https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

Spatiotemporal Dual-Stream Transformers for Cloud Microphysical Parameterization

Yijun Huang[2], Qi Zhang[3], Hoiio Kong[3], Chan-Seng Wong[3], Huan Zhao[4], and Ting Shu[1]

[1]School of Artificial Intelligence, Shenzhen University, Shenzhen, China
[2]Hainan International College, Minzu University of China, Hainan, China
[3]Faculty of Data Science, City University of Macau, Macau, China
[4]Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong, China

Correspondence: Ting Shu (tingshu@szu.edu.cn)

Abstract. Accurate precipitation forecasting is essential for mitigating weather-related disasters. Numerical Weather Prediction (NWP) precipitation forecasting accuracy is largely constrained by microphysical parameterization schemes, which rely on simplifying assumptions that introduce uncertainties. Deep learning provides a promising approach for data-driven modeling of complex microphysical relationships. We propose to model the cloud microphysical process via the Learned Microphysics Transformer (LMP-Tr). LMP-Tr employs a hybrid Convolutional Neural Network (CNN)–Transformer architecture that alternately integrates multi-scale convolutional modules and dual-pathway attention modules to capture both local cloud-scale features and long-range atmospheric dependencies. The key innovation lies in the systematic alternation of multi-scale convolutional modules for local feature extraction and dual-pathway attention modules for global dependency modeling. The proposed model enables progressive refinement of atmospheric representations through height-variable attention pathways and cross-module attention mechanisms. Extensive evaluation on a WRF simulation dataset demonstrates superior performance of the proposed method. LMP-Tr provides a practical and effective solution for enhancing cloud microphysics representation in operational NWP systems, offering improved accuracy and physical consistency compared to other Artificial Intelligence (AI)-based parameterization approaches.

1 Introduction

Accurate precipitation forecasting is essential for addressing escalating challenges posed by global climate change and increasingly frequent extreme weather events. Numerical Weather Prediction (NWP) Shuman (1978) has emerged as the predominant approach for operational weather forecasting worldwide. However, precipitation prediction accuracy in NWP systems is largely constrained by cloud microphysical parameterization schemes, which represent a critical yet inherently uncertain component of atmospheric modeling. Traditional schemes, including single-moment Cotton et al. (1995); Lin et al. (1983), double-moment Morrison et al. (2005), and advanced multi-moment approaches such as Milbrandt-Yau Milbrandt and Yau (2005a, b), Seifert-Beheng Seifert and Beheng (2006), Thompson Thompson et al. (2008); Morrison and Gettelman (2008), P3 Morrison and Milbrandt (2015); Milbrandt and Morrison (2016, 2021), and grid-scale cloud schemes Ferrier et al. (2002), face fundamental limitations from lack of comprehensive observational constraints, inherent closure problems, and numerous simplifying as-

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

sumptions Liu et al. (2023). Comprehensive evaluations Köcher et al. (2023); Segele et al. (2013); Jin and et al. (2023); Shi et al. (2013); Morrison et al. (2020) demonstrate that no existing scheme adequately represents all physical processes across diverse meteorological conditions.

Machine learning offers transformative potential for atmospheric modeling through exceptional nonlinear approximation capabilities Ren et al. (2021). Deep learning-based weather prediction frameworks Weyn et al. (2020) demonstrate substantial potential for enhancing atmospheric process representation. Applications to cloud microphysics include warm rain processes Gettelman et al. (2021); Seifert and Rasp (2020), autoconversion Alfonso and Zamora (2021), raindrop formation Takeishi et al. (2024), and emulation frameworks Sharma and et al. (2025); Arnold and et al. (2024). Integration frameworks like Weather Research and Forecasting-Machine Learning (WRF-ML) Zhong and et al. (2023) and radiative transfer coupling Mu et al. (2023) demonstrate practical feasibility, while basis function approaches Rodríguez-Genó et al. (2022) expand the methodological toolkit. Recent advances Yuval and O'Gorman (2020); Beucler et al. (2021); Brenowitz et al. (2020) show promise for incorporating physical constraints and specialized processes such as convective momentum transport Bryan et al. (2003) and electrification Mansell et al. (2010). However, current approaches face limitations in capturing global-scale coherent structures and complex inter-variable dependencies that characterize highly nonlinear, multi-scale atmospheric systems.

Building upon Shu et al. (2024) Shu et al. (2024), which introduced a pioneering One-Dimensional Convolutional Neural Network (1D-CNN) framework, we address the limitation of capturing long-range dependencies by introducing LMP-Tr (Learned Microphysics Transformer), a Transformer-based cloud microphysical parameterization scheme. Our approach combines 1D-CNN's local feature extraction with global attention mechanisms, employing alternating multi-scale convolutional modules and dual-pathway attention modules (TR_HEI, TR_VAR, and TR_CROSS) to capture vertical and inter-variable dependencies. The primary contributions of this paper are summarized as follows:

1. A Transformer-based cloud microphysical parameterization scheme (LMP-Tr) is proposed, which provides another implementation approach for cloud microphysics parameterization in operational NWP systems.
2. The dual-pathway attention mechanisms (TR_HEI, TR_VAR, and TR_CROSS) in LMP-Tr are developed to capture both local atmospheric features and global vertical–variable dependencies.
3. Two versions of LMP-Tr are implemented and evaluated on this dataset; their performance differences highlight the importance of matching architectural complexity to physical process requirements.

2 Methodology

2.1 LMP-Tr

The overall architecture of LMP-Tr (Learned Microphysics Transformer) is illustrated in Figure 1(a). This AI-based parameterization adopts a hybrid CNN–Transformer structure to simultaneously capture local cloud-scale features and long-range atmospheric dependencies. The core design principle is the systematic alternation of multi-scale convolutional modules for local feature extraction and dual-pathway attention modules for global dependency modeling, enabling progressive refinement of

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

atmospheric representations. The architecture features three pairs of alternating multi-scale convolutional modules (pink) and dual-pathway attention modules (blue), followed by a Conv1D layer (green), with extensive skip connections (black arrows) enabling dense feature reuse. Each sample is one vertical column at a single grid point and time step, represented as $\mathbf{X} \in \mathbb{R}^{H \times V_{\text{in}}}$ ($H = 50$ layers; $V_{\text{in}} = 15$ input variables) with output $\mathbf{Y} \in \mathbb{R}^{H \times V_{\text{out}}}$ ($V_{\text{out}} = 12$). Here, "sequence" denotes the vertical profile, not a multi-time history. LMP-Tr learns the column-wise, one-step Markovian update $\Delta \mathbf{x}_{t+\Delta t} = f(\mathbf{x}_t, \mathbf{e}_t)$ without lagged time steps. In this study, "spatiotemporal" refers to the height–variable dual-stream representation and its one-step update in time-evolving NWP fields, rather than multi-time network inputs. This alternating design enables iterative refinement: convolutional modules extract multi-scale local features while attention modules model global dependencies across vertical and variable dimensions. To enhance information flow, a dense feature reuse mechanism propagates features from earlier pairs to later stages at three levels: (1) input features propagate directly to subsequent pairs, (2) features from earlier pairs integrate into later pairs, and (3) intermediate features from all pairs concatenate into the final Conv1D layer. For each convolutional-attention pair i:

$$F_{\text{dense}}^{(i)} = \text{Concat}(F_{\text{input}}, F_{\text{IT}}^{(1)}, F_{\text{IT}}^{(2)}, \ldots, F_{\text{IT}}^{(i)}) \tag{1}$$

where F_{input} represents original input features and $F_{\text{IT}}^{(i)}$ denotes the output of the i-th pair. This mechanism ensures stable gradient flow, accelerates training convergence, and enhances generalization. LMP-Tr employs *Logarithmic-Quadratic Activation* (LQA) in the multi-scale convolutional layers of both model variants, defined as:

$$\sigma_{\text{LQA}}(x) = \ln(1 + e^x) + \frac{x^2}{2} \tag{2}$$

Compared to ReLU, LQA provides smoother nonlinear characteristics that facilitate capturing complex meteorological relationships while mitigating gradient vanishing and ensuring numerical stability. In the remaining modules, multi-head self-attention employs the built-in softmax; the first feed-forward layer uses ReLU, whereas the second feed-forward layer, layer normalization, and the output head are linear without additional activation. For fair comparison, all deep learning models in the comparative and ablation experiments—including LMP-Tr variants, baseline architectures, and ablated counterparts—use LQA in their convolutional layers wherever convolutions are present; other modules retain their standard activations (e.g., ReLU or softmax).

