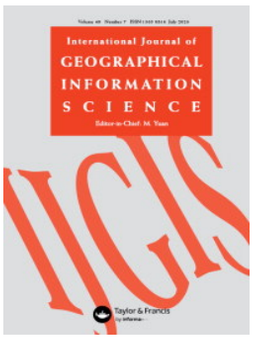

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

# Spatiotemporal adaptive multiscale transformer for prediction

Lai Chen, Zeqiang Chen, Yongze Song, Chao Yang, Sijia He, Zhiqing Li, Wenfeng Guo & Nengcheng Chen

**To cite this article:** Lai Chen, Zeqiang Chen, Yongze Song, Chao Yang, Sijia He, Zhiqing Li, Wenfeng Guo & Nengcheng Chen (2026) Spatiotemporal adaptive multiscale transformer for prediction, International Journal of Geographical Information Science, 40:7, 2475-2499, DOI: 10.1080/13658816.2025.2602536

**To link to this article:** https://doi.org/10.1080/13658816.2025.2602536

Published online: 08 Jan 2026.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2026, VOL. 40, NO. 7, 2475–2499
https://doi.org/10.1080/13658816.2025.2602536

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE



# Spatiotemporal adaptive multiscale transformer for prediction

Lai Chen[a,b], Zeqiang Chen[a], Yongze Song[b], Chao Yang[a], Sijia He[a], Zhiqing Li[a], Wenfeng Guo[c] and Nengcheng Chen[a]

[a]National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan, China; [b]School of Design and Built Environment, Curtin University, Perth, Australia; [c]Hubei Institute of Land Surveying and Mapping, Wuhan, China

**ABSTRACT**

Spatiotemporal processes, such as floods, rainfall-runoff, and land-use changes, continuously evolve over space and time with high dynamism and complex nonlinearity. Accurate and efficient spatiotemporal process prediction is crucial for understanding their underlying patterns. Recently, deep learning has effectively addressed spatiotemporal prediction issues in Earth science. However, most existing studies address either short-term or long-term dependencies, but ignore the multiscale characteristics and spatial heterogeneity inherent to spatiotemporal processes and critical for practical applicability. This study develops a Spatiotemporal Adaptive Multiscale Transformer (SAMT) model for spatiotemporal process prediction. First, we design an enhanced multiscale spatial heterogeneity module to extract multiscale spatial heterogeneity. Then, we introduce the adaptive scale selection that assigns weights to features at different scales based on their contributions. In addition, we incorporate a spatiotemporal transformer block to simultaneously capture short-term and long-term dependencies. We conduct extensive experiments on three representative spatiotemporal datasets of rainfall, temperature, and flood. Compared to state-of-the-art models, the SAMT model achieves significant improvements across all evaluation metrics. The developed SAMT model critically improves the performance of spatiotemporal process prediction for more accurate and effective modelling of spatiotemporal evolution patterns in the field of Earth sciences.

**ARTICLE HISTORY**
Received 14 May 2025
Accepted 8 December 2025

**KEYWORDS**
Spatiotemporal processes; multiscale; spatial heterogeneity; rainfall; flood inundation

## 1. Introduction

Spatiotemporal processes refer to the dynamic evolution of various geographic phenomena in space over time, such as flood inundation, rainfall-runoff, and land use changes (Zheng *et al*. 2022, Yao *et al*. 2023). These processes are not only highly dynamic and nonlinear but also characterized by multiscale characteristics and spatial heterogeneity. (Song *et al*. 2020, Hu *et al*. 2024). Meanwhile, these inherent geographic

**CONTACT** Zeqiang Chen chenzeqiang@cug.edu.cn; Yongze Song Yongze.Song@curtin.edu.au

characteristics present significant challenges in uncovering knowledge and identifying patterns within spatiotemporal processes (Song 2022, Ren *et al*. 2025). As a critical approach to uncovering spatiotemporal process patterns, accurate and efficient spatiotemporal process prediction plays a crucial role in understanding these patterns (Panahi *et al*. 2021, Speight *et al*. 2021).

Spatiotemporal process prediction aims to predict future sequence variations based on historical observation data, emphasizing the integration of temporal and spatial dynamics (Ge *et al*. 2019). Current approaches include physics-based models, shallow machine learning models, and deep learning models (Brunner *et al*. 2021). Physics-based models constitute a traditional yet widely adopted approach for spatiotemporal process prediction and employ complex mathematical equations to describe and quantify spatiotemporal dynamics. For different spatiotemporal processes, researchers have developed various models. For example, the Xin'anjiang model (Lü *et al*. 2013), the variable infiltration capacity model (Su *et al*. 2024), and the soil and water assessment tool model (Zhao *et al*. 2024) are commonly employed for flood simulation and prediction. In addition, numerical weather prediction model and optical flow-based algorithms are utilized for radar-based precipitation nowcasting. Although physics-based models are effective in predicting flood inundation and rainfall-runoff processes, they typically require extensive input data for model calibration and parameter adjustment. Moreover, the interdependencies among parameters make these methods both computationally expensive and time-consuming.

Compared to physics-based models, shallow machine learning models can address simple nonlinear problems without explicitly modelling the physical process. Early methods included linear regression (Zhang *et al*. 2018), autoregressive models, and their variants (Pulukuri *et al*. 2018). Subsequently, more advanced techniques, including support vector machines (Campolo *et al*. 1999), random forest (Tang *et al*. 2021), decision trees (Bui *et al*. 2019), and extreme learning machines (Adnan *et al*. 2019), have been applied to spatiotemporal process prediction. However, the inherent complexity and variability of spatiotemporal processes present significant challenges for machine learning methods in capturing temporal, spatial, and spatiotemporal dependencies (Song *et al*. 2025). Their relatively simple architectures limit their ability to handle highly nonlinear relationships.

