which seem promising for future operational use. However, I have two main concerns about the paper. The first one is the rationale of the model itself, as I do not understand how the fine structure of rainfall may be solely explained with the additional information (on top of the NWPs) provided to the GAN. The second one is related to the experimental design and the fairness of the comparisons and analyses provided. In my opinion, the paper requires an improvement of the rationale and the experimental design, as well as additional analyses before being considered for publication in a scientific journal.

1.2 Answer 1

Thank you for your thoughtful comments and constructive feedback on our paper. Regarding your concern about how the GAN-based model improves the fine structure of precipitation forecasts:

1. As introduced in the Introduction and Method sections, the principle of GAN lies in the adversarial training between the generator and discriminator. The goal is to train the generator to produce outputs that are indistinguishable from real labels, as judged by the discriminator. This process effectively allows the generator to refine the predictions to a level that achieves "realistic" quality.

2. In the field of short-term precipitation forecasting, the scientific validity of GANs has been demonstrated in two prominent works published in Nature (Ravuri et al., 2021; Zhang et al., 2023). These studies highlight GANs' ability to capture complex spatial and temporal structures, making them a promising tool for fine-scale rainfall prediction. Building on these findings, our work explores the application of GANs specifically for short-term precipitation forecasting.

I hope this addresses your concern. Please let me know if you would like further clarification or additional details.

1.3 Question 2

Below, I provide more details about the issues that I observe with the work.

My main concern with the rationale of the paper is that it is not immediately evident how the information about Elevation, Latitude, Longitude, Cycle, and Lead Hour contributes independently to improving forecasts when much of this information may already be embedded in the NWPs. A mechanism should be presented or outlined to justify the gains in accuracy. If none exists, then all the required information about the fine structure of precipitation is already included in the original NWPs, and thus the methods presented are just extracting this information.

1.4 Answer 2

Thank you for your insightful comments regarding the rationale of our paper. Below, we address your concerns regarding the contributions of elevation, latitude, longitude, cycle, and lead hour information to improving forecasts.

While it is true that NWPs already encapsulate fine-scale precipitation structures, each individual NWP model exhibits specific strengths and weaknesses. The goal of our deep learning model is to dynamically integrate these complementary features in a way that achieves a synergistic effect, effectively making $1 + 1 > 1$.

We fully agree with your view that directly analyzing the contribution of these sources and features to precipitation correction is essential and can provide us with significant insights. Through a series of ablation experiments, we systematically evaluated the impact of each feature and input on the precipitation forecasting performance of the GFRNet model. These experiments were designed to verify the contribution of key model components (such as ECMWF, CMA-SH9, CMA-3KM, META features, and temporal features) and quantify their impact through a Relative Importance Score (RIS). The results not only validated the rationality of GFRNet's design but also offered valuable insights for future research.

$$\text{Relative Importance}(x) = \frac{\text{TS(GFRNet)} - \text{TS(GFRNT\_wo\_x)}}{\text{TS(GFRNet)}} \tag{1}$$

- Model Stability and Correction Effect
  - Stability: Ablation experiments show that even when any single feature is removed, GFRNet consistently outperforms the three NWPs in moderate, heavy, and storm precipitation forecasting and generally performs better than the NWPs themselves in precipitation forecasting. This indicates that GFRNet has high stability and can provide positive correction effects even in the absence of certain data sources.
  - Correction Effect: By analyzing the Relative Importance Score (RIS) of each feature, we found that, except for META and temporal features having a slight negative impact on light rain forecasting, all features contribute positively to the correction across all precipitation levels.
- Feature Contribution Analysis
  - ECMWF Input: Makes a significant contribution to moderate and heavy rain forecasting but has less impact on light rain and storm forecasting. As a global high-resolution model, ECMWF outperforms other models in mesoscale precipitation events. Its advanced physical process simulation capabilities (such as cloud physics and boundary layer processing) enable it to excel in moderate and heavy rain forecasting.
  - CMA-3KM Input: Contributes significantly to moderate, heavy, and storm precipitation forecasting, particularly excelling in moderate and heavy rain forecasting. As a high-resolution regional model, CMA-3KM can capture finer precipitation structures and evolutionary processes, especially when dealing with moderate and heavy rain events, where its ability to simulate local convection and precipitation processes is stronger.
  - CMA-SH9 Input: Contributes to some extent to moderate and heavy rain forecasting but has less impact on light rain and storm forecasting. Compared to CMA-3KM, the resolution and physical process simulation capabilities of CMA-SH9 may be less refined, especially when dealing with the complex structures of moderate and heavy rain events, resulting in a lower contribution.
  - META and Temporal Features: Significantly contribute to moderate, heavy, and storm precipitation forecasting but may introduce noise in light rain forecasting. Heavy and storm precipitation events typically have clearer structures and more intense variations, which are more correlated with topography and temporal features, thus helping the model better capture these changes.

GFRNet effectively enhances the accuracy and resolution of precipitation forecasting by integrating multiple NWP models and featurs such as topography and time embeddings. The results of the ablation experiments show that GFRNet has significant advantages in moderate, heavy, and storm precipitation forecasting and demonstrates high model stability. Future work will further optimize feature selection and model architecture to improve the model's forecasting accuracy and generalization ability.

We hope this explanation provides clarity. Please let us know if additional details or analyses are required.

We analyzed whether GFRNet's predictions align with meteorological principles, particularly the orographic precipitation patterns near the Taihang Mountains.