2.2 Alternating Inception-Transformer Architecture

The LMP-Tr scheme employs an alternating architecture where multi-scale convolutional modules and dual-pathway attention modules are systematically interleaved to achieve progressive feature refinement and global dependency modeling. This design addresses the need to simultaneously capture fine-grained local features (e.g., phase transitions at specific layers) and long-range dependencies (e.g., vertical stratification effects and inter-variable interactions). The processing flow follows a three-stage pattern: (1) convolutional modules extract multi-scale local features, (2) attention modules model global dependencies across vertical and variable dimensions, and (3) refined features pass to the next stage. Through three pairs of alternating modules,

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

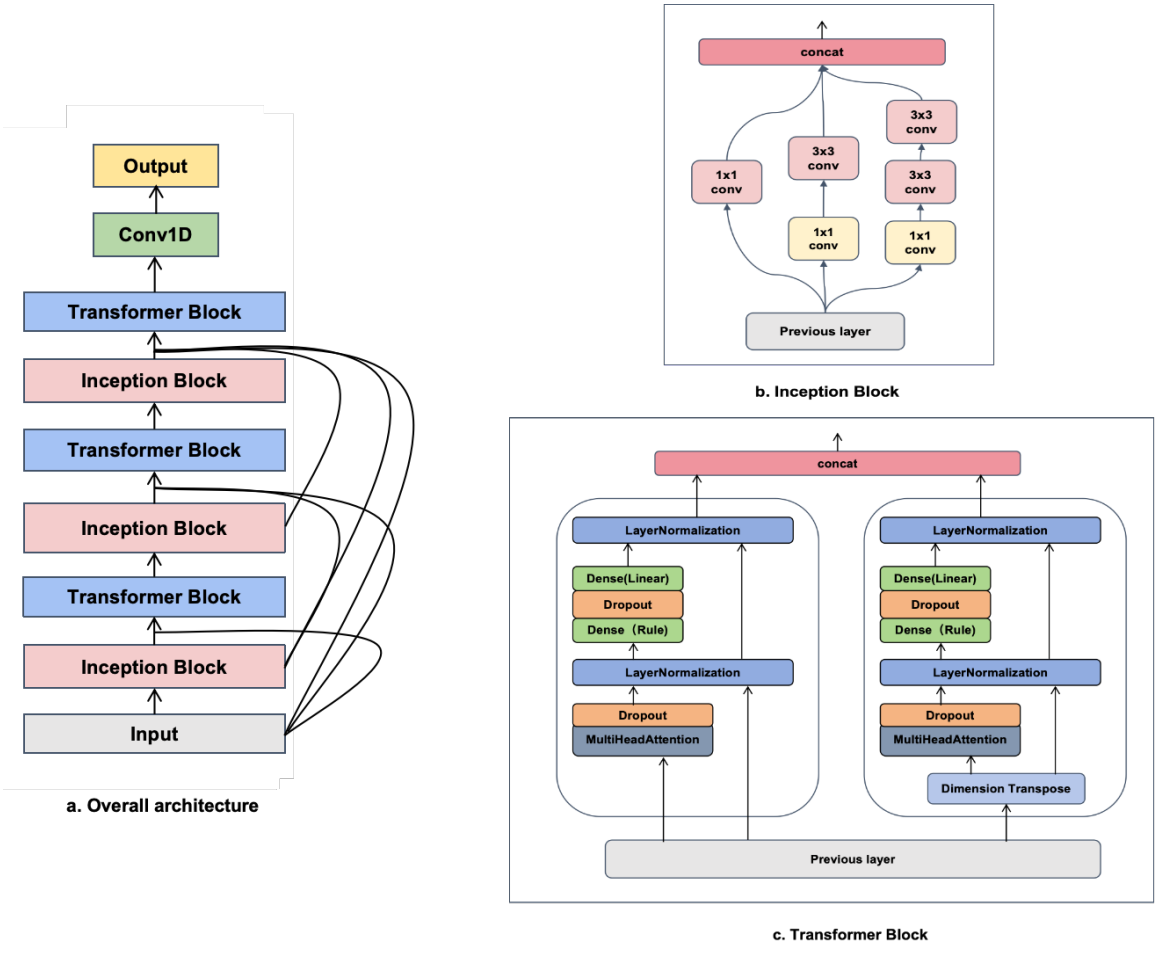


Figure 1. Architecture of the LMP-Tr parameterization scheme. (a) Overall architecture. (b) Multi-scale convolutional module. (c) Dual-pathway attention module.

features undergo progressive refinement—early stages capture basic local patterns, intermediate stages integrate multi-scale information, and later stages establish comprehensive global dependencies.

2.3 Multi-Scale Convolutional Architecture

The multi-scale convolutional architecture serves as the core local feature extraction module, designed to capture multi-scale atmospheric patterns through parallel convolutional pathways (Figure 1(b)). The architecture employs parallel branches at different vertical scales: Branch-1 uses 1×1 convolutions, Branch-2 combines 1×1 and 3×3 convolutions, and Branch-3 (in LMP-Tr-2) applies sequential 3×3 convolutions. All branch outputs are concatenated along the channel dimension. Convolutions are 1D along the vertical axis (H): 1×1 kernels mix variables/features at the same level, whereas 3×3 kernels aggregate three adjacent vertical layers. Global height–variable dependencies are handled by the attention modules (TR_HEI, TR_VAR).

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

We introduce two variants: LMP-Tr-1 employs a two-branch structure where Branch-1 uses 1×1 convolution for dimensionality reduction and Branch-2 adds 1×1 followed by 3×3 for larger-scale features, providing focused extraction suitable for well-defined patterns; LMP-Tr-2 extends this with a third pathway where Branch-3 applies 1×1 followed by two sequential 3×3 convolutions for large-scale context, enabling sophisticated multi-scale integration for complex atmospheric processes. The architectural difference reflects a trade-off: the dual-pathway design excels for well-structured processes, while the three-pathway design benefits tasks requiring sophisticated multi-scale interactions.