Recently, deep learning has gained significant attention in spatiotemporal process prediction. The ability of deep learning to automatically extract features, provide strong nonlinear representation, and provide diverse network architectures has shifted research from physics-based and traditional machine learning models toward deep learning–based intelligent methods. For example, temporal dynamics are modelled using RNNs (Sadeghi Tabas *et al*. 2023) and LSTMs (Graves 2012), while spatial relationships are effectively captured by CNNs (Gu *et al*. 2018), GNNs (Li *et al*. 2018), and attention mechanisms (Zhu *et al*. 2021). Current research is focusing on integrating temporal and spatial representation models to capture the spatiotemporal dynamics, with approaches like ConvLSTM (Xu *et al*. 2024), PredRNN (Wu *et al*. 2022), and SAST-GCN (Jin *et al*. 2025) being employed to model these complex processes. These deep learning methods have significantly advanced the field of spatiotemporal process prediction.

However, the high dynamism, multiscale characteristics, and spatial heterogeneity of spatiotemporal processes pose significant challenges for accurate prediction (Zhang *et al*. 2021). First, existing methods usually ignore the inherent multiscale nature of spatiotemporal processes and typically capture interactions at a single scale when using multiscale contextual information (Zhang *et al*. 2024). This results in the loss of crucial spatiotemporal information and insufficient feature representation. Second, from a temporal perspective, attribute values at a given spatial location evolve over time, while from a spatial perspective, attribute values exhibit variations across different locations at the same time. The combined effect of these temporal and spatial variations gives rise to spatiotemporal heterogeneity. Despite these challenges, current research on spatiotemporal process prediction has yet to adequately explore these issues.

To overcome these challenges, we develop a spatiotemporal adaptive multiscale transformer (SAMT) for spatiotemporal process prediction. Given the Transformer's strong capability for modeling long-range dependencies, which has led to its successful application in object detection, image segmentation, and video recognition tasks (e.g., ConvFormer (Gu *et al*. 2023), PKI-net (Cai *et al*. 2024), and Uniformer (Li *et al*. 2023)), its application in spatiotemporal process prediction still requires further development. SAMT model adopts Transformer architecture as the fundamental framework to effectively capture spatiotemporal dependencies. Specifically, we design an enhanced multiscale spatial heterogeneity block that integrates both multiscale characteristics and spatial heterogeneity to improve feature extraction. In addition, we design an adaptive scale selection block, which dynamically selects the most informative multiscale feature representation for spatiotemporal prediction, assigning adaptive weights to different scales to enhance model prediction performance.

The study makes the following key contributions. First, we propose an enhanced multi-scale spatial heterogeneity module to capture spatial patterns at multiple scales. This module strengthens the representation capability of spatiotemporal process models by explicitly modeling complex spatial variations that are usually ignored in the exsiting studies involving multi-scale spatial heterogeneity. Second, we introduce a scale-adaptive selection mechanism that dynamically assigns weights to features at different scales based on their relative importance. In contrast to conventional approaches that simply sum multi-scale features, this mechanism enables the model to effectively investigate multi-scale spatial heterogeneity in a data-driven manner and improve prediction performance. Third, we replace the feed-forward network in the standard Transformer block with a 3D CNN to jointly capture both short-term and long-term spatiotemporal dependencies. This integration further enhances the model's ability to represent intricate spatiotemporal dynamics compared with traditional architectures. Finally, extensive experiments conducted on three spatiotemporal datasets demonstrate that the SAMT model is consistently better than state-of-the-art models due to the effectiveness of its hybrid architecture and adaptive multiscale modeling in spatiotemporal process prediction.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 describes the proposed spatiotemporal process prediction model. Section 4 presents the experimental setup, results, and analysis. Section 5 concludes the paper and outlines future work.

## 2. Related work

The spatiotemporal process prediction aims to predict future sequences based on historical sequence data from a specific region. The current research on spatiotemporal process prediction can generally be categorized into three main categories: physics-based methods, shallow machine learning methods, and deep learning methods (Brunner *et al*. 2021).

Physics-based methods describe spatiotemporal processes using mathematical formulations grounded in physical laws. Noori and Kalin (2016) employed the SWAT to predict the daily rainfall process across nearly 29 watersheds. Leskens *et al*. (2014) utilized different flood models to assess the predictive performance of the physical model in capturing flood dynamics. Bližňák *et al*. (2017) employed an extrapolation-based model to forecast rainfall patterns. Although physics-based approaches provide notable advantages, their reliance on high-resolution input data and computationally intensive, inefficient processing significantly constrains their scalability and applicability in spatiotemporal process forecasting.

As a viable alternative, shallow machine learning methods can model the nonlinear interactions among available data and construct input-output relationships without necessitating an in-depth comprehension of the underlying physical characteristics. Yan *et al*. (2018) developed a physics-based model to simulate data and constructed two support vector machine models to predict flood warning and maximum flood depth, respectively. Adnan *et al*. (2019) investigated a novel heuristic approach and predicted the daily streamflow process using an extreme learning machine model. Li *et al*. (2016) utilized a random forest algorithm to forecast daily variations in lake water levels. Compared to physics-based models, machine learning methods have demonstrated improvements in predictive performance. However, due to their relatively simple structure, these models usually prioritize temporal information while ignoring spatial dependencies when handling multiscale, high-dimensional data. In addition, they face challenges in effectively capturing the intricate nonlinear relationships in spatiotemporal processes.

Deep learning-based methods are black-box models that can establish end-to-end relationships between inputs and outputs. Leveraging their powerful nonlinear fitting capabilities, deep learning approaches have become valuable tools for spatiotemporal process prediction. Shi *et al*. (2015) initially proposed the ConvLSTM model for prediction, which integrates CNN and LSTM to effectively capture spatiotemporal features, marking a shift from modelling temporal dependencies to modelling spatiotemporal relationships. Later, Wang *et al*. (2017) extended this idea by proposing PredRNN, which incorporates a Spatiotemporal LSTM unit capable of simultaneously modelling spatial and temporal features within a unified unit. Furthermore, Wang *et al*. (2018) proposed the PredRNN++ model for spatiotemporal process prediction, which introduces the causal LSTM unit and combines temporal and spatial structures in a concatenated form to better capture short-term dependency features. Wang *et al*. (2019) proposed a Memory in Memory (MIM) network, which improves the forget gate in the ST-LSTM unit for modelling both stationarity and non-stationarity in spatiotemporal dynamics. To enhance the ability to capture spatiotemporal relationships, Gao *et al*. (2022) developed the SimVP model, which incorporates a gated spatiotemporal attention transformer for improved representation learning. In addition, Some studies have adopted attention mechanisms for their effectiveness in capturing long-range spatial and temporal correlations through global context modelling