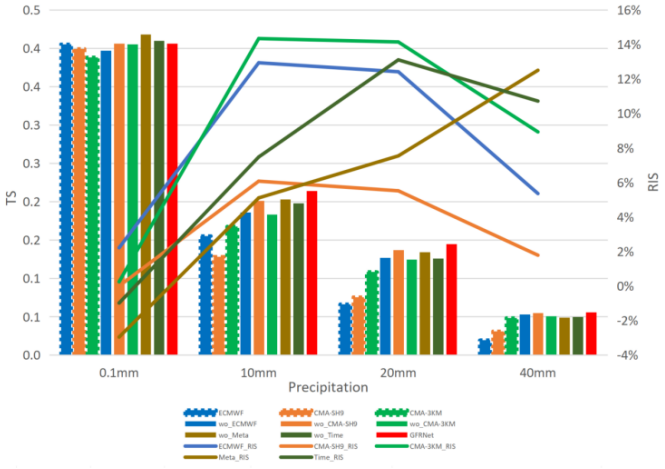

Figure 1. Model performance and IRS of features on ablation experiments

Table 1. TS scores for different rain thresholds in features ablation experiments

Model/Threshold	TS_0.1	TS_10	TS_20	TS_40
wo_ECMWF	0.397	0.186	0.127	0.053
wo_CMA-SH9	0.406	0.201	0.137	0.055
wo_CMA-3KM	0.405	0.183	0.124	0.051
wo_META	**0.418**	0.203	0.134	0.049
wo_Time	0.410	0.198	0.126	0.050
GFRNet	0.406	**0.214**	**0.145**	**0.056**

– Ablation Experiments: These show that topographic features significantly boost GFRNet's performance. Specifically, GFRNet with topographic features outperforms variants without them in multiple precipitation intensity thresholds, with TS improvements of about 12.41% at 20 mm and 5.36% at 40 mm. This indicates GFRNet understands the relationship between topography and precipitation, enhancing NWP outputs through terrain-aware fusion.

– BIAS Spatial Distribution: CMA-SH9 and CMA-3KM show significant over-forecasting (BIAS > 2) west of the Taihang Mountains. In contrast, GFRNet maintains a BIAS close to 1 in these areas, demonstrating its effective correction of NWP biases in complex terrains.

In summary, GFRNet excels in both statistical metrics and physical interpretability, with advantages in complex terrains. We'll keep enhancing the model by integrating more physical knowledge.

1.5 Question 3

If this is the case, as I believe (although I may be wrong), there may be other alternative methods that could improve forecasting with reduced complexity. To verify this point, I suggest that the authors include additional models, such as SVMs or Random Forests, to test if simpler bagging methods with far fewer parameters could also improve forecasting accuracy. In my experience, basic machine learning methods tend to perform similarly to deep learning ones (in this kind of application) at a significantly reduced level of complexity. Including these additional comparisons may serve to justify whether a GAN is an optimal strategy and to show if the improvement in forecasting accuracy comes from the deep learning techniques or from the combination of different sources of information.

Table 2. RIS for different rain thresholds in features ablation experiments

Model/Threshold	0.1 mm	10 mm	20 mm	40 mm
ECMWF	**2.22%**	12.93%	12.41%	5.36%
CMA-SH9	0%	6.07%	5.52%	1.79%
CMA-3KM	0.25%	**14.34%**	**14.14%**	8.93%
META	-2.96%	5.11%	7.54%	**12.50%**
Time	-0.99%	7.48%	13.10%	10.71%

Using FRNet as a benchmark may not be entirely fair, as GFRNet is essentially an enhanced version of the same model with a more advanced training procedure. It would also be interesting to report training and inference times for all models used (I know that the authors have included part of this information in their manuscript).

Thus, with respect to my concern with the rationale of the paper, the authors should include additional simpler models to check if GANs are justified for their complexity level or if other simpler methods may work similarly. Additionally, they should provide some insight into why the improvement occurs. This point leads to my concerns with the experimental design.

1.6 Answer 3

Thank you for your detailed feedback and suggestions regarding the experimental design and the rationale for using GANs in our study. Below, we address your concerns point by point:

1. **Comparison with Simpler Models**. Numerous existing studies have applied machine learning models, such as SVMs or Random Forests, to precipitation correction tasks. These methods, with careful design and parameter tuning, have been shown to outperform NWPs in accuracy. We acknowledge that simpler models with reduced complexity and computational efficiency could achieve high accuracy. However, a common limitation across these non-generative models is the blurring effect in their predictions, as highlighted in prior works.

2. **Rationale for Using GANs**.The primary motivation of our work is to explore whether adopting a generative approach, such as GANs, can address this blurring issue without compromising accuracy. By introducing adversarial training, we aim to produce sharper and more realistic precipitation predictions, especially for fine-scale structures.

3. **Fair Comparison Between Paradigms**.To ensure a fair comparison between non-generative and generative paradigms, we used the same model architecture (FRNet) for the generator component of the GAN. This approach isolates the impact of the adversarial training process, providing a controlled evaluation of the benefits of generative modeling over traditional methods.

4. **Combining Information Sources.** While the quality and diversity of data sources play a critical role, effectively combining these inputs requires leveraging advanced AI techniques. Our ablation studies have demonstrated that the deep learning model in our study successfully integrates these features to achieve positive correction results. It is important to note that the performance of these post-processing models inherently depends on the quality of NWPs. Thus, advancements in both NWPs and deep learning techniques can collectively enhance post-processing outcomes.

We appreciate your suggestion to include simpler models for comparison. While this would undoubtedly add value, the primary focus of our study is to address the limitations of non-generative models and evaluate the advantages of GAN-based techniques. Please let us know if additional analyses or clarifications are required.

1.7 Question 4

My first concern with the experimental design is related to the limited amount of data used in the study. Using only four years of data may not be sufficient to fully verify the accuracy and robustness of the forecasting method. I understand that data

limitations are difficult to overcome, but given the complexity of the models used, it is difficult to ensure that overfitting is not playing a role in the analysis. This concern is exacerbated by the fact that the original time series must be split into training, validation, and test sets.

Moreover, the initial data selection may bias the results. NWPs provide continuous forecasts, so alternative methods should ideally also deliver continuous predictions to ensure a fair comparison. If a heavy data selection procedure is implemented, the comparison may not be entirely fair to the NWPs.