2.4 Dual-Pathway Attention Architecture

Following each multi-scale convolutional module, dual-pathway attention modules capture global dependencies across different physical dimensions (Figure 1(c)). The module comprises three sub-modules: TR_HEI for height-dimension attention, TR_VAR for variable-dimension attention, and TR_CROSS for cross-attention fusion, each employing multi-head self-attention and feed-forward networks with residual connections and layer normalization. To capture different relationship types simultaneously, the model employs Multi-Head Self-Attention (MHSA, Figure 2), which projects queries, keys, and values h times with learned linear projections (W_i^Q, W_i^K, W_i^V), then performs scaled dot-product attention in parallel. Outputs from all h heads are concatenated and linearly transformed:

$$\text{MHSA}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \tag{3}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$. The core component is scaled dot-product attention (Figure 2), operating through: (1) computing similarity scores between query (Q) and key (K) vectors, (2) scaling by $\frac{1}{\sqrt{d_k}}$ and applying softmax, and (3) weighting value (V) vectors by attention weights. Mathematically:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

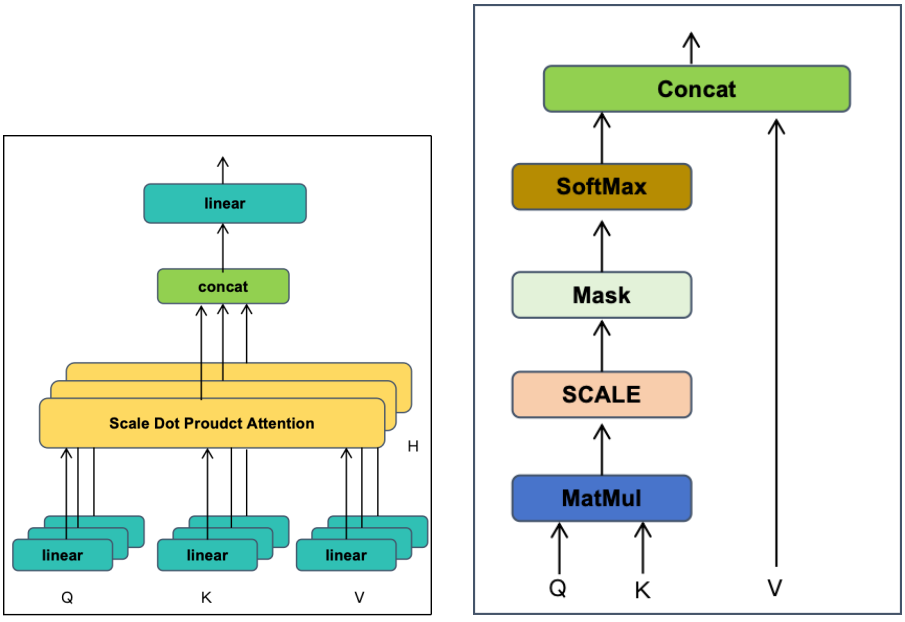
where d_k represents the key dimension. The scaling factor prevents large dot products that would push softmax into low-gradient regions, ensuring stable gradient flow and enabling dynamic focus on relevant input parts.

Dual-pathway dependency modeling employs two parallel Transformer modules (Figure 1(c)). TR_HEI (Height-Dimension Attention) captures dependencies along the vertical dimension by treating each layer as a sequence element, enabling understanding of vertical atmospheric structure and long-range dependencies; this is crucial where vertical stratification plays fundamental roles. TR_VAR (Variable-Dimension Attention) models interactions among cloud microphysical variables by treating each variable as a sequence element, capturing inter-variable relationships such as coupling between temperature, humidity, and hydrometeor species. Both modules employ identical Transformer architectures but operate on different dimensional views, enabling complementary feature extraction. After projection to $\hat{F}_H$ and $\hat{F}_V$, TR_CROSS fuses the two views via cross-attention ($\hat{F}_H$ as Query; $\hat{F}_V$ as Key and Value):

$$F_{\text{cross}} = \text{MHSA}(\hat{F}_H, \hat{F}_V, \hat{F}_V), \quad F_{\text{norm}} = \text{LN}(F_{\text{cross}} + F_H) \tag{5}$$

followed by a residual feed-forward network, yielding the fused output F_{IT} for the next stage.

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026



(a) Multi-head self-attention. (b) Scaled dot-product attention.

Figure 2. Attention mechanisms: (a) Multi-head self-attention; the input is split into multiple heads with learned Q, K, V projections. (b) Scaled dot-product attention; the three-step flow: Q-K similarity, softmax scaling, V weighting.

3 Experiments

3.1 Dataset

Following Shu et al. (2024) Shu et al. (2024), we constructed a WRF simulation dataset of cloud microphysical parameters over Southern China (2018–2020). All samples are model-generated outputs rather than observational measurements. The dataset includes 30 precipitation events selected based on meteorological criteria, with each event comprising over 270,000 samples, totaling more than eight million instances. Data were generated using the Thompson microphysical scheme Thompson et al. (2008), preserving complete spatiotemporal characteristics of cloud variables.

3.1.1 WRF Model Configuration

We adopt WRF model version 4.2.1 with the Thompson microphysical scheme. Initial and boundary conditions are derived from European Centre for Medium-Range Weather Forecasts (ECMWF) operational analysis data. Parameter settings are consistent with real-time operational forecasts of the Shenzhen Meteorological Bureau.

The horizontal domain (Figure 3) is centered at 22.8877°N and 113.6719°E, utilizing Lambert conformal conic projection. Horizontal resolution is 3 km × 3 km, with grid size 577 × 481 points, covering 104.8895°E–122.4543°E and 16.2589°N–29.3032°N. The vertical coordinate system employs terrain-following configuration with 50 layers. The integration time step is 18 sec-

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

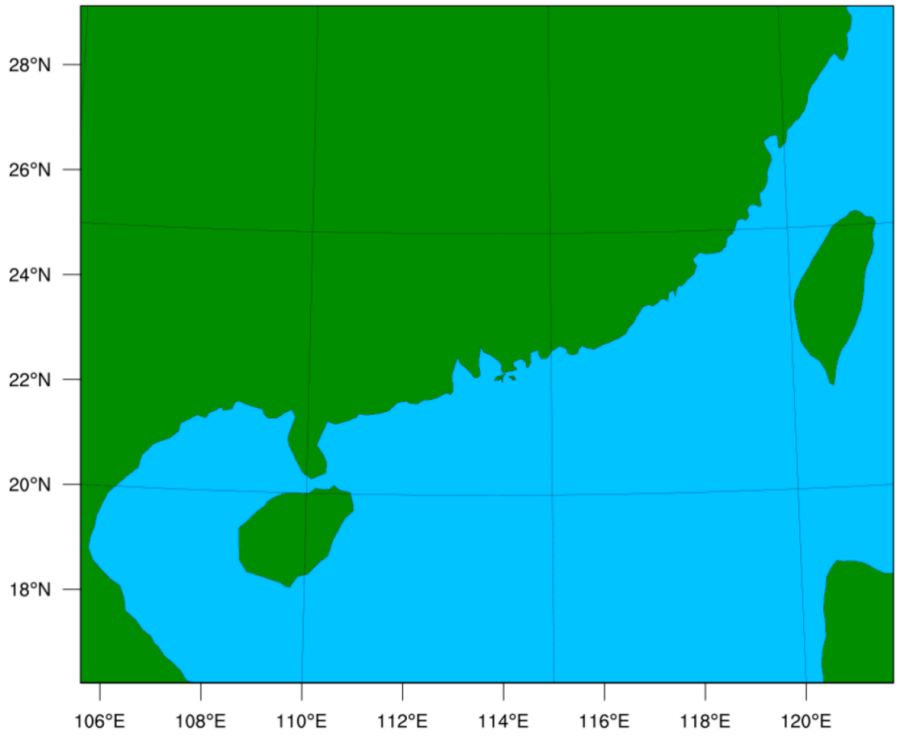


Figure 3. The horizontal domain of Weather Research and Forecasting Preprocessing System (WPS) and WRF in this work. The longitude ranges from 104.8895°E to 122.4543°E and the latitude ranges from 16.2589°N to 29.3032°N, covering the Southern China region.

onds. Physical parameterization schemes include RRTMG for longwave and shortwave radiation, Yonsei University scheme for boundary layer, revised MM5 scheme for surface layer, and unified Noah model for land-surface processes; no cumulus parameterization is employed (cu_physics = 0), as 3-km resolution resolves convective processes explicitly.