(Xiong *et al*. 2021, Liu *et al*. 2022, Tang *et al*. 2024). Luo *et al*. (2021) proposed IDA-LSTM, which incorporates an interaction framework and dual attention mechanisms to improve rainfall prediction. Tang *et al*. (2023) introduced SwinLSTM by embedding Swin Transformer blocks into a simplified LSTM structure, effectively capturing global spatiotemporal dependencies. Tang *et al*. (2024) presented PredFormer, a recurrent-free Transformer framework that achieves improvements in both prediction accuracy and computational efficiency. Seo *et al*. (2023) proposed IAM4VP for weather–climate prediction that integrates the strengths of both autoregressive and non-autoregressive approaches. Li *et al*. (2023) propose UniFormer, a unified transformer architecture that integrates convolution and self-attention to efficiently handle both local redundancy and global dependency. He *et al*. (2025) developed STMixGAN, a radar-based precipitation nowcasting model that effectively captures spatiotemporal rainfall evolution and outperforms conventional and deep learning–based methods.

In summary, deep learning methods primarily focus on modelling short- and long-term spatiotemporal but usually ignore the multiscale characteristics and spatial heterogeneity in spatiotemporal variations. The complex linear relationships within spatiotemporal processes remain insufficiently explored. In response to the limitations, we present an adaptive multiscale Transformer model. The proposed approach offers a more reliable and practical method for accurately predicting spatiotemporal processes.

## 3. Methodology

This section presents a comprehensive overview of the proposed Spatiotemporal Adaptive Multiscale Transformer (SAMT) model. The SAMT model comprises five key components (see Figure 1): a shallow feature block, an enhanced multiscale spatial heterogeneity feature extraction block, an adaptive scale selective block, a spatiotemporal feature extraction block and a prediction block.

The shallow feature extraction block initializes input features using a 2D CNN layer. The remaining, more complex components are detailed in the subsequent subsections.

### 3.1. Preliminaries

Before introducing the detailed methodology, we first provide a statement of the spatiotemporal process prediction problem. The spatiotemporal process prediction involves forecasting future sequences based on historical observed sequences within a spatial region. Formally, given an observation sequence of length $T$, it is represented as $\boldsymbol{X} = (x_1, x_2, ..., x_T)$, where $x_T \in \mathbb{R}^{H\times W\times C}$. H and W correspond to the spatial region, and C is the channel numbers of the feature map. The prediction task aims to produce a future sequence $\hat{Y}$ of length $T^{'}$ by maximizing its probability, which can be defined using Eq. (1) (Shi *et al*. 2015):

$$\hat{y}_{t+1}, \hat{y}_{t+2}, ..., \hat{y}_{t+T'} = \underset{y_{t+1}, y_{t+2}, ..., y_{t+T'}}{\arg\max} \mathcal{P} \quad = y_{t+1}, y_{t+2}, ..., y_{t+T'} X\} \tag{1}$$

where $\mathcal{P}(\cdot|\cdot)$ denote the conditional probability. The predicted sequence $\hat{Y}$ is set of $\hat{y}_{t+1}, \hat{y}_{t+2}, ..., \hat{y}_{t+T'}$, which is expressed as $\hat{Y} = (\hat{y}_{t+1}, \hat{y}_{t+2}, ..., \hat{y}_{t+T'})$, where $\hat{y}_i$ has the same shape as $x_i$.