1.8 Answer 4

Thank you for raising this important concern regarding the amount of data and the potential for overfitting in our study. We acknowledge the limitations inherent in using four years of data and have taken steps to ensure the robustness and fairness of our experimental design. Below, we address your points in detail:

1. **Data Limitations and Generalization**.We agree that collecting and preparing sufficient data for deep learning models is challenging and resource-intensive. It is indeed difficult to guarantee that the model will maintain strong generalization capability on entirely new datasets, as patterns in new data may change. This is a common challenge faced by all AI methods.

2. **Principles for Dataset Partitioning**.Given the constraints of limited data, it is essential to follow strict principles to ensure that the model's performance is both accurate and robust: a.**There should be no data leakage** between the training, validation, and test sets. b.**The test set should be unbiased**, representing real-world scenarios as comprehensively as possible.

3. **Our Dataset Partitioning and Testing Strategy**. **a. No Data Leakage:** As shown in Table 3, our dataset was partitioned chronologically, ensuring no overlap between training, validation, and test sets, thereby eliminating the risk of data leakage. **b. Unbiased and Realistic Test Set**: While we employed a sampling strategy on the training set to address data imbalance, the validation and test sets consisted of continuous, unaltered natural samples. Specifically, the test set included 77 consecutive days of real-world summer weather samples, providing an unbiased representation of the study region's conditions. **c. Validation of Generalization**: The stopping criterion for model training was defined by the minimum loss observed on the validation set, with no further improvement for 20 consecutive epochs. Good performance on the validation set is, therefore, not unexpected. However, if the model performs well on the validation set but poorly on the test set, it would indicate overfitting and poor generalization.

In our study, all evaluation results were conducted on the independent test set, and the model demonstrated strong performance. This suggests that, within the limitations of the available data, the model possesses good generalization capability on unseen data.

Table 3. Sample distribution across training, validation, and test sets.

Dataset	Time Period	Samples	
		Pre-sampling	Post-sampling
Training set	2019-06-01 - 2019-10-10 2020-06-01 - 2020-10-10 2021-03-15 - 2021-07-09 2021-08-21 - 2021-10-10 2022-03-15 - 2022-06-14	4645	2885
Validation set	2021-07-10 - 2021-08-20	637	No sampling
Test set	2022-06-16 - 2022-08-31	1204	No sampling

1.9 Question 5

A second concern with the experimental design is that I would have appreciated a clearer discussion in the methods' section about how the evaluation of accuracy was performed. Table 3 presents evaluation metrics and highlights the best and second-best performers. However, more attention should be paid to the differences. Are they significantly different? Or could all the methods (NWPs and nets) perform similarly given the amount of information used? What would be the expected distribution of the accuracy metrics? I am not fully convinced that part of the results are not an analysis of statistical fluctuations. Additionally, Table 4 seems to contradict the abstract, which states that "GFRNet outperforms all models in terms of Root Mean Square Error (RMSE)," but I may have missed something.

1.10 Answer 5

Thank you for your detailed review and insightful questions. Below, I address your concerns regarding the evaluation metrics, their significance, and the apparent discrepancy in Table 4.

Evaluation Metrics and Fairness. For short-term hourly precipitation forecasting, metrics such as TS, FAR, BIAS, and FSS are widely recognized and commonly used evaluation standards. In this study, these metrics were computed using standard statistical formulas and were evaluated under the same spatial and temporal resolutions to ensure fairness across all methods.

Significance of Differences. To determine whether the observed differences are meaningful, two key questions need to be addressed:

1. Are the differences statistically significant?
2. If so, how can we confirm that these differences are not due to statistical fluctuations but instead reflect the true performance of the models?

Model Performance Gains. Considering recent advancements in NWPs, achieving improvements in hourly short-term forecasts for heavy or torrential rainfall is exceptionally challenging. The TS and FSS improvements achieved by NETs (deep learning models) over the best-performing NWPs were approximately 20% for moderate, heavy, and torrential rainfall. Such improvements are statistically significant and represent meaningful performance gains.

Temporal and Spatial Analyses. We also analyzed the differences between NETs and NWPs in more detail from the two dimensions of time and space.

1. **Temporal Analysis:** As shown in Figure 4, we analyzed the TS and FSS scores of NWPs and NETs across different forecast lead times. The results demonstrate stable performance for all five models, particularly for thresholds of 10 mm and 20 mm. GFRNet and FRNet consistently showed significant advantages over NWPs.
2. **Spatial Analysis:** Figures 5 and 6 illustrate that GFRNet outperformed NWPs in most forecast regions, highlighting its spatial robustness. Additionally, since NETs are post-processed from NWPs, the performance trends of NETs and NWPs over forecast lead times exhibit similar patterns.

Clarification on Table 4 and Abstract.Thank you for pointing out the discrepancy regarding RMSE. You are correct that GFRNet does not achieve the lowest RMSE; ECMWF has the lowest RMSE. However, ECMWF's low RMSE is primarily due to its conservative predictions for moderate and heavy rainfall, which lack forecasting skill. In contrast, GFRNet achieves the second-lowest RMSE while maintaining high forecast skill, demonstrating a balance between accuracy and robustness. We will revise the abstract to accurately reflect this point.

1.11 Question 6

A third concern is related to the case studies, which I believe should be justified and presented in a more detailed way. This point may be related to my concern about the rationale of the paper: if a mechanism by which the GAN strategy improves the forecast is provided, then the case studies may focus on clear examples of this mechanism at play. Without this, I believe a general statistical evaluation would provide a clearer representation of the model's advantages. A detailed analysis of specific situations may not be as illuminating.

1.12 Answer 6

Thank you for your thoughtful comments regarding the case studies. We acknowledge the importance of a well-justified and detailed presentation of case studies, particularly in demonstrating the mechanisms by which the GAN strategy improves forecasts.

