Reply to the Reviewers

Re: Manuscript egusphere-2026-46

Precipitation Nowcasting Based on Convolutional LSTM with Spatio-Temporal Information Transformation Using Multi-Meteorological Factors

Dufu Liu, Feihu Huang, Peng Zheng, Xiaomeng Huang, Xi Wu, Xia Yuan, Jiafeng Zheng, Xiaojie Li, and Jing Hu*

To chief Editor, Editor and Reviewers,

We are pleased to resubmit our revised manuscript, titled “Precipitation Nowcasting Based on Convolutional LSTM with Spatio-Temporal Information Transformation Using Multi-Meteorological Factors,” for further consideration by the journal. We extend our sincere gratitude to the editor and reviewers for their valuable time, insightful comments, and constructive suggestions, which have significantly enhanced the quality of our work.

In this revision, we have systematically addressed all the reviewers' comments and comprehensively optimized the paper's clarity, rigor, and completeness. We have refined the descriptions of the methodology, added explanations where necessary, and improved the structure of the results presentation. Additionally, we have corrected the grammar and enhanced the overall readability of the text.

We believe that the revised manuscript now more clearly communicates our contributions and meets the high standards expected by your journal. We respectfully submit it for your further evaluation.

Yours sincerely,
Dufu Liu, Jing Hu

Chief Editor

Main Judgment #1

Dear authors and Dr. Fita,

I would like to note that the type for this manuscript, currently labelled as "Model experiment description paper", does not correspond to the presented work, and

therefore it should be modified and labelled as a "Model description paper". Please, do it accordingly during the next steps of the review process.

Juan A. Añel

Geosci. Model Dev. Executive Editor

Our response #1

Dear Dr. Juan A. Añel,
We sincerely thank the Executive Editor for this critical and professional correction. We have fully followed your official instruction, and accurately modified the manuscript type from the original "Model experiment description paper" to "Model description paper" in the revised manuscript submission system.
We confirm that this revision will be completed in strict accordance with your requirements. We will fully cooperate with all the subsequent procedures of the review process, and thank you again for your time and rigorous guidance on our manuscript.
Sincerely yours,
Dufu Liu

1. Referee 1

Main Judgment

The manuscript addresses an important nowcasting problem and reports promising results. However, several central claims require clearer framing, better metric justification, and stronger experimental documentation.
The topic is relevant to short-term weather nowcasting. The experiments are substantial, with comparisons against optical-flow, RNN/CNN, Transformer, and generative baselines. The multimodal SEVIR setup is valuable. The ablations on meteorological variables, the STI-related framework, and the loss function are useful. The results suggest improved short-term skill for moderate-to-heavy VIL structures.

Our response

Dear Reviewer,

We would like to express our sincere gratitude to the Editor and the Reviewers for dedicating their time to evaluating our manuscript. We deeply appreciate the Reviewer's encouraging summary, acknowledging the relevance of our multimodal SEVIR setup and the value of our ablations. The constructive and meticulous feedback provided has been immensely helpful in improving the rigor, clarity, and overall quality of our paper.

We have carefully considered all the comments and revised the manuscript accordingly.

Below, we provide a point-by-point response to each of the Reviewer's concerns.

Best regards

Reviewer #1, comment #1

1.The manuseript invokes delay embedding, STI equations, and conjugate duality, but the implementation appears to be a dual encoder-decoder with a consistency relation CTS (CST (z)=z. This may be useful STI-motivated regularization, and Table 6 provides relevant empirical support. However, the current manuscript should not imply that the network solves STI equations. Please reframe the method as STI-inspired/STI-motivated dual learning and clearly separate the mathematical motivation from the learned implementation.

Our response #1

We sincerely thank the reviewer for this mathematically rigorous observation. We

completely agree that our neural network acts as a data-driven approximator rather than an explicit numerical solver for the STI equations. Following your highly constructive suggestion, we have reframed our methodology throughout the manuscript. We now accurately describe our model as an "STI-inspired dual learning framework" that "implicitly approximates spatiotemporal mappings," ensuring a clear boundary between the mathematical motivation and the neural implementation.

To clarify our original thought process: because deep learning fundamentally operates as an implicit function-fitting mechanism, we utilized the model to approximate the STI mapping function itself. We originally viewed this approximation process as implicitly "solving" the conjugate STI function. However, we concede that the term "solve" can be mathematically misleading in this context, and your suggested phrasing is much more accurate and appropriate for a deep learning audience.

We have revised the manuscript, and the specific changes are as follows:

- **Abstract, Page 1:** In response, the present work proposes a dual encoder-decoder training framework based on the STI equation and the idea of dual learning, which can map multidimensional spatial features to the temporal prediction of future precipitation variables.
- **Introduction, Page 4:** To bridge this gap, we propose a Spatio-Temporal Information-Dual Encoder-Decoder Network (STI-DEDN) that enables a more accurate approximate of STI equations inspired by the inherent conjugate duality of the STI equations.

Reviewer #1, comment #2

2.The forecast target is VIL, not direct surface precipitation. VIL is a useful radar-derived proxy for convective intensity, and using VIL is fair within SEVIR because all methods are evaluated on the same target. However, the manuscript should consistently describe the task as VIL nowcasting, not rainfall-rate or surface-precipitation forecasting. Converting SEVIR values to VIL units does not validate the model against rainfall rate or accumulation.