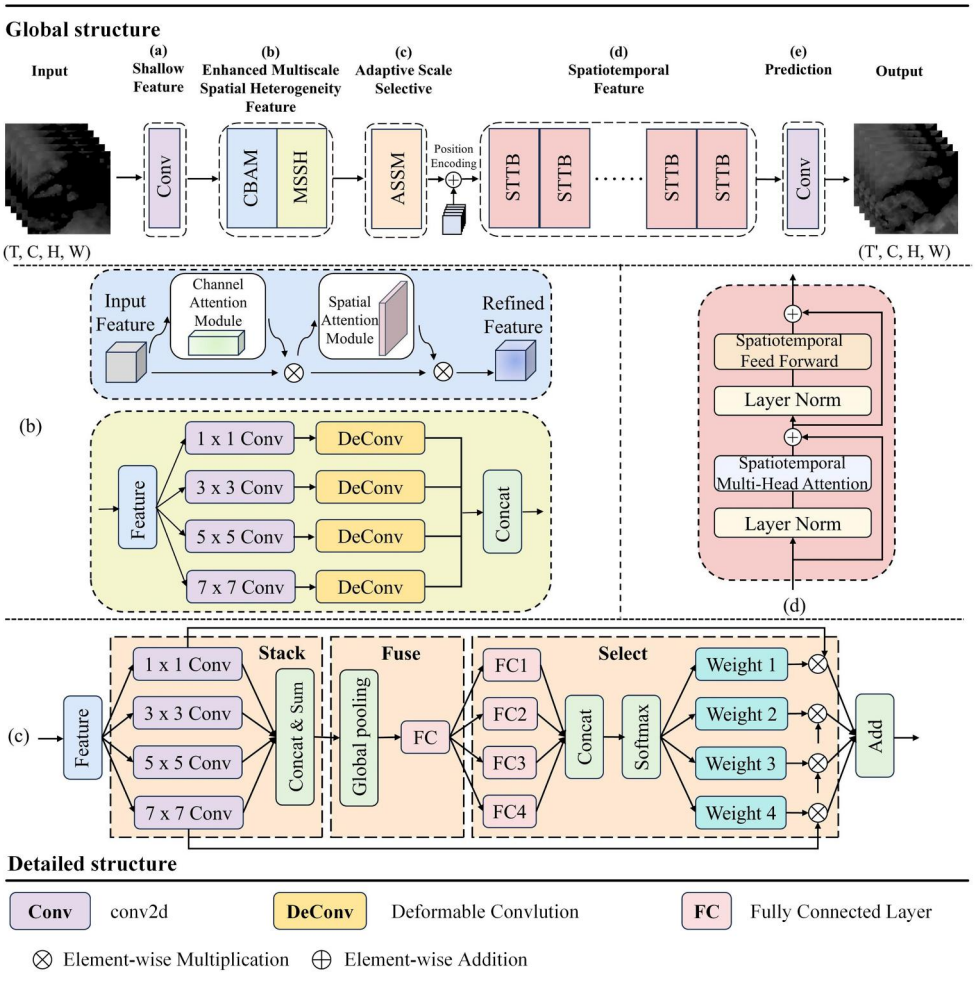


**Figure 1.** Overall framework of the proposed Spatiotemporal Adaptive Multiscale Transformer (SAMT) model. (a) Shallow feature block, (b) top: convolution block attention module (CBAM), bottom: Multiscale Spatial Heterogeneity Module (MSHM), (c) Adaptive Scale Selection Module (ASSM), (d) Spatiotemporal Transformer Block (STTB), (e) Prediction Block.

## *3.2. Adaptive multiscale transformer network with spatial heterogeneity*

### *3.2.1. Enhanced multiscale spatial heterogeneity feature extraction*

Spatial heterogeneity is widely present in various geographic processes and is one of the most common spatial phenomena (Li and Reynolds 1995, Fotheringham and Sachdeva 2022). Spatial heterogeneity refers to variations in spatial attributes across different locations. In addition, spatial heterogeneity varies across different scales, tending to decrease as the spatial scale increases (Riera and Magnuson 1998). This makes spatiotemporal feature extraction challenging. To address this problem, inspired by inception Architecture (Szegedy *et al.* 2017) and deformable convolution (Zhu *et al.* 2018), the enhanced multiscale spatial heterogeneity feature extraction block (EMSHB) was proposed.

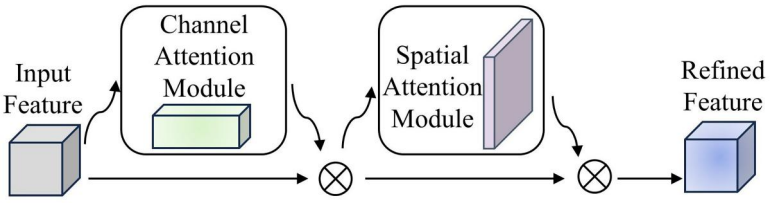


**Figure 2.** The architectural structure of the convolutional block attention module (CBAM).

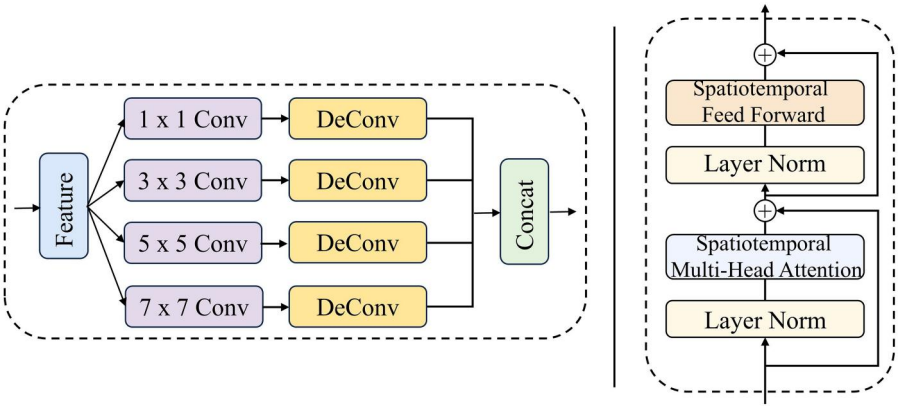


**Figure 3.** Structure diagram of MSHM (left) and STTB (right). where $\oplus$ indicates element-wise addition.

As a core component of SAMT, the EMSHB incorporates both the Convolutional Block Attention Module (CBAM) and the Multiscale Spatial Heterogeneity Module (MSHM). CBAM is a lightweight attention module composed of two submodules: channel attention and spatial attention, with its structure shown in Figure 2. CBAM can filter out unimportant information from feature maps while retaining important information, without increasing the number of parameters or computational cost (Yin *et al*. 2023). When the initialised input feature is $F \in R^{C\times T\times H\times W}$, the input feature map is sequentially processed by the channel and spatial attention modules to compute the corresponding weights. These weights are then applied to the input via element-wise multiplication to generate the enhanced feature map $F_{enh} \in R^{C\times T\times H\times W}$. The process can be expressed by Equations (2) and (3) (Woo *et al*. 2018).

$$F' = M_c(F) \otimes F \tag{2}$$

$$F_{enh} = M_s(F') \otimes F' \tag{3}$$

To capture spatial heterogeneity at different scales in spatiotemporal processes, the MSHM is designed. As shown in Figure 3 (left), the enhanced spatiotemporal features from the previous step are first processed through different convolution layers. These outputs are then passed through a shared deformable convolution operation, followed by feature fusion across different scales. This process can be expressed as follows:

$$M_{mshm}(F) = \text{Concat}(\text{DeConv}(\text{Conv}_k(F_{enh}))) \tag{4}$$

where $k \in \{1 \times 1, 3 \times 3, 5 \times 5, 7 \times 7\}$, $F_{enh}$ represents the enhanced feature, Conv represents the convolution operation, DeConv represents the deformation convolution operation and Concat represents the feature connection operation.

Specifically, the MSHM utilises convolution kernels of sizes 1, 3, 5, and 7 to capture spatial heterogeneity at multiple scales. Unlike standard convolution kernels that extract features using only regular windows, which significantly limits their ability to capture heterogeneous information, we introduce deformable convolution inspired by the deformable convolution network. MSHM can learn offset parameters for adaptive extraction of heterogeneous spatial information, which is expressed in Equation (5) (Zhu *et al*. 2018).

$$\mathrm{DeConv}(x) = \sum_{p_n \in \mathcal{R}} w(p_n) * x(p_0 + p_n + \Delta p_n) \tag{5}$$

Taking the 3 x 3 convolution as an example, the grid $\mathcal{R}$ with the receptive field is defined as $\{(-1, -1), (-1, 0), (-1, 1), \ldots, (1, -1), (1, 0), (1, 1)\}$. Given an input feature map $x$, the convolution output at location $p_0$ is computed as a weighted sum over locations $(p_0 + p_n + \Delta p_n)$, where $p_n \in \mathcal{R}$ and $\Delta p_n$ represents the learnable offset. It is worth noting that $\Delta p_n$ is typically a fractional value, requiring bilinear interpolation to determine the new feature value at the location of $(p_0 + p_n + \Delta p_n)$.