While general statistical evaluations provide an overall representation of model performance, they may obscure certain limitations or strengths of the models. Based on suggestions from meteorological experts, we selected three distinct precipitation events characterized by different dynamic and thermodynamic conditions. By visualizing these cases and providing detailed case-specific scores, we aimed to illustrate the stability of NETs, particularly GFRNet, across various precipitation scenarios.

Moreover, compared to FRNet, which tends to produce blurred predictions, GFRNet consistently delivers forecasts with clearer precipitation structures in these case studies. This highlights the GAN-based model's ability to address the blurring issue and capture fine-scale precipitation details more effectively.

We believe these case studies complement the statistical analysis by providing deeper insights into the model's performance under diverse conditions.

1.13 Question 7

A fourth concern about the experimental design is related to the selection of three NWPs. If two models similar to ECMWF were available, would it make sense to include both? How would the results change? This raises questions about generalizability. In many machine learning applications, the data exert a closer control on accuracy than the algorithms themselves. I understand that a paper cannot address every concern, but some guidance from the authors would be appreciated.

1.14 Answer 7

Thank you for raising this insightful question. The selection of NWPs and its impact on generalizability is indeed a critical consideration.

From the principles of ensemble forecasting design and prior studies on multi-model post-processing, an important guideline for selecting ensemble members is that they should be "**high-quality and diverse.**" This ensures that the input information provided to the correction model is both accurate and comprehensive, maximizing the potential for improved forecast performance. If two models similar to ECMWF were included, they would likely offer redundant information, providing little additional benefit beyond what a single ECMWF model could contribute. Consequently, the improvements in forecast accuracy might be limited.

The results of the source ablation experiments above also confirm one point: both CMA-3KM and CMA-SH9 are regional numerical models, and when CMA-3KM has already been added, the model gains from adding CMA-SH are not as apparent.

Regarding the relationship between data and algorithms, we share your perspective that data defines the upper limit of performance, while the iterative improvement of models seeks to approach this limit. For smaller datasets, simpler models may suffice, whereas larger datasets often require more complex models to fully exploit the information available. The development of both data and models should ideally evolve in tandem to achieve optimal results.

We hope this addresses your concern. Please let us know if further clarification or additional discussion would be helpful.

1.15 Question 8

My final concern relates to the generality of the conclusions and the reproducibility of the results in other locations. How robust are the results to the data selection procedure or the structure of the ANN? Would the same structure work well in other locations, or would changes be required? If a less intense data selection procedure were used, how would the results change? If 20 years of data were available, would GFRNet perform similarly? I believe the study would be much more robust if extended to other regions with more data available. Currently, the method seems to work, but the evidence may not yet be robust enough to fully support the claims made in the paper.

1.16 Answer 8

Thank you for your thoughtful comments and for raising concerns about the generality and reproducibility of the study's conclusions. These are critical points that deserve careful discussion.

Currently, due to the substantial effort required for data collection and preprocessing, we have not yet validated the generalization capability of the model trained in the North China region to other geographic locations. Precipitation patterns are indeed highly region-specific, influenced by local geography, climate, and dynamics.

That said, one of the key contributions of our study lies in proposing and demonstrating a methodology: **using generative deep learning models to improve both the accuracy and fine-scale structure of precipitation forecasts**. This approach, while tailored to the North China region, is intended to serve as a guiding framework. With this methodology, similar models could be trained or fine-tuned for other regions or extended datasets, such as longer time periods or more extensive geographic areas. We are optimistic that this approach would yield comparable results, though further studies are needed to verify this.

If more extensive datasets, such as 20 years of data, were available, we believe GFRNet would continue to perform well. A larger dataset could enable the model to better capture long-term patterns and variability, potentially improving its robustness and generalizability further.

We appreciate your suggestions and agree that extending this work to other regions with more data would significantly strengthen the study's conclusions. Such an extension is a valuable direction for future research.

1.17 Question 9

Finally, I present some comments about minor issues:

1. *References should be enclosed in parentheses. The way they are written now complicates the reading of the paper.*
2. *Figures 4 and 5 are difficult to interpret, particularly due to their complex visual layout. A more intuitive representation could enhance their clarity. For the maps, since topography seems to play such an important role, residuals might provide better insights.*
3. *Some references to equations are incomplete.*
4. *A better discussion on ensemble forecasts and deterministic quantitative forecasts may be in order. In my opinion, ensemble forecasts may convey a much better idea of severe storm potential, especially when combined with synthetic generation, so focusing on deterministic forecasts may be a disadvantage.*

1.18 Answer 9

Thank you for your detailed comments and suggestions regarding minor issues. Below, we address each of your points:

1. **References.** We appreciate your observation regarding the formatting of references. We will revise the manuscript to ensure that all references are properly enclosed in parentheses, improving the consistency and readability of the text.
2. **Figures 4 and 5**. Thank you for your comments on Figures 4 and 5. To address their complexity, we will simplify the visual layout and explore more intuitive representations. Indeed, residual maps can better illustrate the impact of topography on precipitation forecasts. As such, we have included the spatial distribution of residuals under the FSS metric in Figure 6, comparing GFRNet with NWPs and FRNet, to provide deeper insights into this effect.
3. **Equations.** Thank you for pointing out the incomplete references to equations. We will carefully review the manuscript to ensure all equation references are complete and formatted correctly.
4. **Discussion on Ensemble Forecasts.** We agree that ensemble forecasts have significant advantages in capturing the potential for severe weather events, particularly when combined with synthetic generation techniques. However, the primary focus of this study is to evaluate the capability of deterministic forecasts in predicting precipitation structures and fine-scale details. In future work, we plan to explore how ensemble forecasting techniques can complement or enhance our deterministic approach, especially for severe storm scenarios.