Our response #2

We appreciate the reviewer pointing out this critical distinction. We fully acknowledge that VIL is a radar-derived proxy for convective intensity and strictly differs from direct surface precipitation or rainfall rates. We have thoroughly revised the manuscript to consistently describe our specific task as "VIL nowcasting" and removed any misleading terminology regarding surface precipitation.

Furthermore, to address the underlying concern regarding the real-world applicability

of our method to actual rainfall, we have proactively conducted an additional validation experiment. We trained our model on an independent, real-world precipitation dataset and compared it against traditional forecasting methods. The results demonstrate that our method maintains its effectiveness and outperforms traditional approaches in predicting true rainfall volumes. We have added this supplementary experiment to the manuscript and explicitly outlined the expansion to broader real-precipitation datasets as a key focus for our future work.

We have revised the manuscript, and the specific changes are as follows:

- **Introduction, Page 4:** In addition, due to the particularity of VIL variables in the experimental dataset, all experiments in our paper are VIL intensity forecasts. However, this does not mean that our method cannot achieve excellent performance on real precipitation data. To validate this, we conducted additional experiments on the MeteoNet dataset containing real precipitation measurements. The results demonstrate that our method also achieves promising performance on real-world precipitation data. Consequently, in subsequent chapters, we use precipitation forecasts as an approximation for VIL forecasting.
- **Introduction, Page 4:** Through comprehensive benchmarking experiments against state-of-the-art methods, we demonstrate a superior performance and generalizability of our framework. The experimental results show that the proposed method has more practical significance for VIL precipitation forecasting in medium and heavy VIL regions.
-

Reviewer #1, comment #3

3.The model appears deterministic, yet CRPS and BSS are reported. Please either:
a. clearly define the forecast distribution or sampling procedure used to compute these metrics,
b. or if they are computed from deterministic point forecasts, state this explicitly and do not interpret them as uncertainty or probabilistic-skill metrics,
c. or remove/demote them from the main results.
In the current version, these metrics are not sufficiently justified.

Our response #3

We appreciate the reviewer's strict assessment of our evaluation metrics. You are entirely correct that our model generates deterministic point forecasts, making the interpretation of continuous probabilistic scores like CRPS and BSS mathematically unjustified in this context. Following your recommendation, we have completely removed the CRPS and BSS metrics from the tables and the main text to maintain the absolute rigor of our deterministic evaluation.

We have revised the manuscript, and the specific changes are as follows:

- **Evaluation Metrics, Page 13:** ~~The continuous ranked probability score (CRPS) \citep{gong2024easecast} serves as another important metric, enabling us to estimate the model's capability to accurately represent the uncertainty associated with precipitation predictions. This is achieved by calculating the cumulative error between the distribution of the predicted image and the distribution of the actual image~~.
- **Experimental Results and Analysis, Page 16:** ~~Furthermore, to validate the efficacy of our method in efficient uncertainty modeling, we employed the CRPS metric to quantify the discrepancy between predicted probability distributions from different methods and the actual distributions of observed images. As illustrated in the Table \ref{tab:comparison}, the STI-DEDN model demonstrated a 9.39$\%$ improvement in CRPS compared to the worst-performing model, and a 4.06$\%$ enhancement in CRPS relative to the second-best PredRNN model. This demonstrates that precipitation predictions produced by the STI-DEDN model display probability distributions closest to observational data indicating enhanced proficiency in uncertainty modeling~~.
- **Case 2: Evaluation of Extreme Precipitation Events, Page 21:** ~~Since case 2 is an extreme precipitation event under complex weather conditions, which has the characteristics of small range and large local precipitation, we introduce the Brier Skill Score (BSS) \citep{choi2022rad} to assess the calibration degree of the predicted probabilities for a weather event's accurate forecast. In case 1, we did not conduct similar experiments due to the nature of the event, which involved typical ordinary rainfall. The precipitation from this event was widespread but light, rendering the categorization of rainfall levels for threshold calculations impractical~~.
-

Reviewer #1, comment #4

4.Please add a baseline setup table. At minimum,it should specify input channels, preprocessing. training loss, retraining status, validation criterion, hyperparameter tuning, and evaluation procedure for generative models such as DGMR. If official implementations were used with minimal changes, state this explicitly. The point is not perfect fairness, but enough detail to determine whether gains come from architecture, multimodal inputs, loss design, or setup choices.

Our response #4

Thank you for highlighting the need for transparency and reproducibility. We have added explicit statements in the manuscript to clarify our baseline setups. To guarantee absolute fairness, all baseline models were trained and evaluated using their exact official open-source implementations without any architectural or hyperparameter modifications. Furthermore, the input datasets (including channels and preprocessing)

fed into the baselines were strictly identical to those used for our proposed method.

Because we adhered 100% to the official configurations for each respective baseline, their hyperparameters and specific training setups vary drastically by architecture. Compiling all these disparate, algorithm-specific parameters into a single coherent table proved impractical and potentially confusing. Instead, we have explicitly documented in the text that we strictly followed the parameters established in the official baseline repositories and corresponding papers. This also applies to the generative model DGMR, where the implementation is completely aligned with the official paper, changing only the input data stream to the SEVIR dataset.