### 3.2.2. Adaptive scale selective module

After obtaining multiscale spatial heterogeneity features, directly adding or concatenating features from different scales usually assigns equal weights to all scales, ignoring their varying importance in spatiotemporal process prediction (Gao *et al*. 2023). Since the contribution of spatial heterogeneity features at different scales varies, it is essential to assign weights adaptively rather than treating all scales equally. As shown in Figure 4, to fully leverage multiscale spatial heterogeneity features, we design an adaptive scale selection module (ASSM) that enables the model to autonomously learn weight parameters during training and assign different weights based on the contribution of heterogeneous spatial features at each scale. To facilitate understanding, this process can be simplified as Equation (6).

$$F_{out} = \sum_{i=1}^{n} w_i \times F_i, \quad \sum_{i=1}^{n} w_i = 1 \tag{6}$$

where $w_i$ represents adaptive weight, $n$ represents the number of scale features, $F_i$ represents the feature at different scales, and $F_{out}$ represents the feature output after weighted fusion.

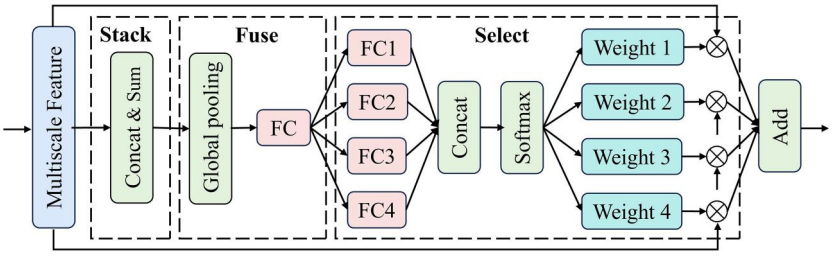


**Figure 4.** The process of adaptive scale selection module (ASSM). FC denotes a fully connected layer, and $\otimes$ represents element-wise multiplication.

Specifically, the adaptive selection of features at more scales can be extended in the three example scales. The ASSM consists of three stages: stacking, fusing, and selecting. The first stage is stacking input features that are passed through different convolution layers to obtain features at four distinct scales. These features are then stacked along a newly introduced dimension, referred to as the scale dimension, resulting in $F_{stack}$. The next stage is fusing features from different scales using element wise addition. A global average pooling operation is then applied to incorporate global contextual information, yielding $F_{global}$, which contains channel-wise statistical information. Subsequently, $F_{global}$ is transformed into a compact representation $F_{compact}$ via a fully connected (FC) layer. The last stage is selecting weights for multi-scale information. To facilitate precise adaptive scale selection and generate weights for each scale, the FC layers matching the number of scales is applied to $F_{compact}$, resulting in $F_1$, $F_2$, $F_3$ and $F_4$. These generated features are concatenated along the scale dimension, aligning with the dimension of $F_{stack}$. The softmax function is then employed to generate scale-specific weights $W$, enabling the model to adaptively select relevant spatial heterogeneity features across different scales. Finally, the weights $W$ are multiplied by $F_{stack}$, applying distinct weights to the feature at each scale, and the resulting features are aggregated through a summation operation to yield the output feature $F_{adaptive}$ after adaptive selection.

### 3.2.3. Spatiotemporal feature extraction

To model spatiotemporal dependencies between different frames in the spatiotemporal process, we employ a transformer block with a 3D convolution operation, referred to as the spatiotemporal transformer block (STTB), as illustrated in Figure 3 (right). In STTB, 3D convolution is employed to extract short-term dependencies, while the transformer mechanism is designed to model long-term dependencies. Compared to stacking multiple ConvLSTM blocks to model long-term dependencies, the transformer component in STTB leverages a multi-head attention mechanism that more effectively captures long-range dependencies and alleviates the gradient vanishing problem.

Specifically, STTB is composed of two key sublayers: a multi-head attention layer (MHA) and a feed-forward network (FFN) layer (Vaswani *et al*. 2017). A residual connection is incorporated between these two sublayers with a layer normalization (LN) applied before each sublayer, as described in Yang *et al*. (2022). First, $F_{adaptive}$ is processed through layer normalization, yielding $\hat{F}_{adaptive} = \mathrm{LN}\left(F_{\mathrm{adaptive}}\right)$. Then it is fed into the spatiotemporal multi-head attention layer, where 3D convolution operations are used to obtain Q, K and V, formulated as follows:

$$Q = \mathrm{Conv3D}_Q\left(\hat{F}_{adaptive}\right) \tag{7}$$

$$K = \mathrm{Conv3D}_K\left(\hat{F}_{adaptive}\right) \tag{8}$$

$$V = \mathrm{Conv3D}_V\left(\hat{F}_{adaptive}\right) \tag{9}$$

where Conv3D represents 3D convolution operation. Compared to the linear transformation in the standard Transformer block, applying 3D convolution allows for capturing short-term dependency in the spatiotemporal process. Meanwhile, the long-term dependencies are captured based on the principles of the attention mechanism, formulated as

follows (Vaswani *et al*. 2017):

$$F_{MHA} = \text{Softmax}\left(\frac{QK^T}{\sqrt{D}}\right)\text{V} \tag{10}$$

where $F_{MHA}$ represents the feature map after the MHA layer, and Softmax represents the softmax function.

Further, the residual connection is applied to obtain:

$$F_{attention} = F_{MHA} + F_{adaptive} \tag{11}$$

After that, the final output of the STTB is obtain as follows:

$$F_{out} = \text{FFN}(\text{LN}(F_{attention})) + F_{attention} \tag{12}$$

Finally, following the configuration of Vaswani *et al*. (2017), we stack six STTB modules to effectively capture and model spatiotemporal dependencies, thereby enhancing the model's representational capacity.