We greatly appreciate your thoughtful feedback and will incorporate these suggestions to improve the manuscript. Please let us know if there are additional areas that require further attention.

2 Comment and Response 2

2.1 Review Comments

Title: Improving the fine structure of intense rainfall forecast by a designed adversarial generation network
Authors: Zuliang Fang, Qi Zhong, Haoming Chen, Xiuming Wang, Zhicha Zhang, and Hongli Liang
Submitted to GMD (open for interactive public discussion as preprint on EGUsphere

2.2 Recommendation

This manuscript proposes a Generative Adversarial Fusion Network (GFRNet) for short-term precipitation forecasting in North China, aiming to improve the accuracy of 3-h accumulated precipitation predictions over a 24-h period. The study optimizes data sampling strategies, loss functions, and model architecture, demonstrating GFRNet's superiority over numerical weather prediction (NWP) models in metrics such as TS, FSS, and RMSE. The paper is well-structured, with a logical experimental design and thorough result analysis, showing both innovation and practical value.

However, limitations remain in model generalization analysis, physical interpretability, and stability in extreme precipitation forecasting, which require further discussion and improvement, and methodological details (e.g.,hyperparameter selection) should be expanded.

The English of the paper is generally good enough to be understood, but with some polishing, the paper could achieve more professional, natural-sounding academic English. The technical content is clear, but the language could be more precise in places. None of the issues seriously impede understanding but correcting them would elevate the paper's professionalism. I would recommend having the English grammar professionally checked by a specialized editing service.

The paper is recommended for publication after major revisions, and future work could explore advanced architectures and broader applications.

2.3 General Comments

I suggest ablation studies or explainability analyses (e.g., SHAP, DeepLIFT, or similar tools) are needed for the AI-based precipitation forecasting paper.

2.3.1 Question 1: blation Studies (Highly Recommended)

- Purpose: Validate the necessity of GFRNet's key components (GAN strategy, SE blocks, weighted loss).
- Suggested Tests:
  - Remove GAN discriminator (compare with FRNet results).
  - Ablate SE attention blocks.
  - Test without weighted loss (standard MSE/MAE).
- Justification: The paper claims GANs address blurring, but quantitative evidence is needed to isolate its contribution versus other components.

2.3.2 Answer 1

Thank you very much for your insightful comments and suggestions. Below is our response to your suggestions regarding the ablation study of GFRNet's key components:

Complexity of Deep Learning Models in Precipitation Correction:

1. **Importance of Model Design and Parameter Tuning**: As you rightly pointed out, achieving effective precipitation correction using deep learning requires careful model design, loss function selection, and meticulous parameter tuning. This is especially true for short-term heavy precipitation events exceeding 40 mm. In this study, the unique loss function design and model architecture of FRNet have played a crucial role in enhancing the positive correction of heavy and strong precipitation.

2. L**imitations of TS Improvement**: While FRNet shows some improvement in TS scores, our in-depth analysis reveals that, similar to many deep learning correction methods, this improvement often stems from over-forecasting and blurry forecasts. This can significantly diminish the model's practical utility. Hence, we introduced the GFRNet framework, which ensures clear and detailed precipitation structures while maintaining correction effectiveness, making it a truly operational high-quality product.

3. **Balancing Role of GAN Strategy**: It is important to note that while the SE architecture and unique loss function design enhance model accuracy, they also introduce the issue of blurry forecasts. The GAN strategy does not directly boost precipitation accuracy (TS) and even involves a certain sacrifice in accuracy compared to FRNet. However, its strength lies in achieving a balanced trade-off between accuracy and practicality (clear structural forecasts) through adversarial training, which is best reflected in the FSS indicator.

We conducted ablation experiments on GFRNet by removing SE attention and using standard MSE/MAE loss. The results are summarized in the table below:

1. **Impact of SE Blocks**: The TS values of GFRNet and GFRNet without SE are similar in light rain forecasting (0.406 vs. 0.408), indicating a minor impact of SE blocks on light rain. However, for thresholds of medium rain and above, GFRNet significantly outperforms GFRNet without SE. For instance, at the 20 mm and 40 mm thresholds, the TS values of GFRNet are 0.145 and 0.056, compared to 0.134 and 0.052 for GFRNet without SE. This highlights the critical role of SE blocks in capturing the detailed structure of heavy precipitation events.

2. **Impact of Weighted Loss Function**: In light rain forecasting, GFRNet without Weighted Loss shows higher TS values than GFRNet (0.431 vs. 0.406), suggesting better performance of standard MSE/MAE loss in this scenario. Nevertheless, for thresholds of medium rain and above, GFRNet significantly surpasses GFRNet without Weighted Loss. For example, at the 20 mm and 40 mm thresholds, the TS values of GFRNet are 0.145 and 0.056, while those of GFRNet without Weighted Loss are 0.115 and 0.028. This underscores the superiority of the weighted loss function in effectively enhancing the model's ability to forecast heavy precipitation.

Table 4. TS scores for different rain thresholds in blocks ablation experiments

Model/Threshold	TS_0.1	TS_10	TS_20	TS_40
ECMWF	0.405	0.155	0.067	0.019
CMA-SH9	0.399	0.128	0.076	0.031
CMA-3KM	0.388	0.168	0.108	0.049
FRNet	0.416	0.216	0.147	0.077
GFRNet	0.406	0.214	0.145	0.056
GFRNet_wo_SE	0.408	0.195	0.134	0.052
GFRNet_wo_WeightedLoss	0.431	0.191	0.115	0.028

Due to space limitations, the detailed results and analysis of the ablation experiments are added to appedix. Once again, thank you for your valuable feedback. We will refine the manuscript further based on your suggestions.

