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 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.

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.

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.

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 1. 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 will not be included in the main text but will be considered for the appendix. Once again, thank you for your valuable feedback. We will refine the manuscript further based on your suggestions.

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).

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.

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

- 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.

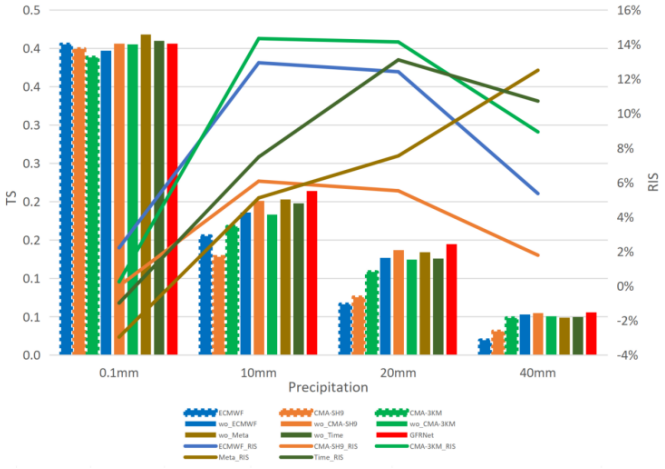

Figure 1. Model performance and IRS of features on ablation experiments

– 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 2. 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 3. 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'll keep enhancing the model by integrating more physical knowledge.

4 Specific Comments

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.

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.

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).

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.

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.

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.

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".

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.

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.

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.