## 4. Experiments and evaluation

### 4.1. Data description and experimental setup

The proposed model was evaluated on three spatiotemporal process datasets (see Table 1), including the Shenzhen radar echo rainfall dataset, the WeatherBench dataset, and the Dongting Lake flood inundation dataset.

The Shenzhen radar echo rainfall dataset originates from Luo *et al*. (2021) and is used for short-term rainfall forecasting. From this dataset, the samples used for training, validation, and testing are 8,000, 2,000, and 4,000, respectively. Each sample consists of 15 radar echo images, recorded at 6-minute intervals, covering an area of 101 km x 101 km. In the training phase, the model takes the first five-time steps as input to predict the radar echo rainfall patterns for the subsequent ten-time steps.

The WeatherBench dataset, provided by Rasp *et al*. (2020), serves as a widely recognized benchmark for climate prediction. It includes gridded climate variables from 1979 to 2018, encompassing a variety of meteorological factors such as temperature, humidity, and others. Due to the dataset's large scale and the abundance of variables, we select temperature as the target variable for our experiments. The data has a spatial resolution of 5.625° (resulting in a $64 \times 32$ global grid) and a temporal resolution of 1 hour, providing global coverage. Following the data split strategy of Rasp *et al*. (2020), we use the years 1979–2015 for training, 2016 for validation, and 2017–2018 for testing. This results in 13,514 training samples, 366 validation samples, and 730 testing samples. Each sample consists of 24 sequential global temperature frames,

**Table 1.** Overview of the spatiotemporal process datasets used in the experiments.

| Dataset | Input size | Train sequence | Validation/test sequence | Input length | Output length |
|---|---|---|---|---|---|
| Shenzhen radar echo rainfall | 128 x 128 | 8,000 | 6,000 | 5 | 10 |
| WeatherBench | 64 x 32 | 13,514 | 1,096 | 12 | 12 |
| Dongting lake flood inundation | 128 x 128 | 10,800 | 2,730 | 12 | 12 |

corresponding to 24 hourly measurements. The prediction task is formulated as a 12-h ahead forecast, using the past 12 hours of temperature data to predict the following 12 h. All temperature values are expressed in Kelvin (K).

The Dongting Lake flood inundation dataset, compiled by Chen *et al*. (2024), records the flood process from 2012 to 2022, featuring an hourly temporal resolution and a spatial resolution of 10 meters. We selected data from the years 2012, 2014, 2016, 2017, and 2019 for model training. Given that each image contains $4{,}350 \times 5{,}610$ grid cells, making direct input to the model computationally expensive and prone to memory overflow. To mitigate this, we cropped the images into $128 \times 128$ patches with a 0.2 overlap ratio, yielding 10,800 training samples and 2,730 testing samples. Each sample comprises 24 flood inundation images, where the initial 12 frames serve as input and the remaining 12 frames are predicted as output, predicting the spatiotemporal evolution of flood inundation.

To ensure a fair evaluation, all experiments were conducted on the same machine to eliminate potential biases caused by varying hardware setups. We selected seven models, ConvLSTM (Shi *et al*. 2015), PredRNN (Wang *et al*. 2017), PredRNN++ (Wang *et al*. 2018), SimVP (Gao *et al*. 2022), IDA-LSTM (Luo *et al*. 2021), SwinLSTM (Tang *et al*. 2023), PredFormer (Tang *et al*. 2024), IAM4VP (Seo *et al*. 2023), STMixGAN (He *et al*. 2025), and Uniformer (Li *et al*. 2023), for comparison with the proposed model. We selected Adam as the optimizer, with a learning rate set to 0.0005 and a maximum of 8,000 iterations. All implementations were developed using PyTorch and Python.

### 4.2. Evaluation metrics

To validate the performance and effectiveness of the proposed model, we used Mean Squared Error (MSE), Mean Absolute Error (MAE), and Structural Similarity Index Measure (SSIM) as evaluation metrics to measure the differences between the prediction and ground truth (Liu *et al*. 2022). Given the observed image $y_i$ and predicted image $\hat{y}_i$, their calculation formulas are as follows (Shi *et al*. 2015):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overset{'}{y}_i\right)^2 \tag{15}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \overset{'}{y}_i\right| \tag{16}$$

$$\mathrm{SSIM}(\hat{\mathrm{y}}, \mathrm{y}) = \frac{\left(2\mu_{\hat{\mathrm{y}}}\mu_{\mathrm{y}} + \mathrm{C}_1\right)\left(2\sigma_{\hat{\mathrm{y}}\mathrm{y}} + \mathrm{C}_2\right)}{\left(\mu_{\hat{\mathrm{y}}}^2 + \mu_{\mathrm{y}}^2 + \mathrm{C}_1\right)\left(\sigma_{\hat{\mathrm{y}}}^2 + \sigma_{\mathrm{y}}^2 + \mathrm{C}_2\right)} \tag{17}$$

where $n$ is the total number of observations, $\mu_{\hat{y}}$ and $\mu_y$ represent the mean values of the predicted and ground truth images, $\sigma_{\hat{y}}^2$ and $\sigma_y^2$ are their respective variances, $\sigma_{\hat{y}y}$ is the covariance between them, C1 and C2 are small constants added to prevent division by zero.

In addition, we also selected the meteorological evaluation metrics, Critical Success Index (CSI) (Liu *et al*. 2022) and Heidke Skill Score (HSS) (Luo *et al*. 2021), to assess the accuracy of radar echo rainfall process prediction. Specifically, we set three thresholds of 5, 20, and 40 to represent different rainfall intensities (light rain, moderate rain, and

heavy rain). The predicted pixel values and ground-truth pixel values are converted to 0/1 using these thresholds. The formulas for CSI and CSI can be as follows (Hogan *et al*. 2010):

$$CSI = \frac{TP}{TP + FN + FP} \tag{18}$$

$$HSS = \frac{TP \times TN - FN \times FP}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)} \tag{19}$$

where TP (True Positive) refers to correctly predicting an event that occurred, and TN (True Negative) refers to correctly predicting the absence of an event. FP (False Positive) means incorrectly predicting an event that did not occur, and FN (False Negative) means failing to predict an event that occurred. CSI and HSS scores lie between 0 and 1, with greater values reflecting superior prediction performance.