2.3.3 Question 2: Explainability Methods (Conditionally Useful)

- SHAP/DeepLIFT Value: Limited for pure precipitation prediction, as:
  - Input variables are homogeneous (all are precipitation forecasts + topography/temporal features).
  - The model's nonlinear fusion process matters more than individual feature importance.
- Alternative Approaches:
  - Sensitivity Analysis: Perturb input NWP models (e.g., mask ECMWF/CMA-3KM inputs) to quantify their relative contributions.
  - Physical Interpretability: Analyze whether GFRNet's predictions align with meteorological principles (e.g., orographic precipitation patterns near Taihang Mountains).

2.3.4 Answer 2

Thank you very much for your valuable comments and suggestions. We fully agree with your view that directly analyzing the contribution of these sources and features to precipitation correction is essential and can provide us with significant insights. Through a series of ablation experiments, we systematically evaluated the impact of each feature and input on the precipitation forecasting performance of the GFRNet model. These experiments were designed to verify the contribution of key model components (such as ECMWF, CMA-SH9, CMA-3KM, META features, and temporal features) and quantify their impact through a Relative Importance Score (RIS). The results not only validated the rationality of GFRNet's design but also offered valuable insights for future research.

$$\text{Relative Importance}(x) = \frac{\text{TS(GFRNet)} - \text{TS(wo\_x)}}{\text{TS(GFRNet)}} \tag{2}$$

- Model Stability and Correction Effect
  - Stability: Ablation experiments show that even when any single feature is removed, GFRNet consistently outperforms the three NWPs in moderate, heavy, and storm precipitation forecasting and generally performs better than the NWPs themselves in precipitation forecasting. This indicates that GFRNet has high stability and can provide positive correction effects even in the absence of certain data sources.
  - Correction Effect: By analyzing the Relative Importance Score (RIS) of each feature, we found that, except for META and temporal features having a slight negative impact on light rain forecasting, all features contribute positively to the correction across all precipitation levels.
- Feature Contribution Analysis
  - ECMWF Input: Makes a significant contribution to moderate and heavy rain forecasting but has less impact on light rain and storm forecasting. As a global high-resolution model, ECMWF outperforms other models in mesoscale precipitation events. Its advanced physical process simulation capabilities (such as cloud physics and boundary layer processing) enable it to excel in moderate and heavy rain forecasting.
  - CMA-3KM Input: Contributes significantly to moderate, heavy, and storm precipitation forecasting, particularly excelling in moderate and heavy rain forecasting. As a high-resolution regional model, CMA-3KM can capture finer precipitation structures and evolutionary processes, especially when dealing with moderate and heavy rain events, where its ability to simulate local convection and precipitation processes is stronger.
  - CMA-SH9 Input: Contributes to some extent to moderate and heavy rain forecasting but has less impact on light rain and storm forecasting. Compared to CMA-3KM, the resolution and physical process simulation capabilities of CMA-SH9 may be less refined, especially when dealing with the complex structures of moderate and heavy rain events, resulting in a lower contribution.

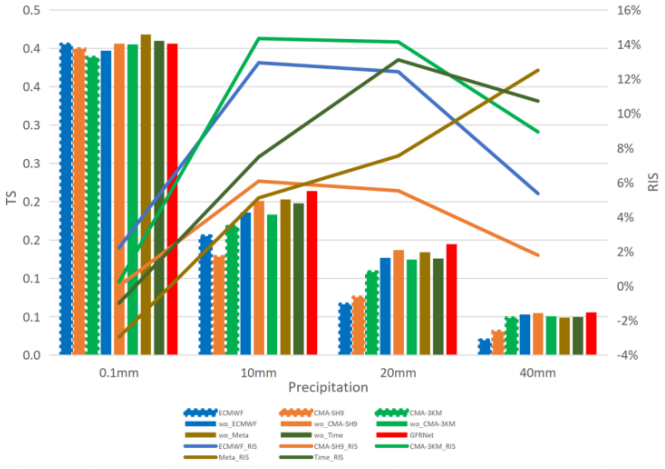

Figure 2. Model performance and IRS of features on ablation experiments

– META and Temporal Features: Significantly contribute to moderate, heavy, and storm precipitation forecasting but may introduce noise in light rain forecasting. Heavy and storm precipitation events typically have clearer structures and more intense variations, which are more correlated with topography and temporal features, thus helping the model better capture these changes.

GFRNet effectively enhances the accuracy and resolution of precipitation forecasting by integrating multiple NWP models and features. The results of the ablation experiments show that GFRNet has significant advantages in moderate, heavy, and storm precipitation forecasting and demonstrates high model stability. Future work will further optimize feature selection and model architecture to improve the model's forecasting accuracy and generalization ability.

Table 5. TS scores for different rain thresholds in features ablation experiments

Model/Threshold	TS_0.1	TS_10	TS_20	TS_40
wo_ECMWF	0.397	0.186	0.127	0.053
wo_CMA-SH9	0.406	0.201	0.137	0.055
wo_CMA-3KM	0.405	0.183	0.124	0.051
wo_META	**0.418**	0.203	0.134	0.049
wo_Time	0.410	0.198	0.126	0.050
GFRNet	0.406	**0.214**	**0.145**	**0.056**

Table 6. RIS for different rain thresholds in features ablation experiments

Model/Threshold	0.1 mm	10 mm	20 mm	40 mm
ECMWF	**2.22%**	12.93%	12.41%	5.36%
CMA-SH9	0%	6.07%	5.52%	1.79%
CMA-3KM	0.25%	**14.34%**	**14.14%**	8.93%
META	-2.96%	5.11%	7.54%	**12.50%**
Time	-0.99%	7.48%	13.10%	10.71%

We analyzed whether GFRNet's predictions align with meteorological principles, particularly the orographic precipitation patterns near the Taihang Mountains.

- Ablation Experiments: These show that topographic features significantly boost GFRNet's performance. Specifically, GFRNet with topographic features outperforms variants without them in multiple precipitation intensity thresholds, with TS improvements of about 12.41% at 20 mm and 5.36% at 40 mm. This indicates GFRNet understands the relationship between topography and precipitation, enhancing NWP outputs through terrain-aware fusion.