We have revised the manuscript, and the specific changes are as follows:

- **Environment Setup, Page 15:** In our experiments, all baseline models employed the same multi-modal inputs, preprocessing pipelines, and dataset partitioning criteria as STI-DEDN, ensuring a fair comparison without architectural modifications. Additionally, Table 5 summarizes the computational efficiency during the inference phase to demonstrate practical operational viability.

Reviewer #1, comment #5

5. The random SEVIR event split may be standard, but it may still overestimate generalization if temporal, seasonal, regional, or storm-system correlations exist, A held-out year, season, or region test would substantially strengthen the paper. If this is not feasible, state it clearly as a limitation. Confine all claims to the specific conditions tested: lead time, metric type, event intensity, and test split. For example, a defensible claim is: "STI-DEDN improves HSS/CSI for 0-60 min forecasts of moderate-to-heavy VIL events on the random SEVIR test split." Avoid unqualified claims such as "outperforms all methods," "generalizes well," or "addresses extreme precipitation" without these qualifications. Report the count and percentage of no-rain events excluded. Discuss how this exclusion affects FAR, CSI, and HSS interpretation, especially whether scores are inflated relative to an operational seting with many null events. A sensitivity analysis with no-rain events retained is recommended if feasible.

Our response #5

(1) Regarding Claims and Generalization: We completely agree with your assessment. A random split might inadvertently capture temporal or storm-system correlations. Because the SEVIR dataset is structured around discrete weather events rather than continuous timelines, isolating specific seasons or geographical regions for a held-out test set poses significant structural challenges. We have now explicitly acknowledged this limitation in the manuscript. As suggested, we have carefully added qualifiers to our conclusions (specifying dataset, forecast horizons, and event intensities)

to avoid overclaiming. Incorporating specific seasonal and regional hold-out evaluations has been formally added to our Future Work section.

(2) Regarding the "No-Rain Events" : We owe the reviewers a deep apology for a serious descriptive error in our original text. In fact, from the original pool of approximately 13,000 SEVIR events, we strictly filtered for events due to a lack of sensor data or a lack of a continuous four-hour time gap. The total number of events excluded was approximately 2% of the dataset, leaving us with 12,000 complete events, of which we selected 10,000 for experimentation. We describe this practice of ours in the text as "we excluded events without rainfall," and we correct this wording in Section 4.1 to accurately reflect our data quality control process and to confirm that the natural category balance of empty events was preserved.

We have revised the manuscript, and the specific changes are as follows:

- **Abstract, Page 1:** Leveraging the SEVIR dataset, the proposed model integrates multiple meteorological variables to generate 1-h precipitation forecasts via Vertically Integrated Liquid (VIL) estimation. Experimental results demonstrate that the STI-driven framework achieves superior predictive accuracy and reduced error rates in hourly multi-step forecasting compared to state-of-the-art deep learning benchmarks.
- **Introduction, Page 4:** The experimental results show that the proposed method has more practical significance for VIL precipitation forecasting in medium and heavy VIL regions.
- **Dataset, Page 4:** We excluded weather events with less than 4 hours of weather duration and missing time points in weather events, accounting for 3\% of total events..
- **Conclusions, Page 29:** It is worth considering that our work uses a SEVIR dataset that cannot be correlated at the temporal, seasonal, regional, or storm system levels. The randomized division of this dataset may overestimate the generalization ability of the model, which provides a direction for our future research..
-

Reviewer #1, comment #6

6.Claims should be qualified by lead time, metric, and target variable. Longer-lead results show that other models achieve better MSE/PSNR, so the manuscript should not imply uniform superiority across all metrics and horizons.

Our response #6

We fully agree with this insightful feedback. We have carefully revised our manuscript to remove any implications of uniform superiority. Specifically, in the Abstract, Introduction, and Conclusion, we have strictly bounded our claims, explicitly stating that our model primarily improves structural and intensity-focused metrics (CSI and HSS) for short-term forecasting. Additionally, we have added a more balanced

discussion in Section 5.3, transparently highlighting the respective advantages of our method alongside the strengths of Transformer-based baselines (which inherently excel in MSE/PSNR over longer horizons).

We have revised the manuscript, and the specific changes are as follows:

- **Introduction, Page 4:** The experimental results show that the proposed method has more practical significance for VIL precipitation forecasting in medium and heavy VIL regions.
- **Results of Long-Term Forecasting, Page 25:** the HSS score for the 2-hour forecast improved by 20.38\% compared to the second-ranked model, Earthformer, while the HSS score for the 3-hour forecast showed a 5.72\% enhancement over the second-ranked model, PastNet. These results further validate the forecast potential of the STI-DEDN in precipitation forecasting tasks.
- **Results of Long-Term Forecasting, Page 25:** However, it exhibits poor performance in terms of indices such as the CSI and HSS, which characterize the distribution and structural details of precipitation regions. By contrast, the STI-DEDN model can predict relatively large local precipitation amounts. The inherent limitations of RNN-based methods restrict STI-DEDN's ability to capture long-term dependencies, while the attention mechanism can effectively model long-range spatiotemporal dependencies by directly establishing correlations among global pixels, thus demonstrating significant advantages in long-lead-time, pixel-level forecasting tasks. These points pave the way for our future research areas.

Reviewer #1, comment #7

7.The introduction is long and verbose. It takes about 140 lines to reach the main contribution. Please reduce it where possible..

