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

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.

Major concerns

1. The manuscript 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.

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.

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.

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.

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 setting with many null events. A sensitivity analysis with no-rain events retained is recommended if feasible.

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.

Minor issues

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

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

3. 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.

4. 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.

5. 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.

Questions for clarification

1. How exactly are CRPS and BSS computed from model outputs?
2. Do all baselines use the same multimodal inputs and preprocessing?
3. Were all baselines retrained using the same split and validation criterion?
4. How many no-rain events were removed, as a count and percentage?
5. Can the method be tested on held-out temporal or spatial subsets?
6. What is the precise implementation of the $\alpha$ weighting in ADGLoss?