- BIAS Spatial Distribution: CMA-SH9 and CMA-3KM show significant over-forecasting (BIAS > 2) west of the Taihang Mountains. In contrast, GFRNet maintains a BIAS close to 1 in these areas, demonstrating its effective correction of NWP biases in complex terrains.

In summary, GFRNet excels in both statistical metrics and physical interpretability, with advantages in complex terrains. We have already update this part into the appendix. We'll keep enhancing the model by integrating more physical knowledge.

2.4 Specific Comments

2.4.1 Question 3: Methodology

- strengths
  - The GAN strategy (WGAN-GP) mitigates blurry predictions, producing more realistic precipitation structures and intensities.
  - The weighted loss function with exponential weighting enhances extreme precipitation learning.
- Suggestions for Improvement:
  - The choice of hyperparameters (e.g., gradient penalty coefficient γ, loss weights a and b) lacks justification (e.g., grid search or ablation studies). Sensitivity analysis should be added.
  - The generator (U-Net + SE block) and discriminator (DCGAN-based) architectures are conventional. Advanced generative models (e.g., Diffusion Models) or spatiotemporal attention mechanisms could further improve fine-scale precipitation capture. Some discussion is necessary.

2.4.2 Answer 3: Methodology

Thank you very much for your insightful comments and suggestions. Below is our response to your suggestions regarding hyperparameter selection and model architecture:

Response to Hyperparameter Selection

- Gradient Penalty Coefficient γ. We experimentally verified the impact of different gradient penalty coefficients on model performance. The results indicate that setting γ to 10 achieves the best balance between the generator and discriminator. This effectively prevents gradient vanishing and exploding, enhancing training stability and sample quality. This aligns with the theoretical support for the gradient penalty term in WGAN-GP, which ensures the discriminator's gradients approach 1, satisfying the Lipschitz continuity condition.
- Theoretical Basis and Reference.
  - Long-tailed Distribution Handling: Given the long-tailed nature of precipitation data, with light rain dominating and heavy rain being rare, assigning higher loss weights to rare events like heavy precipitation is a common and effective approach. This method is widely used in addressing class imbalance and has been validated in related studies.

- Experimental Validation and Grid Search: The specific weight parameters (a=4.3, b=0.8) were determined through extensive grid search experiments on the validation set. This combination showed the best TS performance across multiple precipitation thresholds, significantly improving the model's ability to forecast heavy precipitation while maintaining reasonable performance for light rain.

In summary, our weight design is well-founded theoretically and validated experimentally. Future work will explore more advanced loss function designs to further enhance model performance.

Response to Model Architecture We agree that advanced generative models (e.g., diffusion models) or spatiotemporal attention mechanisms could enhance performance. However, as this study aims to explore GAN strategies in precipitation forecasting, we used the mature U-Net and SE block architectures. Moving forward, we will investigate more advanced architectures, such as:

- Diffusion Models: For their excellence in generating high-quality images.
- Spatiotemporal Attention Mechanisms: To better capture the temporal and spatial evolution of precipitation events.

Thank you again for your feedback. We will refine the manuscript further based on your suggestions.

2.4.3 Question 4: Results and Analysis

- Strengths
  - Quantitative metrics (TS, FSS, RMSE) show GFRNet outperforms NWP models, particularly for heavy precipitation (e.g., significant TS improvement at 20 mm threshold).
  - Case studies demonstrate GFRNet's ability to capture precipitation band evolution and intensity changes.
- Suggestions for Improvement:
  - Some results are contradictory: e.g., GFRNet's TS for 40 mm (0.056) is lower than FRNet's (0.077). The authors should analyze whether this is due to GAN's conservative generation strategy (missed events).
  - Spatial analysis shows higher BIAS in mountainous regions (e.g., west of Taihang Mountains). Terrain effects on model performance should be further discussed (e.g., elevation-dependent constraints).

2.4.4 Answer 4

Thank you for your comments. Here's our response to the issues you raised:

TS Value Concern:

- The lower TS value of GFRNet (0.056) compared to FRNet (0.077) for 40mm precipitation might be due to GAN's conservative strategy. GFRNet, through adversarial training, aims to produce more realistic and detailed precipitation structures, which may lead to missed events and a lower TS. In contrast, FRNet might generate more intense forecasts without GAN's mechanism, potentially inflating TS but at the cost of structural accuracy.
- Notably, FRNet shows a high BIAS (over-forecasting) for 40mm precipitation, while GFRNet's BIAS is closer to 1, indicating a better balance between under- and over-forecasting. Moreover, GFRNet surpasses FRNet in the FSS, highlighting its superior performance in capturing precipitation's spatial structure. Despite potential pixel-wise misses, GFRNet better represents overall precipitation patterns, which is crucial for practical applications.
- Future work will focus on optimizing GFRNet for extreme precipitation events by refining the GAN architecture, adjusting loss functions, and incorporating additional meteorological features. We'll also explore more comprehensive evaluation metrics to better assess extreme event forecasting performance.

Impact of Terrain on Model Performance:

- We've observed that CMA-SH9 and CMA-3KM exhibit significant over-forecasting (BIAS > 2) in high-altitude areas west of the Taihang Mountains for heavy and storm precipitation. This implies terrain considerably influences these models' performance. The complex topography causes air uplift, increasing precipitation chances and intensity. However, NWP, especially mesoscale models, may have biases in simulating these effects, particularly in high-altitude regions.
- Interestingly, both FRNet and GFRNet effectively correct the high bias of regional numerical models in these areas. GFRNet shows a more uniform BIAS distribution across the entire region, with values mostly close to 1, indicating excellent correction of the models' biases. This underscores the ability of deep learning models, particularly GFRNet within the GAN framework, to effectively amend traditional numerical model biases. We plan to include these analyses in the manuscript.