Our response #7

We have thoroughly revised and streamlined the Introduction, removing overly verbose meteorological background information and bringing the focus to the core contributions much earlier in the text.

We have revised the manuscript, and the specific changes are as follows:

- **Introduction Page 2:** ~~Globally, varied precipitation intensities may trigger diverse natural disasters, resulting in significant socioeconomic disruptions and irreversible strain on ecological resources. Consequently, accurate precipitation prediction is indispensable for optimizing natural resource management, enabling early warning systems for extreme weather events, and mitigating their socioeconomic and environmental impacts.~~
- **Introduction, Page 2:** ~~This computational approach derives meteorological~~

~~forecasts by applying mathematically defined initial conditions and physical boundary constraints in its temporal simulations. With the nonlinear method, NWP generates probabilistic precipitation forecasts. However, computational demands escalate exponentially posing persistent challenges to practical implementation under complex synoptic conditions and sensitivity to uncertainties in initial conditions.~~

- **Introduction, Page 3:** ~~Furthermore, radar observational data, with their intricate multi-scale spatiotemporal dependencies, pose greater challenges than conventional image sequence prediction.~~
- **Introduction, Page 3:** ~~CNNs capture local spatial correlations through convolutional operations while stacking multiple layers allows for the hierarchical representation of features from local to global spatial scales. In contrast,~~

Reviewer #1, comment #8

8. Define NWP, LSTM, SOTA, and other abbreviations at first use.

Our response #8

We have carefully proofread the manuscript and ensured that all abbreviations (including NWP, LSTM, and SOTA) are fully spelled out and defined upon their first occurrence.

We have revised the manuscript, and the specific changes are as follows:

- **Introduction, Page 2:** Over the past decades, numerical weather prediction (NWP), which is grounded in solving physical equations, has served as the foundational methodology driving the evolution of meteorological forecasting.
- **Introduction, Page 3:** Long Short-Term Memory (LSTM) networks are inherently suitable for modeling temporal dynamics in sequential data, capturing long-term dependencies, and addressing the nonlinear evolution patterns present in time series.
- **Introduction, Page 2:** Recent existing state-of-the-art (SOTA) approaches in deep learning have addressed numerous challenges in traditional machine learning tasks, prompting researchers to adapt these techniques to precipitation forecasting.

Reviewer #1, comment #9

9. Add +1σ confidence intervals or 95% confidence intervals to the metrics in Tables 4 and 5, computed over the test samples or the relevant hold-out subset.

Our response #9

We have updated our results to include confidence intervals for the evaluation metrics, utilizing the testing data from our test set to compute the variance, thereby providing a clearer picture of the statistical stability of our performance.

Reviewer #1, comment #10

10. Improve the readability of large multi-panel figures. Figure 3 is useful as a schematic, but reproducibility requires a layer-by-layer table with channel counts, tensor shapes, activation functions, loss weights, and validation criterion.

Our response #10

We entirely agree that a layer-by-layer breakdown is vital for reproducibility. We have added a comprehensive architectural parameter table to the manuscript detailing the channel counts, tensor shapes, and activation functions for each layer of the network.

We have revised the manuscript, and the specific changes are as follows:

- **Method, Page 9:** Table 1. Detailed architectural configuration of STI-DEDN

Methods	MSE ↓	PSNR ↑	FAR-M ↓	CSI-M ↑	HSS-M ↑
Rainymotion	4.0654±1.1439	44.5022±1.5833	0.5570±0.2556	0.3346±0.2315	0.4290±0.2307
ConvLSTM	2.7321±0.8672	46.0102±1.6259	0.5686±0.2838	0.3730±0.2413	0.4742±0.2352
PredRNN	2.4193±0.7457	46.4704±1.5367	0.5190±0.2966	0.4008±0.2431	0.5038±0.2388
PhyDNet	2.5362±0.7526	46.1420±1.5120	0.5285±0.2626	0.3877±0.2364	0.4897±0.2348
Rainformer	2.6323±0.7129	46.0267±1.3811	0.5569±0.2775	0.3768±0.2446	0.4785±0.2410
Earthformer	2.5603±0.6216	46.4447±1.1418	0.5604±0.2978	0.3742±0.2501	0.4755±0.2489
DGMR	2.5312±0.6368	46.2833±0.8236	0.5481±0.3004	0.3775±0.2450	0.4766±0.2438
TAU	2.5993±0.6657	46.0232±1.1414	0.5385±0.2818	0.3698±0.2352	0.4708±0.2342
SimVP	2.6796±0.6609	46.2224±1.0962	0.5954±0.3085	0.3475±0.2549	0.4403±0.2571
PastNet	2.5256±0.6956	46.4227±1.2967	0.5488±0.3007	0.3788±0.2493	0.4768±0.2488
STI-DEDN	**2.3794±0.6775**	**46.5231±1.4467**	**0.4686±0.2403**	**0.4410±0.2236**	**0.5595±0.2034**

Reviewer #1, comment #11

11. Section 4.3 reports training time and inference speed;, add a compact comparison with key baselines, at least ConvLSTM, PredRNN, and Earthformer, or state clearly why this comparison is not feasible..

Our response #11

We have added the requested computational efficiency data. We expanded our discussion in the text and included the training epochs and inference speeds (in frames per second/seconds per sequence) for all evaluated baseline models alongside our own, providing a clear operational comparison.