## 4.3. Performance comparison

### 4.3.1. Comparison with state-of-the-art (SOTA) methods

Table 2 presents the performance comparsion between the proposed model and the SOTA models on Shenzhen radar echo data. Bold values indicate the best performance. Spatiotemporal Multiscale Transformer (SMT) model refers to the model without the adaptive scale selection module, while SAMT represents the model incorporating this module. Specifically, our model outperforms others in HSS and CSI across different thresholds. At HSS thresholds of 5, 20, and 40 dBZ, the proposed model surpasses the best benchmark model (IDA-LSTM) by 0.8, 1.5, and 8.4%, respectively. Similarly, for CSI at the same thresholds, SAMT demonstrates improvement of 1, 2.3, and 5.9% over IDA-LSTM, respectively. In addition, compared to other models, the SAMT model achieves the second-best result in MAE, with only a minor difference from the optimal result, while achieving the best performance in MSE. Notably, SwinLSTM, PredFormer, Uniformer, IAM4VP and STMixGAN performed similarly poorly across all metrics. This may be due to the fact that both models struggle to capture the spatiotemporal evolution patterns of radar-based rainfall processes, especially under conditions of limited historical input and long prediction horizons. Furthermore, the results indicate that

**Table 2.** Quantitative comparison results between our model and SOTA models on the Shenzhen radar echo rainfall dataset.

| | HSS ↑ | | | CSI ↑ | | | | |
|---|---|---|---|---|---|---|---|---|
| dBZ threshold | 5 | 20 | 40 | 5 | 20 | 40 | MAE ↓ | MSE ↓ |
| ConvLSTM | 0.696 | 0.491 | 0.127 | 0.771 | 0.421 | 0.070 | 15.00 | 24.34 |
| PredRNN | 0.710 | 0.494 | 0.096 | 0.773 | 0.408 | 0.059 | 14.15 | 23.67 |
| PredRNN++ | 0.708 | 0.515 | 0.148 | 0.772 | 0.437 | 0.091 | 14.35 | 23.8 |
| SimVP | 0.675 | 0.474 | 0.129 | 0.755 | 0.401 | 0.082 | 15.60 | 24.76 |
| IDA-LSTM | 0.715 | 0.518 | 0.171 | 0.778 | 0.440 | 0.106 | **14.13** | 23.66 |
| SwinLSTM | 0.542 | 0.348 | 0.003 | 0.656 | 0.297 | 0.002 | 21.30 | 30.75 |
| PredFormer | 0.530 | 0.346 | 0.007 | 0.644 | 0.290 | 0.004 | 21.54 | 31.33 |
| IAM4VP | 0.529 | 0.348 | 0.012 | 0.644 | 0.297 | 0.007 | 21.79 | 32.13 |
| STMixGAN | 0.535 | 0.356 | 0.014 | 0.650 | 0.303 | 0.008 | 21.46 | 30.93 |
| Uniformer | 0.538 | 0.351 | 0.014 | 0.648 | 0.300 | 0.008 | 21.76 | 30.41 |
| SMT (ours) | 0.723 | 0.515 | 0.237 | 0.784 | 0.450 | 0.150 | 14.47 | 23.00 |
| SAMT (ours) | **0.723** | **0.533** | **0.255** | **0.788** | **0.463** | **0.161** | 14.31 | **22.56** |

most models effectively identify low-value regions (5 dBZ). However, for medium-to-high-value regions (20 and 40 dBZ) SAMT outperforms the the weakest-performing model, improving prediction accuracy by 18.7 and 25.2% in HSS and by 17.3 and 15.9% in CSI. The results domonstrate that the SAMT model provides more accurate predictions of the spatiotemporal rainfall process across different rainfall thresholds.

Figure 5 depicts the variations in HSS and CSI prediction curves at different time steps under various thresholds. As the number of future time frames increases, the prediction performance of most models gradually declines, highlighting the challenges of long-term forecasting. However, as shown in Figure 5, our model achieves the best prediction accuracy at all threshold values, with its prediction curves consistently remaining above those of other models. These results indicate the enhanced stability and reliability of the proposed model.