Thank you for your feedback. We'll continue to enhance the manuscript based on your suggestions.

2.4.5 Question 5: Discussion and Future Work

- Strengths
  - Clear future directions (e.g., higher resolution, physics-informed learning, ensemble forecasting) are proposed.
- Suggestions for Improvement:
  - Limitations are under-discussed: e.g., GFRNet relies on multi-model inputs—how does it handle systematic biases in individual models (e.g., CMA-3KM)?
  - Computational costs (3-h training on A100 GPU) are not evaluated for operational feasibility. Real-time deployment constraints should be addressed.

2.4.6 Answer 5

Thank you for your comment.

Model Feature Preparation and Processing

- The model's input features encompass precipitation forecasts from multiple Numerical Weather Prediction (NWP) models, static topographic features, and time-encoded information.
- Processing static topographic features and time-encoded information involves no computational overhead.
- Extracting and processing precipitation forecasts from NWP into the input format required by the Deep Learning (DL) model incurs minimal computational overhead, contingent on the platform's data storage method and CPU performance. The data processing time for eight samples per cycle does not exceed 60 seconds.

Model Inference Time

- Inferring eight samples (hourly precipitation forecasts for the next 24 hours) on a GPU takes less than 1 second.
- Inferring the same samples on a CPU takes under 20 seconds.

Deployment and Real-Time Performance

- During model deployment, the optimal model can be selected based on real-time data source availability. Even when relying solely on a single NWP, the DL model demonstrates superior performance compared to the NWP itself.
- The sole constraints are the computational time required for NWP forecasts and the transmission time of forecast data to the platform. For instance, downstream platforms typically receive 24-hour forecast data from ECMWF, initialized at 00UTC, after 06UTC to 09UTC.

- The inference time for downstream users employing FRNet or GFRNet does not exceed 2 minutes, which is negligible compared to the computation and transmission time of NWP.
- Ablation experiments indicate that the model outperforms the NWP itself even when based on any single NWP. Thus, during deployment, the optimal model can be selected in real-time according to data source availability. Specifically, the platform can choose which model to use based on the time it takes to obtain forecasts from three NWPs, balancing speed and accuracy.

2.5 Question: Minor Issues

- The title emphasizes "intense rainfall," but the paper does not justify the 40 mm/3h threshold (is it a standard benchmark?).
- In Figure 6 (FSS spatial gain), white regions (FSS=0) are unexplained and could be misleading.
- Consistency Check:
  - Ensure consistent use of terms (e.g., "deep learning" vs. "Deep Learning").
  - Check all acronyms are defined at first use.
  - Verify all citations follow the same style.
- Line 2: "the accuracy of precipitation forecasts remains significantly inadequate" –> "the accuracy...remains inadequate" or "is significantly inadequate".
- Line 6: "based on the outputs of multiple numerical weather models" –>"based on outputs from multiple numerical weather models".
- Line 74: "We apply the GAN strategy in developing the GFRNet model" –>"We implement a GAN strategy to develop the GFRNet model".
- Line 81: "The target area features a complex topography" –> "The target area features complex topography".
- Line 85: "CMA Multi-source merged Precipitation Analysis System(CMPAS)" –> Needs spaces: "CMA Multi-source merged Precipitation Analysis System (CMPAS)"; Similarly, a space is required between the preceding English word and the opening parenthesis. Please check all parts of the paper.
- Line 101-102: "Let r3(T) denote the accumulated precipitation over the past 3 hours at time T, with the learning target being r3(T) from CMPAS." –> "Let r3(T) denote the 3-hour accumulated precipitation at time T, where the learning target is the corresponding CMPAS r3(T) observation."
- Line 143: "a U-Net with encoder-decoder architecture" –> "a U-Net with an encoder-decoder architecture".
- Table 1 Title: "Data Sources and Features Used in Model" –> "Data sources and features used in the model".

2.6 Answer: Minor Issues

Why we set 40 mm/3h as threshold as intense rainfall?

- Operational Requirements: In practical meteorological operations, precipitation events of 40 mm/3 hours are typically regarded as heavy rainfall events that require special attention. This threshold is chosen based on operational requirements and the practical experience of forecasters.
- Research Comparison: In related studies, different thresholds have been used to define heavy rainfall events. For example, (Zhou et al., 2022) and (Ravuri et al., 2021) used thresholds of 20 mm/3 hours and 5 mm/hour, respectively, in their studies. Our study selects a threshold of 40 mm/3 hours to more precisely focus on more destructive rainfall events.

- Data Support: By analyzing the distribution of precipitation intensity in the training data, we found that 40 mm/3-hour precipitation events are significantly representative in the dataset and are key targets for operational forecasting.

We have carefully revised the manuscript and addressed the improper expressions you mentioned. Please refer to the revised version for more details.

Thank you for your meticulous feedback on the writing of the article. This has been extremely helpful for future paper writing. We will address and revise the article as needed based on your suggestions. Thank you.

We greatly appreciate your thoughtful feedback and will incorporate these suggestions to improve the manuscript. Please let us know if there are additional areas that require further attention.

References

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., and Mohamed, S.: Skillful Precipitation Nowcasting Using Deep Generative Models of Radar, Nature, 597, 672–677, https://doi.org/10.1038/s41586-021-03854-z, 2021.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J.: Skilful Nowcasting of Extreme Precipitation with NowcastNet, Nature, pp. 1–7, https://doi.org/10.1038/s41586-023-06184-4, 2023.

Zhou, K., Sun, J., Zheng, Y., and Zhang, Y.: Quantitative Precipitation Forecast Experiment Based on Basic NWP Variables Using Deep Learning, ADVANCES IN ATMOSPHERIC SCIENCES, 39, 1472–1486, https://doi.org/10.1007/s00376-021-1207-7, 2022.