3.1.2 Precipitation Event Selection

Five selection criteria ensure diverse and representative precipitation conditions: (1) preferential selection during months with higher rainfall, (2) 80%–90% of events during rainy seasons (April–September), (3) time interval between events within the same month exceeding 48 hours, (4) hourly sliding average precipitation exceeding 10 mm, and (5) approximately even distribution of rainfall amounts.

Thirty precipitation events spanning 2018–2020 are selected (Figure 4). Results show 27 (90%) events occurred during April–September. Maximum sliding accumulated precipitation reaches 136.5 mm h^{-1}, with most events exceeding 50 mm h^{-1}.

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

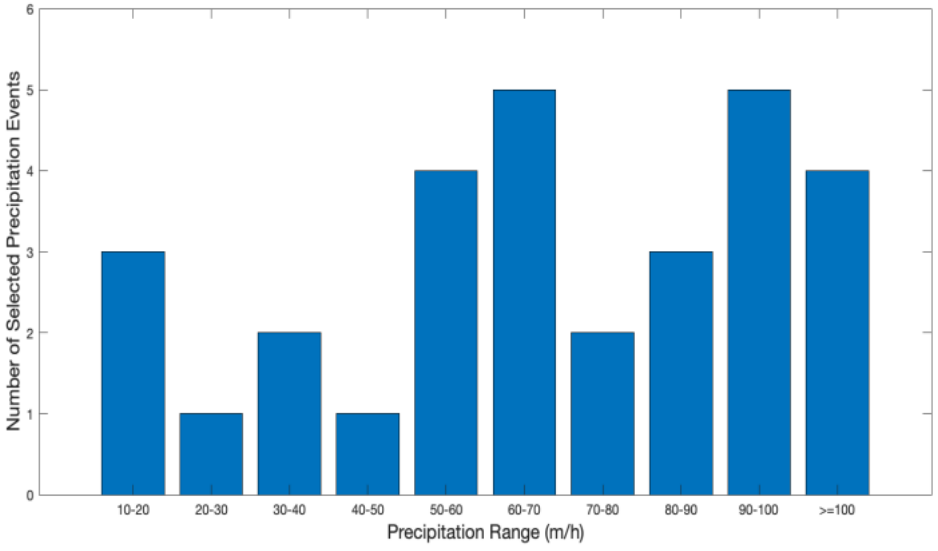


Figure 4. Month and precipitation distribution of the selected thirty precipitation events. There were 27 precipitation events (90%) from April to September, corresponding to the rainy seasons in the South China region. All thirty precipitation events exhibited hourly sliding accumulated precipitation greater than 10 mm h^{-1}.

3.1.3 Dataset Variables

The dataset contains comprehensive cloud microphysical variables (Table 1). Mass mixing ratios (kg/kg) include water vapor (QV), cloud water (QC), cloud ice (QI), rainwater (QR), snow (QS), and graupel (QG). Number mixing ratios include cloud ice (NI), rainwater (NR), and cloud water (NC). Aerosol concentrations (#/kg) include water-friendly (NWFA) and ice-friendly (NIFA) aerosols. Thermodynamic variables include temperature (T; K), pressure (P; Pa), vertical wind speed (W; m/s), and layer thickness (DZ; m). All variables are defined across 50 vertical layers.

3.1.4 Sample Independence Analysis

A correlation analysis using Pearson correlation coefficient validates dataset suitability. Since samples from different grid points are calculated independently, even adjacent grid columns can be considered independent. A block-based approach divides samples into blocks. The correlation coefficient ρ between samples X and Y is:

$$\rho = \frac{\sum_{i=1}^{D}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{D}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{D}(Y_i - \bar{Y})^2}} \tag{6}$$

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

Name	Description	I/O	Name	Description	I/O
QV	Water vapor mass mix. [kg/kg]	I/O	NR	Rain no. [#/kg]	I/O
QC	Cloud water mass mix. [kg/kg]	I/O	NC	Cloud water no. [#/kg]	I/O
QI	Cloud ice mass mix. [kg/kg]	I/O	NWFA	Water-friendly aerosol [#/kg]	I/O
QR	Rain mass mix. [kg/kg]	I/O	NIFA	Ice-friendly aerosol [#/kg]	I/O
QS	Snow mass mix. [kg/kg]	I/O	T	Temperature [K]	In
QG	Graupel mass mix. [kg/kg]	I/O	P	Pressure [Pa]	In
NI	Cloud ice no. [#/kg]	I/O	W	Vertical wind [m/s]	In

Table 1. Input and output variables of the WRF-4.2.1 Thompson scheme. Notation: x^c (mass), y (number/aerosol), x^e (env). I/O = input and output.

where $D = 50$ denotes the vertical dimension. Results (Figure 5) demonstrate exceptional sample independence: 83.36% of correlation values range from 0 to 0.1, with only 31 values exceeding 0.5 and maximum correlation of 0.7416, confirming high sample independence suitable for machine learning applications.

3.2 Data Preprocessing

Given minimal variation between consecutive time steps, prediction targets are set as incremental changes rather than absolute values. Preserving vertical structural consistency across 50 layers is crucial for maintaining physical characteristics.

Inputs are normalized using Min-Max normalization:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \tag{7}$$

where $X_{\max}$ and $X_{\min}$ are computed across the entire dataset.

Outputs are scaled using standardized normalization:

$$y_{\text{norm}} = y \times y_{\text{scale}} = (x_{t+1} - x_t) \times y_{\text{scale}}, \tag{8}$$

with scaling factors: QV–QG (10^7), NI (10^{-3}), NR (10^0), NC (10^{-8}), NWFA (10^{-7}), NIFA (10^{-2}), and T (10^3), x_{t+1} and x_t represent the variables at the current and next time step, respectively, which are "Input and Output" in Table 1.

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

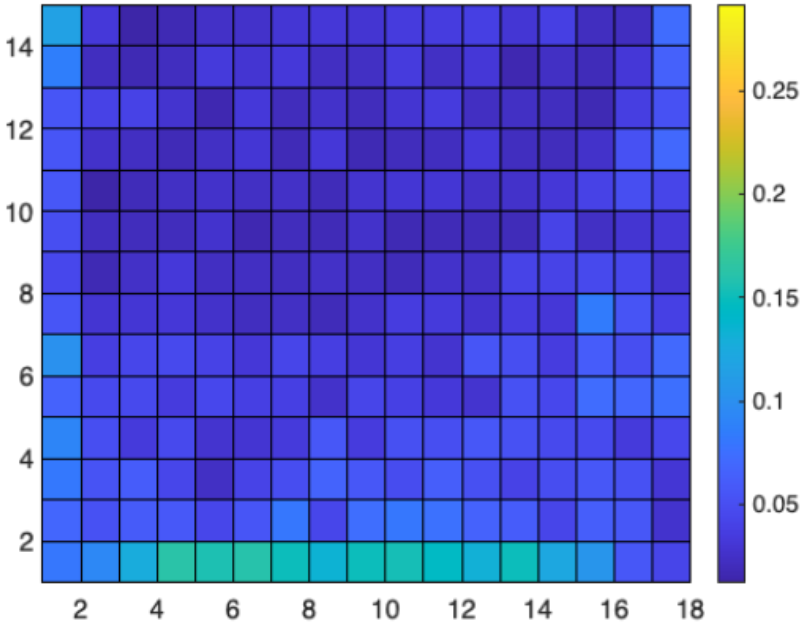


Figure 5. The pseudocolor plot and range distribution of all block correlation values from the whole dataset: (a) pseudocolor plot of element-averaged correlation value matrix and (b) range distribution. The analysis demonstrates that approximately 83.36% of correlation coefficient values range from 0 to 0.1, confirming strong sample independence suitable for machine learning applications.