We have revised the manuscript, and the specific changes are as follows:

- **Environment Setup, Page 15:** In our experiments, all baseline models employed the same multi-modal inputs, preprocessing pipelines, and dataset partitioning criteria as STI-DEDN, ensuring a fair comparison without architectural modifications. Additionally, Table 3 summarizes the computational efficiency during the inference phase to demonstrate practical operational viability.

Methods	Training Epochs	Inference Time (s)	FPS
Rainymotion	0	1.48	12.16
ConvLSTM	200	0.52	34.61
PredRNN	200	2.37	7.59
PhyDNet	200	0.85	21.18
Rainformer	200	0.45	40.00
Earthformer	200	0.51	35.29
DGMR	200	0.48	37.50
TAU	200	0.65	27.69
SimVP	200	0.43	41.86
PastNet	200	0.22	81.82
STI-DEDN	200	0.54	33.33

Reviewer #1, comment #12

12. How exactly are CRPS and BSS computed from model outputs?

Our response #12

As acknowledged in our response to Major Concern 3, interpreting these metrics was mathematically inappropriate for our deterministic point forecasts. We have completely removed CRPS and BSS from the manuscript.

Reviewer #1, comment #13

13. Do all baselines use the same multimodal inputs and preprocessing?.

Our response #13

Yes. All baseline models utilized the exact same multimodal inputs and preprocessing steps as our proposed method.

Reviewer #1, comment #14

14. Were all baselines retrained using the same split and validation criterion?.

Our response #14

Yes. All baselines were retrained from scratch using our exact same train/validation/test splits and identical validation criteria.

Reviewer #1, comment #15

15. How many no-rain events were removed, as a count and percentage?.

Our response #15

We have revised the relevant statement in the text. As detailed in our response to Main Concern 5, the statement in the original text is a description error. We excluded about 300 events (about 2% of the original sample pool) strictly based on missing sensor data, rather than excluding weather events due to lack of rainfall.

Reviewer #1, comment #16

16. Can the method be tested on held-out temporal or spatial subsets?.

Our response #16

Because the SEVIR dataset is event-based rather than a continuous chronological or mapped geospatial stream, creating strict temporal (seasonal) or spatial held-out subsets is not structurally supported without fundamentally altering the dataset. We have noted this as a limitation and earmarked it for future validation on continuous datasets.

Reviewer #1, comment #17

17. What is the precise implementation of the a weighting in ADGLoss?

Our response #17

We apologize if the practical implementation was not sufficiently clear in the text. While Equation (7) provides the mathematical definition, we have added a clarification in Section 3.3 detailing that, programmatically, the adaptive weight alpha is computed dynamically per batch during the forward pass. This allows the gradient to scale adaptively based on the specific precipitation intensity of the batch currently being processed.

- **Loss Function, Page 11:** Then it uses Eq.7 to calculate the ratio weight between the predicted image x_i and ground truth y_i. \alpha then acts on the MAE loss component in backpropagation. This mechanism ensures that the model adapts to underestimating and overestimating precipitation, forcing the model to pay more attention to the absolute error between the prediction and the true value.

2. Referee 2

Main Judgment

This study proposes using spatial and temporal information from multiple meteorological variables to improve precipitation nowcasting. The authors develop a dual encoder–decoder ConvLSTM architecture based on the spatio-temporal information transformation framework, consisting of a spatio-temporal converter for precipitation prediction and a temporal-spatial converter that provides an inverse mapping to promote spatio-temporal consistency. The paper is generally well organised. The model architecture and loss function are clearly described, and the experiments are well designed to evaluate the performance of the proposed model. The results are presented clearly, with effective illustrations and descriptions.

Our response

Dear Reviewer,
We would like to express our deepest gratitude for your time and effort in reviewing our manuscript. We are highly encouraged by your positive assessment of our work, particularly your recognition of our model architecture, loss function, and overall experimental design. Your constructive feedback has been invaluable in helping us refine our narrative, clarify our experimental setups, and strengthen the overall rigor of the paper.
Below, we provide a detailed, point-by-point response to each of your comments, outlining the corresponding modifications made to the revised manuscript.
Best regards

Reviewer #2, comment #1

1.Introduction and Related Work sections: These sections are somewhat descriptive and read more like a listing of previous studies. The manuscript would benefit from additional synthesis and transition and summary sentences. This would help guide readers more effectively and better motivate the proposed approach.

Our response #1

We sincerely thank the reviewer for this constructive structural advice. We have revised both the Introduction and Related Work sections. Specifically, we added comprehensive synthesis paragraphs and improved the transitional phrasing between different methodological families. This new structure better summarizes the common limitations of existing models and provides a much clearer, logical progression that directly

motivates our proposed approach.

We have revised the manuscript, and the specific changes are as follows:

- **Introduction, Page 3:** Despite the robust feature learning capabilities of CNNs and RNNs, they still exhibit degraded performance when relying solely on short-term data and tend to underestimate intense precipitation events.
- **Introduction, Page 4:** Although current precipitation forecasting methods account for spatio-temporal features, the nonlinear characteristics and causal spatio-temporal relationships embedded in high-dimensional meteorological variables within historical observational data remain underutilized.
- **Related Work, Page 5:** Despite the robust feature learning capabilities of CNNs and RNNs, they still exhibit degraded performance when relying solely on short-term data and tend to underestimate intense precipitation events

Reviewer #2, comment #2

2. Training with precipitation events only: At line 259, the authors mention that events with no rainfall are excluded from training. This choice requires further discussion. Although including many no-rain samples may lead to class imbalance, completely removing them may bias the training distribution and may require an additional strategy to handle rainfall occurrence and no-rain situations in operational applications.