3.3 Evaluation Metrics

Given substantial magnitude differences across predicted variables, Mean Squared Error (MSE) is adopted as the primary evaluation metric:

$$\mathrm{MSE}_j = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{D}\left(y_{\mathrm{true},i,j,k} - y_{\mathrm{pred},i,j,k}\right)^2, \tag{9}$$

where $y_{\mathrm{true},i,j,k}$ and $y_{\mathrm{pred},i,j,k}$ denote reference and predicted values for variable j at layer k of sample i, $D = 50$ is the vertical dimension, and N is the total number of test samples.

3.4 Benchmark Models and Training Protocol

We conduct comparative experiments using baseline models spanning different methodological approaches (Table 2). These benchmark models provide a comprehensive evaluation framework from classical machine learning (Linear, MLP) to modern deep learning architectures (Autoencoder, 1D-CNN, DenseNet, ResNet, Transformer, Inception). The 1D-CNN baseline serves as an important reference point, representing our proven methodology and providing the foundation for LMP-Tr.

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

Model	Pros / Cons	Model	Pros / Cons
Linear (1805)	+Simple −No nonlinear	DenseNet (2017)	+Reuse −High memory
MLP (1986)	+Nonlinear −No spatial	ResNet (2015)	+Deep −Limited atm.
Autoencoder (1987)	+Dim. reduct. −Not predict	Transformer (2017)	+Long-range −Compute-heavy
SVM (1995)	+Classic −Limit. complex	Inception (2014)	+Multi-scale −May overfit
1D-CNN (2024)	+Local −Limit. long-range		

Table 2. Benchmark models: + = strength, − = limitation.

The dataset is partitioned with an 80/20 training-validation split. An independent test set composed exclusively of 2020 precipitation events is used for evaluation. We randomly selected 60,000 samples from 20 events spanning 2018–2019 (3,000 samples per event) for training and validation. Training employs the Adam optimizer with initial learning rate 0.001, progressively reduced to 0.0001. Batch size is 64, and training runs for 500 epochs with early stopping to prevent overfitting.

4 Results

4.1 Effectiveness Evaluation

Performance evaluation across 12 cloud microphysical variables reveals substantial improvements with the proposed LMP-Tr variants. Quantitative results are summarized in Table 3, while Figure 6 visualizes performance profiles across all 12 evaluated models and variables, including baseline models (1D-CNN, Autoencoder, Linear, MLP, DenseNet, ResNet, Transformer, Inception), proposed models (LMP-Tr-1, LMP-Tr-2), and ablation variants (w/o Height Attention, w/o Cross Attention). The two proposed variants consistently outperform baseline methods and existing deep learning architectures.

Number concentration variables exhibit particularly strong improvements. Cloud water number concentration (NC) errors decrease from 0.188 to 0.120 (LMP-Tr-1) and 0.020 (LMP-Tr-2), with LMP-Tr-2 achieving an 89.4% reduction. Rain number concentration (NR) errors fall from 25.3 to 16.4 (LMP-Tr-1) and 15.7 (LMP-Tr-2), representing 35.2% and 38.0% reductions. Temperature (T) prediction errors decrease from 24.0 to 17.7 (LMP-Tr-1) and 15.8 (LMP-Tr-2), showing 26.3% and 34.2% improvements.

Performance differences between the two variants reflect distinct architectural characteristics and their alignment with physical process requirements. LMP-Tr-1, with its dual-pathway design, performs particularly well on ice-phase hydrometeors (QI, QS), where focused feature extraction captures well-defined patterns effectively. In contrast, LMP-Tr-2's three-pathway architecture shows advantages for precipitation-related variables (QR, QG) and temperature prediction, where multi-scale feature integration becomes critical. These observations suggest that architectural complexity should be matched to the underlying physical processes: simpler designs suffice for well-structured patterns, while complex interactions benefit from more sophisticated multi-scale representations.

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

Model	QV	QR	QI	QC	QS	QG	NI	NR	NC	NWFA	NIFA	T
1D-CNN	3.73	332.4	31.9	129.0	45.1	132.9	991.8	25.3	0.188	5.07	0.72	24.0
Autoencoder	3.68	274.9	20.1	89.4	45.7	112.0	902.0	16.9	0.310	5.60	2.03	22.5
Linear	5.20	441.5	19.7	149.2	81.9	116.4	943.4	19.4	0.330	5.43	0.97	29.3
MLP	7.80	484.7	27.8	240.2	173.2	172.4	949.5	23.0	0.620	6.15	1.34	44.2
DenseNet	3.38	308.7	27.2	115.6	41.3	122.4	938.5	23.2	0.168	5.02	0.68	21.8
ResNet	3.61	327.4	30.8	125.9	44.7	131.2	961.8	24.6	0.185	5.08	0.73	23.9
Transformer	3.42	312.3	26.5	119.7	42.1	127.8	948.2	23.5	0.172	5.03	0.66	22.1
Inception	3.74	335.6	31.4	130.2	46.3	133.7	970.1	25.0	0.189	5.09	0.75	24.5
LMP-Tr-1	**2.85**	**244.2**	**10.9**	**78.7**	**25.4**	**78.1**	**823.9**	**16.4**	**0.120**	**5.00**	**0.92**	**17.7**
LMP-Tr-2	**2.67**	**226.1**	**12.4**	**73.8**	**29.7**	**82.5**	**837.4**	**15.7**	**0.020**	**5.04**	**0.79**	**15.8**
w/o Height Attention	5.44	393.7	17.2	66.5	30.4	74.6	893.0	15.5	0.040	4.96	1.79	32.7
w/o Cross Attention	3.11	242.2	31.7	74.6	48.6	102.9	899.9	16.8	0.030	4.74	0.64	19.6

Table 3. Comparative effectiveness analysis of different models across cloud microphysical variables. MSE values are reported for each variable, with lower values indicating better performance.

Relative to the 1D-CNN baseline, both variants show consistent improvements across all variables. Water vapor (QV) errors decrease from 3.73 to 2.85 (LMP-Tr-1) and 2.67 (LMP-Tr-2), representing 23.6% and 28.4% reductions, respectively. Rainwater (QR) errors drop from 332.4 to 244.2 (LMP-Tr-1) and 226.1 (LMP-Tr-2), corresponding to 26.5% and 31.9% improvements. The most substantial gains occur in ice-phase hydrometeors. Cloud ice (QI) errors fall from 31.9 to 10.9 (LMP-Tr-1) and 12.4 (LMP-Tr-2), representing 65.8% and 61.1% reductions. Cloud water (QC) errors decrease from 129.0 to 78.7 (LMP-Tr-1) and 73.8 (LMP-Tr-2), showing 39.0% and 42.8% improvements. Snow (QS) errors reduce from 45.1 to 25.4 (LMP-Tr-1) and 29.7 (LMP-Tr-2), with 43.7% and 34.1% reductions. Graupel (QG) errors drop from 132.9 to 78.1 (LMP-Tr-1) and 82.5 (LMP-Tr-2), corresponding to 41.2% and 37.9% improvements.

4.2 Ablation Study

We conduct ablation studies by selectively removing critical components from LMP-Tr-2 (Table 3), examining two variants: w/o Height Attention and w/o Cross Attention.

Height-Dimension Attention Removal: Removal results in substantial accuracy degradation. QR experiences 74.1% MSE increase (393.7 vs. 226.1), QV shows 103.7% increase (5.44 vs. 2.67), QI increases by 38.7% (17.2 vs. 12.4), and QS by 2.4% (30.4 vs. 29.7). These results underscore the critical importance of height-dimension attention in capturing vertical atmospheric dependencies.

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

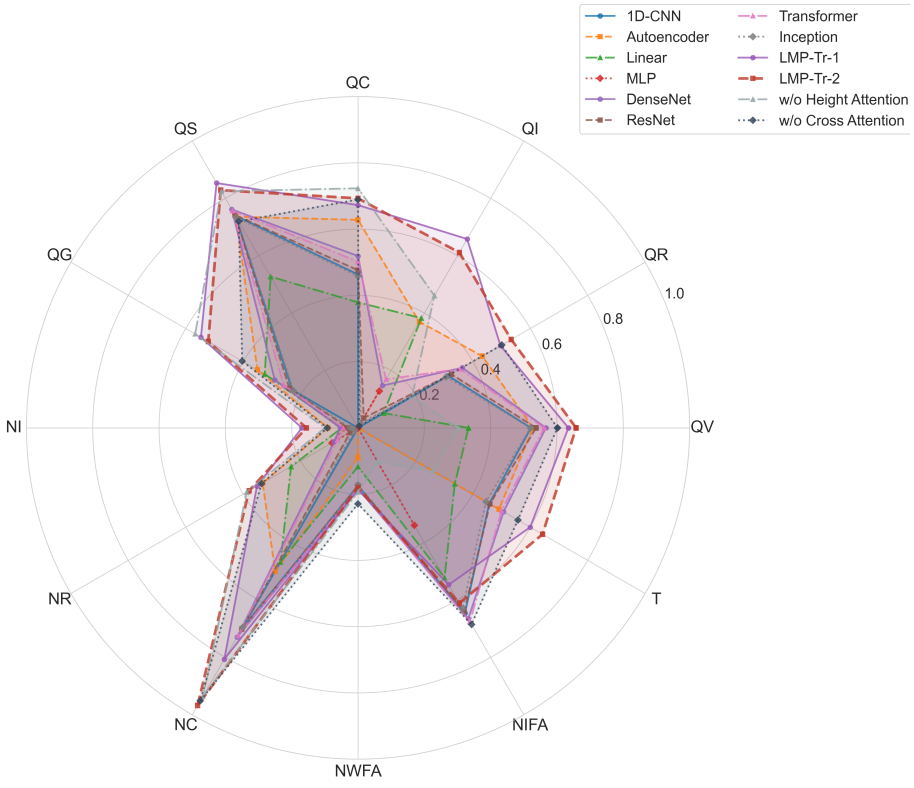


Figure 6. Radar chart comparing performance profiles of all evaluated models across 12 cloud microphysical variables. The normalized performance score (1-MSE/Max) is plotted for each variable, where MSE values are normalized relative to the maximum MSE across all models for each variable. Larger areas and higher scores (closer to 1.0) indicate better relative performance. All 12 models are included: baseline models, proposed models (LMP-Tr-1, LMP-Tr-2), and ablation variants (w/o Height Attention, w/o Cross Attention).

Cross-Attention Mechanism Removal: Ablation reveals severe impacts on variables requiring strong inter-variable coupling. QI experiences 155.6% MSE increase (31.7 vs. 12.4), QS shows 63.6% increase (48.6 vs. 29.7), and QG exhibits 24.7% degradation (102.9 vs. 82.5), highlighting the essential role of cross-attention in modeling processes that depend on multiple atmospheric variables.

Component Complementarity: Height-dimension attention is crucial for variables with strong vertical dependencies (QR, QV, NI), while cross-attention is essential for variables requiring complex inter-variable interactions (QI, QS, QG). Temperature prediction (T) benefits substantially from both components, showing 107.0% and 24.1% increases when height-dimension attention and cross-attention are removed, respectively.

The comparison between LMP-Tr-1 and LMP-Tr-2 further illustrates the importance of matching architectural complexity to task requirements. LMP-Tr-1 achieves superior performance in ice-phase variables (QI, QS) where focused feature extraction proves more effective, while LMP-Tr-2 excels in variables requiring complex multi-scale interactions (QR, QG, T, NC). The

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

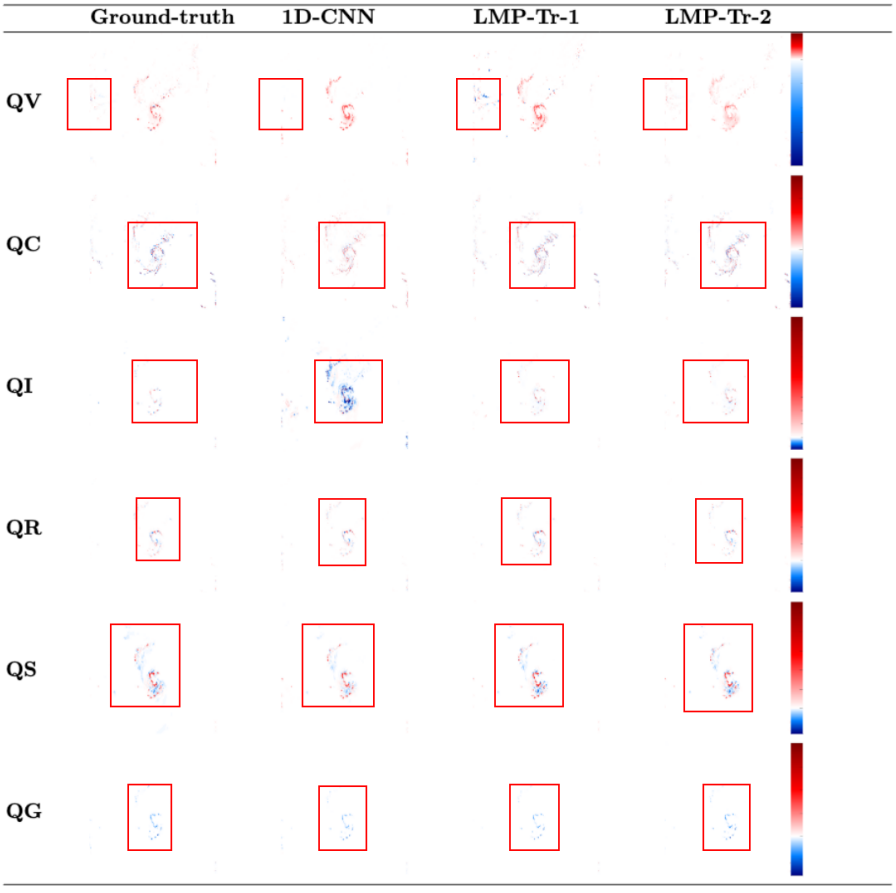


Figure 7. Comparison of cloud microphysical variable changes across different approaches.

complete LMP-Tr-2 achieves superior or competitive performance across all 12 variables, confirming that both mechanisms jointly contribute to model accuracy and generalization capacity.

4.3 Case Visualization

Figure 7 provides visual assessment of three methods (1D-CNN, LMP-Tr-1, LMP-Tr-2) against ground-truth data. The figure displays six cloud microphysical variables (rows: QV, QC, QI, QR, QS, QG) across four columns (Ground-truth, 1D-CNN, LMP-Tr-1, LMP-Tr-2). The reference data exhibit coherent spatial structures with sharp gradients and physically consistent distributions. Detailed analysis of key regions across different variables reveals distinct performance characteristics.

Water Vapor (QV): In the highlighted regions, ground-truth QV exhibits relatively weak amplitude changes with subtle local perturbations. The 1D-CNN baseline maintains low-amplitude distributions but fails to capture clearer local structures. LMP-Tr-1 shows localized enhancement responses, producing blue feature blocks not present in ground-truth, indicating over-

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

amplification tendencies. In contrast, LMP-Tr-2 effectively suppresses such local artifacts, with predictions closer to ground-truth in both amplitude levels and spatial distributions, demonstrating superior robustness in modeling weak-signal variables.