Our response #2

We owe you a profound apology for a critical typographical error in our original text. We mistakenly wrote that "we excluded events with no rainfall." In reality, from the original pool of approximately 13,000 SEVIR events, we filtered out events strictly due to missing sensor data or severe gaps in continuous time intervals. We did not filter out no-rain events. The total number of excluded events accounted for only about 2% of the dataset. We utilized 10,000 of the remaining complete events for our experiments. We have corrected this phrasing in Section 4.1 to accurately reflect our data quality control process and to reassure readers that the natural class balance of the operational environment was preserved.

We have revised the manuscript, and the specific changes are as follows:

- **Dataset, Page 4:** We excluded weather events with less than 4 hours of weather duration and missing time points in weather events, accounting for 3\% of total events.

Reviewer #2, comment #3

3. Performance of the proposed model: The conclusion drawn in lines 324–325 could be too strong. The experimental results show that the proposed architecture performs better than other models under the specific experimental setting (e.g., the chosen dataset). Do authors think the proposed architecture is generally superior beyond the tested experimental conditions? For operational applications, a fair comparison between different models should allow each model to use its preferred input data, loss function, and training strategy. The forecast skill should then be evaluated against independent observational sources that are separate from the data used to train any of the models.

Our response #3

We appreciate your diligence in ensuring academic rigor.

First, we have carefully reviewed our manuscript and added necessary qualifiers to all conclusive statements, explicitly bounding our claims (e.g., stating that the model demonstrates "improved CSI for 1-hour VIL nowcasting within the SEVIR dataset").

Second, regarding general superiority beyond the tested conditions: theoretically, our architecture is designed to generalize. To validate this, we have proactively conducted an additional experiment using an external, real-world precipitation dataset for training and testing. The preliminary results confirmed the effectiveness of our model. We have included a discussion on this out-of-distribution generalization in the Future Work section. This experiment is not the main experiment of this paper, but only a verification experiment.

Finally, regarding fair comparison: to guarantee absolute fairness, all baseline models were trained using their official, optimal parameter settings and recommended training strategies. We made zero modifications to the underlying baseline architectures. Furthermore, both our model and the baselines were fed the exact same input data and evaluated on the same test sets, thereby ensuring a strictly level playing field for deep learning model benchmarking.

We have revised the manuscript, and the specific changes are as follows:

- **Experimental Results and Analysis, Page 16:** In the foundational one-hour precipitation forecast, ConvLSTM and SimVP exhibit relatively MSE scores and relatively low PSNR scores. This indicates that the capabilities of CNN and RNN-based models in extracting precipitation information are comparatively weak.

- **Environment Setup, Page 15:** In our experiments, all baseline models employed the same multi-modal inputs, preprocessing pipelines, and dataset partitioning criteria as STI-DEDN, ensuring a fair comparison without architectural modifications. Additionally, Table 5 summarizes the computational efficiency during the inference phase to demonstrate practical operational viability.
- **Experimental Results and Analysis, Page 17:** Finally, Table 7 is an additional experiment on the MeteoNet dataset - a high-resolution meteorological dataset from France containing multi-source meteorological data including real radar precipitation observations with 5-minute temporal resolution. This supplementary experiment was not part of our primary evaluation but served to validate STI-DEDN's effectiveness for real precipitation forecasting. The experiment uses one hour of radar precipitation observations as input to predict precipitation for the following hour. The experiment is compared with the traditional optical flow method Rainymotion, and the results show that STI-DEDN is still valid for real precipitation forecasting, which is an effective step for us to use real-world precipitation as a forecasting condition in future studies.

Table 7. Comparison of index results of each method under MeoteoNet test set

Methods	MSE ↓	PSNR ↑	FAR-M ↓	CSI-M ↑	HSS-M ↑
Rainymotion	13.9703	27.4228	0.4880	0.3798	0.5175
STI-DEDN	**10.9466**	**29.1939**	**0.4563**	**0.4339**	**0.5708**

Reviewer #2, comment #4

4. Training with interpolated data: Table 2 indicates that training with interpolated data slightly outperforms training with the original-resolution data. However, the stated motivation for interpolation is reduced training complexity, suggesting an efficiency-driven rather than accuracy-driven choice. Since the authors also conducted experiments at the original resolution, training without interpolation appears feasible. If efficiency is the main motivation, the authors could provide a direct comparison of computational cost, such as training time and memory usage, between the interpolated and original-resolution settings..

Our response #4

We agree with your observation. To directly address the efficiency motivation, we have added a comprehensive comparison of computational costs to the manuscript. This new addition explicitly details the differences in training time and model parameter volume

between the interpolated and original-resolution settings, thereby providing concrete evidence for our preprocessing choices.

We have revised the manuscript, and the specific changes are as follows:

- **4.1 Dataset, Page 13: Table 3.** Comparison of results with and without interpolation for STI-DEDN on the SEVIR test set. NOP represents the number of parame-ters of the model, and ET represents the training time of the model for each round of epoch

Methods	MSE ↓	PSNR ↑	FAR-M ↓	CSI-M ↑	HSS-M ↑	NOP ↓	ET ↓
STI-DEDN(Original resolution)	2.3847	46.5070	0.4781	0.4388	0.5579	34.2017M	1.897s
STI-DEDN(Interpolation)	**2.3794**	**46.5231**	**0.4686**	**0.4410**	**0.5595**	**34.2016M**	**1.697s**

Reviewer #2, comment #5

5.Line 29: The phrase "… recent observational data more effectively capture precipitation at the next time step…" reads a bit confusing.

Our response #5

We agree that the phrasing was awkward. We have rephrased this sentence in the revised manuscript to be much clearer and easier to understand.
We have revised the manuscript, and the specific changes are as follows:

- **4.1 Introduction, Page 12:** In the atmospheric system, the closer the observed data to the occurrence time of precipitation events can help forecasting methods to make more effective predictions

Reviewer #2, comment #6

6. Line 51: As the acronym SOTA only appears once in the manuscript, the full name should be used instead.

Our response #6

Corrected. We have replaced the acronym with "state-of-the-art".

We have revised the manuscript, and the specific changes are as follows:

- **Introduction, Page 2:** Recent existing state-of-the-art (SOTA) approaches in deep learning have addressed numerous challenges in traditional machine learning tasks, prompting researchers to adapt these techniques to precipitation forecasting.

Reviewer #2, comment #7

7. Line 63: The full name of LSTM should be given for the first time.

Our response #7

Corrected. We now provide the full name, Long Short-Term Memory, at its first occurrence.

We have revised the manuscript, and the specific changes are as follows:

- **Introduction, Page 3:** Long Short-Term Memory (LSTM) networks are inherently suitable for modeling temporal dynamics in sequential data, capturing long-term dependencies, and addressing the nonlinear evolution patterns present in time series.

Reviewer #2, comment #8

8. Lines 137–138: The paper "Predicting future dynamics from short-term time series using an Anticipated Learning Machine" (https://doi.org/10.1093/nsr/nwaa025) seems relevant to the discussion of using spatio-temporal information.

Our response #8

We sincerely thank you for recommending this highly relevant literature. The Anticipated Learning Machine (ALM) paper introduces an elegant framework capable of extracting temporal evolution rules directly from short-term observations without relying on traditional numerical integration. This conceptual framework aligns perfectly with our discussion on leveraging high-dimensional spatiotemporal information to predict future dynamics. We have synthesized its core innovations and incorporated this important reference into our Related Work section to enrich the theoretical background of spatiotemporal modeling.

Reviewer #2, comment #9

9. Lines 257-258: Should the contribution of lightning data to precipitation forecasting be casedependent, for example depending on rainfall type, convective intensity, and the availability of other data types?

Our response #9

We completely agree with your nuanced perspective. The role of lightning data is indeed highly case-dependent and strongly correlated with intense convective events. We have revised our statement in Section 4.1 to acknowledge this. We clarified that our decision to exclude lightning was primarily due to its sparse point-event format.

Aligning point data with our dense, rasterized radar framework causes spatial mismatch issues and introduces interpolation artifacts. This exclusion does not imply that lightning lacks meteorological value. Furthermore, as noted in reference[1] of our manuscript, incorporating lightning data within the SEVIR dataset did not yield a positive impact on VIL forecasting specifically.

We have revised the manuscript, and the specific changes are as follows:

- **Experiments Basis, Page 12:** Lightning data is stored as site events (precise spatiotemporal coordinates), while the other modalities (IR, VIS, VIL) are all rasterized radar images. Lightning events that cannot be captured by meteorological stations are set as no-lightning supervision information. Converting sparse point events into raster format requires additional interpolation and aggregation methods, which inevitably introduce artificial patterns and noise, deviating from the true physical process. This spatiotemporal mismatch can interfere with model learning and reduce the reliability of predictions. Lightning is a byproduct of strong convective activity, and strong convective activity does not necessarily accompany lightning events

Reviewer #2, comment #10

10. Line 266: What is the spatial resolution after interpolation?

Our response #10

After interpolation, the spatial resolution of the data is 1.5 km. We have explicitly added this detail to the text.

We have revised the manuscript, and the specific changes are as follows:

- **Experiments, Page 12:** To resolve spatial resolution inconsistencies among data modalities in the SEVIR dataset, we uniformly resampled all data types via bicubic interpolation, standardizing their dimensions to 256 × 256 (with a spatial resolution of 1.5km) to reduce model training complexity.

Reviewer #2, comment #11

11. Line 321: I wonder whether the same loss functions were used across all models.

Our response #11

To ensure a fair comparison, all baseline models were trained using the standard, optimal loss functions recommended in their respective official implementations. Conversely, our STI-DEDN was trained using our proposed ADGLoss. We have

clarified this in the text.

We have revised the manuscript, and the specific changes are as follows:

- **Environment Setup, Page 15:** In our experiments, all baseline models employed the same multi-modal inputs, preprocessing pipelines, and dataset partitioning criteria as STI-DEDN, ensuring a fair comparison without architectural modifications.

Reviewer #2, comment #12

12. Line 330: The word "efficient" usually implies reduced computational cost. However, in this context, the authors seem to be referring mainly to the accuracy or effectiveness of the model in representing uncertainty.