Cloud Water (QC): Ground-truth QC displays clear vortex-like structures with alternating positive and negative changes in the highlighted regions. The 1D-CNN baseline reconstructs overall morphology but exhibits weakened local contrast and smoothed detail structures. Both LMP-Tr-1 and LMP-Tr-2 preserve the spatial morphology and sign distribution of vortex structures well, with LMP-Tr-2 showing superior local contrast and background cleanliness, achieving higher consistency with ground-truth.

Cloud Ice (QI): This variable provides the most critical evidence of model performance. Ground-truth QI shows low-amplitude, spatially dispersed changes in the highlighted regions. However, the 1D-CNN baseline produces pronounced strong blue aggregation responses with significantly amplified amplitudes, deviating substantially from ground-truth and indicating local over-estimation problems. In contrast, both LMP-Tr-1 and LMP-Tr-2 effectively suppress such anomalous enhancement, bringing predictions closer to ground-truth in amplitude levels and spatial contours. LMP-Tr-2 results are more stable with reduced noise, demonstrating superior reconstruction accuracy and stability.

Rainwater (QR): Ground-truth QR displays weak and compact local structures in the highlighted regions, with relatively small differences among methods in overall morphology. The 1D-CNN baseline, LMP-Tr-1, and LMP-Tr-2 all reconstruct main spatial features. Comparatively, LMP-Tr-2 shows slight advantages in local structural continuity and background smoothness, exhibiting more stable consistency with ground-truth.

Snow (QS): Ground-truth QS exhibits relatively clear alternating positive-negative structures in the highlighted regions. The 1D-CNN baseline preserves overall structural morphology but shows weakened local contrast. Both LMP-Tr-1 and LMP-Tr-2 better maintain the clarity of positive-negative change structures, with LMP-Tr-2 results showing sharper boundaries and less background noise, making local change features closer to ground-truth.

Graupel (QG): Ground-truth QG displays fine-scale, continuous local change structures in the highlighted regions. The 1D-CNN baseline exhibits smoothing effects during reconstruction, slightly weakening local details. LMP-Tr-1 recovers fine structural morphology well, while LMP-Tr-2 further reduces background noise while maintaining structural continuity, bringing predictions closer to ground-truth in both spatial distribution and amplitude characteristics.

Overall, across key regions of all cloud microphysical variables, LMP-Tr-1 demonstrates marked improvements over the 1D-CNN baseline, while LMP-Tr-2 further enhances performance in suppressing anomalous responses, maintaining weak-signal consistency, and improving structural stability, achieving the highest overall consistency with ground-truth.

5 Conclusions

We present LMP-Tr (Learned Microphysics Transformer), an AI-based cloud microphysical parameterization scheme that integrates Transformer mechanisms with our prior 1D-CNN framework. LMP-Tr combines local feature extraction with global dependency modeling through alternating multi-scale convolutional modules and dual-pathway attention modules. We introduce two variants: LMP-Tr-1 with dual-pathway multi-scale convolutional architecture and LMP-Tr-2 with enhanced three-pathway

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

multi-scale convolutional architecture, offering different levels of architectural complexity to match diverse task requirements. Extensive experiments demonstrate that both variants deliver substantial accuracy gains across all twelve evaluated cloud microphysical variables, significantly outperforming traditional parameterization schemes and conventional machine learning approaches. The experimental results validate the effectiveness of the proposed architecture in capturing complex cloud microphysical processes, with LMP-Tr-1 demonstrating superior performance in ice-phase hydrometeors where focused feature extraction proves most effective, and LMP-Tr-2 showing enhanced capabilities in precipitation-related variables and temperature prediction where complex multi-scale feature integration is essential. These findings highlight the potential of AI-based parameterization schemes as another implementation approach for cloud microphysics parameterization, while demonstrating the importance of matching architectural complexity to physical process requirements. Future work should focus on online coupling with full numerical weather prediction models, long-term stability analysis, physical consistency validation, and generalization tests across diverse meteorological regimes.

Code and data availability. The large-scale training dataset used in this study was generated using the Weather Research and Forecasting (WRF) model version 4.2.1. Due to the massive volume of the full output files, the complete dataset is preserved via a reproducibility package rather than direct hosting. This package includes the model configuration files (namelist.wps and namelist.input), the specific version of the geography data, and the Python scripts for post-processing the raw outputs into the machine learning training set. All these materials, along with a representative sample of the processed data, are openly available at https://doi.org/10.5281/zenodo.19177453 (Shuting (2026)). The source code for LMP-Tr from this research is openly available at https://doi.org/10.5281/zenodo.20481658 (HUANG (2026)). The driving force fields were obtained from the ECMWF data portal (https://data.ecmwf.int/). The source code of WRF 4.2.1 is available at https://github.com/wrf-model/WRF/releases/tag/v4.2.1.

Competing interests. The authors declare there are no conflicts of interest for this manuscript.

Author contributions. YH and TS designed the study. YH developed the model, performed the experiments, analyzed the results, and wrote the original draft. QZ, HK, CSW, and HZ contributed to data analysis, interpretation, and manuscript revision. TS supervised the work, acquired funding, and revised the manuscript. All authors reviewed and approved the manuscript.

Acknowledgements. This work was supported by the Guangdong Basic and Applied Basic Research Foundation (2024A1515510031, 2023A1515011438), the National Natural Science Foundation of China (42105145), the Scientific Foundation for Youth Scholars of Shenzhen University (868-000001033384), and the China Meteorological Administration Youth Innovation Team (CMA2024QN01). This work is supported by the Intelligent Computing Center of Shenzhen University.

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

References

Alfonso, L. and Zamora, J. M.: A two-moment machine learning parameterization of the autoconversion process, Atmospheric Research, 249, 105 269, https://doi.org/10.1016/j.atmosres.2020.105269, 2021.

Arnold, C. and et al.: Efficient and stable coupling of the SuperdropNet deep-learning model for warm-rain microphysics with ICON v2.6.5, Geoscientific Model Development, 17, 4017–4032, https://doi.org/10.5194/gmd-17-4017-2024, 2024.

Beucler, T., Rasp, S., Pritchard, M., and Gentine, P.: Enforcing analytic constraints in neural networks emulating physical systems, Physical Review Letters, 126, 098 302, https://doi.org/10.1103/PhysRevLett.126.098302, 2021.

Brenowitz, N. D., Beucler, T., Pritchard, M. S., and Bretherton, C. S.: Interpreting and stabilizing machine-learning parametrizations of convection, Journal of the Atmospheric Sciences, 77, 4357–4375, https://doi.org/10.1175/JAS-D-20-0082.1, 2020.

Bryan, G. H., Morrison, H., and Fritsch, J. M.: Convective momentum transport and its sensitivity to microphysics in squall lines, Journal of the Atmospheric Sciences, 60, https://doi.org/10.1175/1520-0469(2003)060<1870:CMTAIS>2.0.CO;2, 2003.

Cotton, W. R., Pielke, R. A., Walko, R. L., Liston, G. E., Tremback, C. J., Jiang, H., and Harrington, J. Y.: New RAMS cloud microphysics parameterization. Part I: The single-moment scheme, Atmospheric Research, 38, 29–62, https://doi.org/10.1016/0169-8095(94)00087-T, 1995.

Ferrier, B. S., Jin, Y., Lin, Y., Black, T., Rogers, E., and DiMego, G.: Implementation of a new grid-scale cloud and precipitation scheme in the NCEP Eta model, Weather and Forecasting, 17, https://doi.org/10.1175/1520-0434(2002)017<0270:IOANGS>2.0.CO;2, 2002.

Gettelman, A., Morrison, H., Santos, S., Bogenschutz, P., and Caldwell, P.: Machine learning the warm rain process, Journal of Advances in Modeling Earth Systems, 13, e2020MS002 268, https://doi.org/10.1029/2020MS002268, 2021.