Our response #12

Thank you for pointing out this lexical inaccuracy. We have removed this part of the text.

We have revised the manuscript, and the specific changes are as follows:

- **Experimental Results and Analysis, Page 16:** ~~Furthermore, to validate the efficacy of our method in efficient uncertainty modeling, we employed the CRPS metric to quantify the discrepancy between predicted probability distributions from different methods and the actual distributions of observed images. As illustrated in the Table6, the STI-DEDN model demonstrated a 9.39% improvement in CRPS compared to the worst-performing model, and a 4.06% enhancement in CRPS relative to the second-best PredRNN model. This demonstrates that precipitation predictions produced by the STI-DEDN model display probability distributions closest to observational data indicating enhanced proficiency in uncertainty modeling~~.

Reviewer #2, comment #13

13. Figures 5 and 6: The proposed method appears to be clearly separated from the other methods in terms of forecast skill, while the remaining methods seem to form a relatively similar group. Could the authors discuss the main factors contributing to this separation?

Our response #13

The significant separation in forecast skill is primarily attributable to the synergistic effect of our STI-driven dual encoder-decoder design and the ADGLoss function.

Standard baseline models typically suffer from progressively smoothed results and severe error accumulation over longer time steps. In contrast, our spatiotemporal dual structure acts as a robust regularizer. Combined with the adaptive gradient penalty of the ADGLoss, the model strictly enforces spatiotemporal consistency during training. This significantly reduces uncertainty amplification and preserves structural fidelity much better than standard RNN/CNN architectures.

Reviewer #2, comment #14

14. Figure 6: Could the authors explain the jumps in the PastNet results at small spatial scales, shown by the purple lines?

Our response #14

The abnormal jumps at small spatial scales for PastNet are likely due to its unique physical prior modules and internal spatial resampling steps. These mechanisms can inadvertently inject unphysical high-frequency noise (artifacts) when forecasting highly complex, non-smooth precipitation patterns. It is worth noting that deep learning models inherently possess stochasticity, and some degree of fluctuation is normal when forecasting highly non-linear precipitation, but PastNet's architectural traits make these artifacts more pronounced at small scales. We have added a brief explanatory note regarding this to the text.

We have revised the manuscript, and the specific changes are as follows:

- **Experimental Results and Analysis, Page 1:** It is worth noting that deep learning models inherently possess stochasticity, and some degree of fluctuation is normal when forecasting highly non-linear precipitation, but PastNet's architectural traits make these artifacts more pronounced at small scales. Overall, the results indicate that the STI-DEDN is capable of providing robust technical support for multi-scale forecasting tasks in the context of conventional precipitation events.

Reviewer #2, comment #15

15. Line 417: The phrase "Overall, STI-DEDN is more applicable in large-scale extreme precipitation events" could be confusing. Forecasting extreme precipitation more accurately at larger spatial scales is not the same as forecasting large-scale extreme precipitation events.

Our response #15

We apologize for the semantic confusion. We have corrected the phrasing to accurately

reflect that our method is "capable of forecasting extreme precipitation more accurately at larger spatial scales," as you kindly suggested.

We have revised the manuscript, and the specific changes are as follows:

- **Case 2: Evaluation of Extreme Precipitation Events, Page 21:** Overall, STI-DEDN is more suitable for forecasting precipitation under large-scale spatial conditions.

Reviewer #2, comment #16

16. Figures 5, 7 and 9: Rainymotion appears to preserve more small-scale texture than other models. Could the authors discuss the reason for this?

Our response #16

Rainymotion preserves high-frequency, small-scale textures because it is a physics-based optical flow method, not a deep neural network. It essentially advects existing pixel intensities rigidly across the grid based on motion vectors. Consequently, it completely avoids the blurring and smoothing effects that are inherently caused by the convolutional downsampling and upsampling operations utilized in deep learning autoencoders.

Reviewer #2, comment #17

17. Section 5.4.2: The meaning of "removing the STI framework" could be defined more explicitly. Does this refer specifically to removing the dual spatio-temporal and temporalspatial converter structure?

Our response #17

Yes, your understanding is exactly correct. We have explicitly clarified this definition in the text. "Removing the STI framework" means downgrading our architecture to a standard, single-stream encoder-decoder network by completely removing the temporal-spatial converter (the inverse mapping constraint) and its associated conjugate training mechanism.

We have revised the manuscript, and the specific changes are as follows:

- **Abstract, Page 1:** To assess the optimization role of the STI framework in precipitation forecasting enhancement, we performed ablation studies through both framework exclusion (remove dual architecture and use only one-way training process) and its integration with other high-performing prediction models.

Reviewer #2, comment #18

1. There is missing space before "(" in several places, for example in lines 79, 85, 89 and 172.

2. Please use lowercase "on" in titles.

3. Please ensure consistent capitalization across all titles.

4. Line 261: There appears to be a duplicated period at the end of the sentence.

Our response #18

We have meticulously implemented all technical corrections. We added the missing spaces, corrected the capitalization of "on" and other words in titles for consistency, and removed the duplicate period. We deeply appreciate your exceptionally thorough review of our manuscript.

References

[1]. Yang, N. and Li, X.: Lightweight AI-powered precipitation nowcasting, The Innovation Geoscience, 2, 100 066, https://doi.org/10.59717/j.xinn-geo.2024.100066, 2024.