HUANG, Y.: LMP-Tr v1.0: Source code and experiment scripts for cloud microphysical parameterization, https://doi.org/10.5281/zenodo.20481658, 2026.

Jin, H.-G. and et al.: Do double-moment microphysics schemes make reliable raindrop number predictions?, Journal of Geophysical Research: Atmospheres, 128, e2022JD038 394, https://doi.org/10.1029/2022JD038394, 2023.

Köcher, G., Chwala, C., Siems-Anderson, A., and Kunstmann, H.: Influence of cloud microphysics schemes on weather model performance – case studies with WRF, Atmospheric Chemistry and Physics, 23, 6255–6282, https://doi.org/10.5194/acp-23-6255-2023, 2023.

Lin, Y.-L., Farley, R. D., and Orville, H. D.: Bulk parameterization of the snow field in a cloud model, Journal of Climate and Applied Meteorology, 22, 1065–1092, https://doi.org/10.1175/1520-0450(1983)022<1065:BPOTSF>2.0.CO;2, 1983.

Liu, Y., Yau, M. K., Shima, S.-I., Lu, C., and Chen, S.: Parameterization and explicit modeling of cloud microphysics: Approaches, challenges, and future directions, Advances in Atmospheric Sciences, 40, 747–790, https://doi.org/10.1007/s00376-022-2077-3, 2023.

Mansell, E. R., Ziegler, C. L., and Bruning, E. C.: Simulated electrification of a supercell thunderstorm with parameterized microphysics, Journal of the Atmospheric Sciences, 67, https://doi.org/10.1175/2009JAS2965.1, 2010.

Milbrandt, J. A. and Morrison, H.: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part III: Introduction of multiple free categories, Journal of the Atmospheric Sciences, 73, 975–995, https://doi.org/10.1175/JAS-D-15-0204.1, 2016.

Milbrandt, J. A. and Morrison, H.: A triple-moment representation of ice in the predicted particle properties (P3) bulk microphysics scheme, Journal of the Atmospheric Sciences, 78, https://doi.org/10.1175/JAS-D-20-0084.1, 2021.

Milbrandt, J. A. and Yau, M. K.: A multimoment bulk microphysics parameterization. Part I: Analysis of the role of the spectral shape parameter, Journal of the Atmospheric Sciences, 62, https://doi.org/10.1175/JAS3534.1, 2005a.

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

Milbrandt, J. A. and Yau, M. K.: A multimoment bulk microphysics parameterization. Part II: A proposed three-moment closure and scheme description, Journal of the Atmospheric Sciences, 62, https://doi.org/10.1175/JAS3535.1, 2005b.

Morrison, H. and Gettelman, A.: A new two-moment bulk stratiform cloud microphysics scheme in the Community Atmosphere Model (CAM3). Part I: Description and numerical tests, Journal of Climate, 21, https://doi.org/10.1175/2008JCLI2105.1, 2008.

Morrison, H. and Milbrandt, J. A.: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part I: Scheme description and idealized tests, Journal of the Atmospheric Sciences, 72, 287–311, https://doi.org/10.1175/JAS-D-14-0065.1, 2015.

Morrison, H., Curry, J. A., and Khvorostyanov, V. I.: A new double-moment microphysics parameterization for application in cloud and climate models. Part I: Description, Journal of the Atmospheric Sciences, 62, 1665–1677, https://doi.org/10.1175/JAS3446.1, 2005.

Morrison, H., van Lier-Walqui, M., Fridlind, A. M., and et al.: Confronting the challenge of modeling cloud and precipitation microphysics, Journal of Advances in Modeling Earth Systems, 12, e2019MS001 689, https://doi.org/10.1029/2019MS001689, 2020.

Mu, B., Chen, L., Yuan, S., and Qin, B.: A radiative transfer deep learning model coupled into WRF with a generic Fortran Torch Adaptor, Frontiers in Earth Science, 11, 1149 566, https://doi.org/10.3389/feart.2023.1149566, 2023.

Ren, X., Li, X., Ren, K., Song, J., Xu, Z., Deng, K., and Wang, X.: Deep learning-based weather prediction: A survey, Big Data Research, 23, 100 178, https://doi.org/10.1016/j.bdr.2020.100178, 2021.

Rodríguez-Genó, C. F., Torri, G., and Kurowski, M. J.: Parameterization of the collision–coalescence process using basis functions, Geoscientific Model Development, 15, 493–510, https://doi.org/10.5194/gmd-15-493-2022, 2022.

Segele, Z. T., Leslie, L. M., Yigzaw, W., and Anquetin, S.: Data assimilation and microphysics sensitivity experiments using the WRF model for a heavy rainfall event over Ethiopia, Tellus A: Dynamic Meteorology and Oceanography, 65, 19 599, https://doi.org/10.3402/tellusa.v65i0.19599, 2013.

Seifert, A. and Beheng, K. D.: A two-moment cloud microphysics parameterization for mixed-phase clouds: Part 1. Model description, Meteorology and Atmospheric Physics, 92(1–2), 45–66, https://doi.org/10.1007/s00703-005-0112-4, 2006.

Seifert, A. and Rasp, S.: Potential and limitations of machine learning for modeling warm-rain cloud microphysical processes, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 301, https://doi.org/10.1029/2020MS002301, 2020.

Sharma, S. and et al.: SuperdropNet: A stable and accurate machine learning emulator of warm-rain microphysics, Journal of Advances in Modeling Earth Systems, 17(—), e2024MS004 279, https://doi.org/10.1029/2024MS004279, 2025.

Shi, X., Wang, H., and Liu, X.: Two-moment bulk stratiform cloud microphysics in the grid-point atmospheric model of IAP LASG, Advances in Atmospheric Sciences, 30, 971–986, https://doi.org/10.1007/s00376-012-2072-1, 2013.

Shu, T., Zhou, X., Xie, Y., Zhao, H., Chen, X., and Wang, R.: Deep learning based cloud microphysics parameterization scheme for numerical weather prediction with 1DD-CNN, in: Proceedings of the 2024 7th International Conference on Artificial Intelligence and Big Data (ICAIBD), pp. 1–6, Chengdu, China, https://doi.org/10.1109/ICAIBD62003.2024.10604645, 2024.

Shuman, F. G.: Numerical weather prediction, Bulletin of the American Meteorological Society, 59, 5–17, https://doi.org/10.1175/1520-0477-59.1.5, 1978.

Shuting: Shuting/Generating-MPS-Dataset-via-WRF-4.2.1: GeneratingMPSDatasetViaWRF4.2.1, https://doi.org/10.5281/zenodo.19177453, 2026.

Takeishi, A., Wang, C., and others.: Parameterizing raindrop formation using machine learning in a cloud microphysics model, Monthly Weather Review, 152, https://doi.org/10.1175/MWR-D-22-0175.1, 2024.

https://doi.org/10.5194/egusphere-2026-3243
Preprint. Discussion started: 18 June 2026

Thompson, G., Field, P. R., Rasmussen, R. M., and Hall, W. D.: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization, Monthly Weather Review, 136, https://doi.org/10.1175/2008MWR2387.1, 2008.

Weyn, J. A., Durran, D. R., and Caruana, R.: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 109, https://doi.org/10.1029/2020MS002109, 2020.

Yuval, J. and O'Gorman, P. A.: Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions, Nature Communications, 11, 3295, https://doi.org/10.1038/s41467-020-17142-3, 2020.

Zhong, X. and et al.: WRF–ML v1.0: A bridge between WRF v4.3 and machine learning parameterizations and its application to atmospheric radiative transfer, Geoscientific Model Development, 16, 199–209, https://doi.org/10.5194/gmd-16-199-2023, 2023.