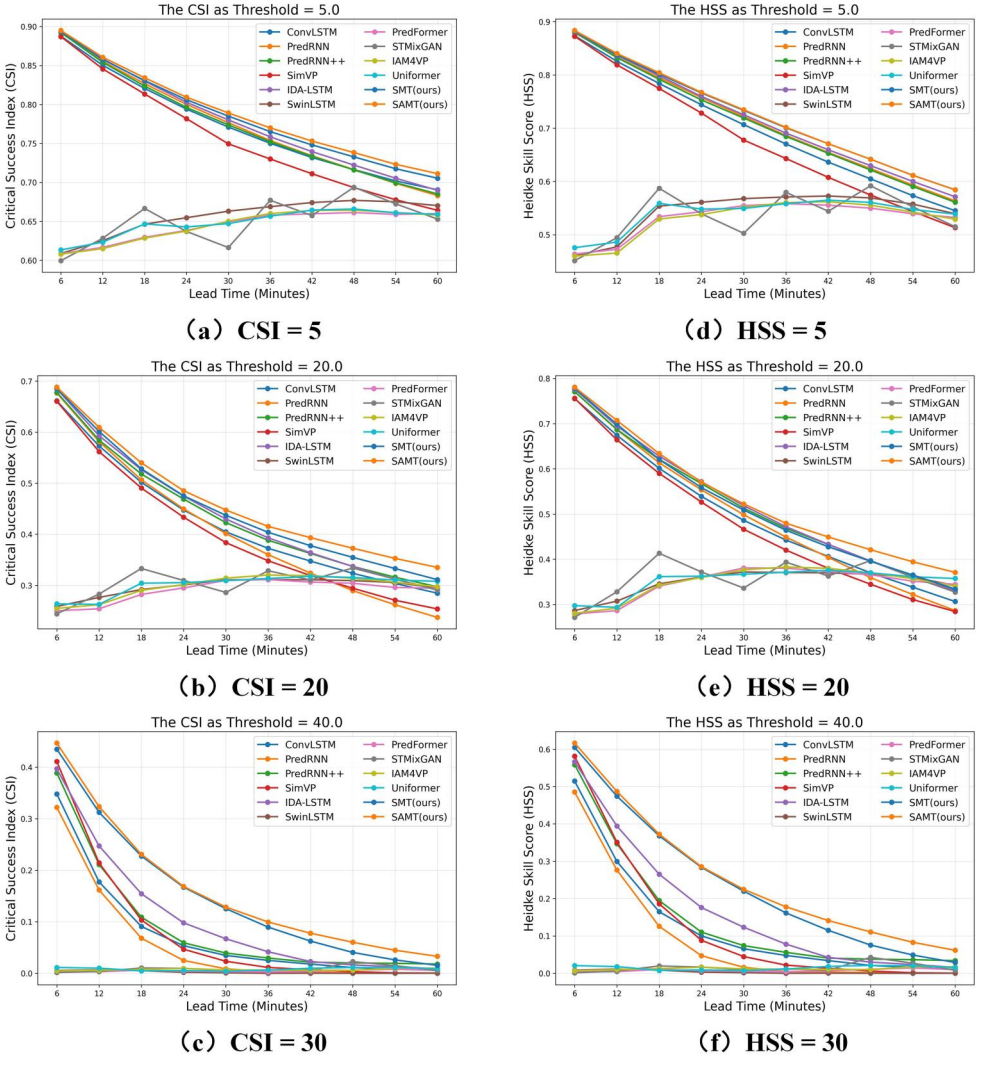


**Figure 5.** Comparison of radar echo rainfall prediction performance across models using CSI and HSS under different thresholds.

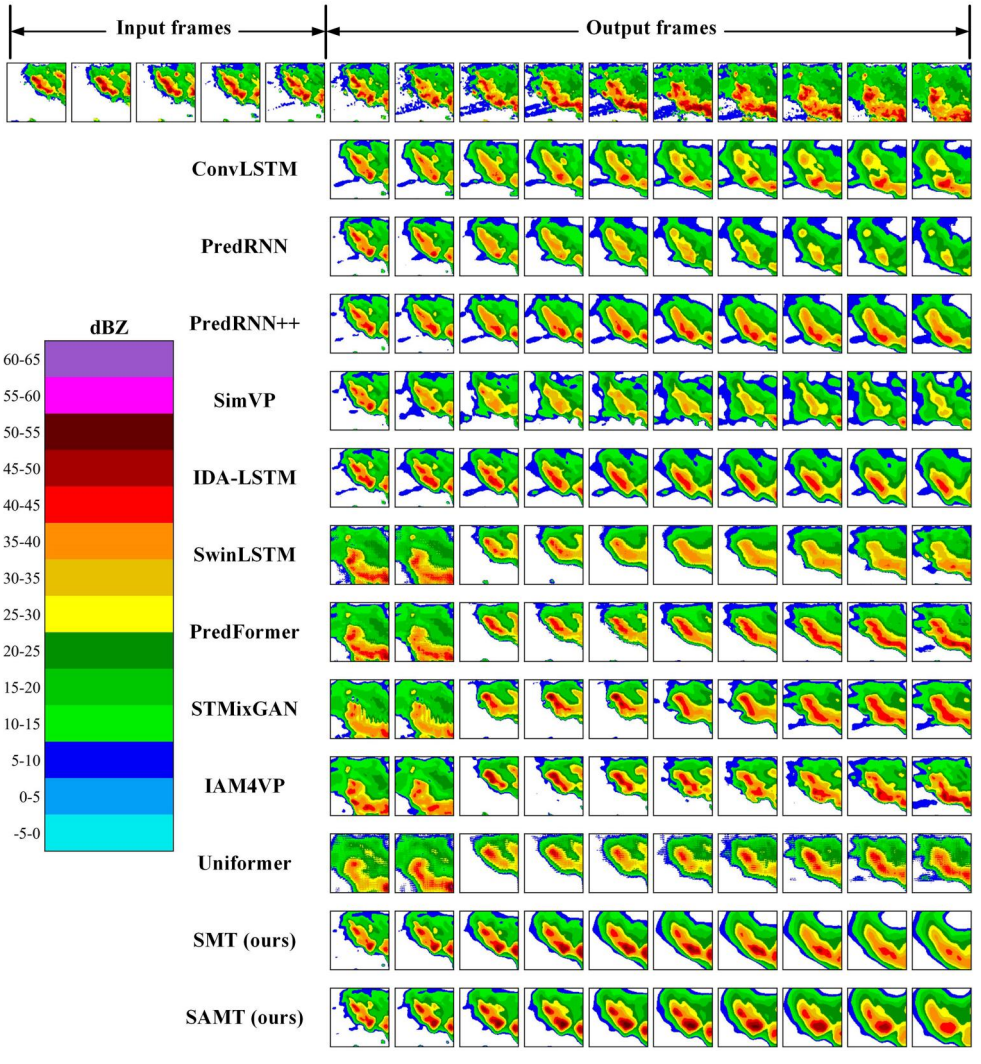


**Figure 6.** An example of visualization results on the Shenzhen radar echo rainfall dataset. The top row displays five input images along with the corresponding ground truth outputs, while the rows below show predictions from different models.

To further demonstrate the predictive capability of the proposed model across different rainfall intensities, Figures 6 and 7 present the visualization of prediction results and the corresponding errors between the predicted and observed values for various models on the Shenzhen radar echo rainfall dataset. As illustrated in Figures 6 and 7, SAMT produces predictions that closely resemble the ground truth and consistently outperforms other models across all regions, achieving lower errors under varying rainfall intensities. SMT ranks second, followed by IDA-LSTM, whereas SwinLSTM, PredFormer, STMixGAN, IAM4VP, and Uniformer perform less effectively. These results underscore the strong ability of SAMT to model spatial heterogeneity and capture heterogeneous patterns in rainfall processes. In addition, SAMT maintains consistently sparse prediction errors across time steps, indicating its capacity to effectively capture

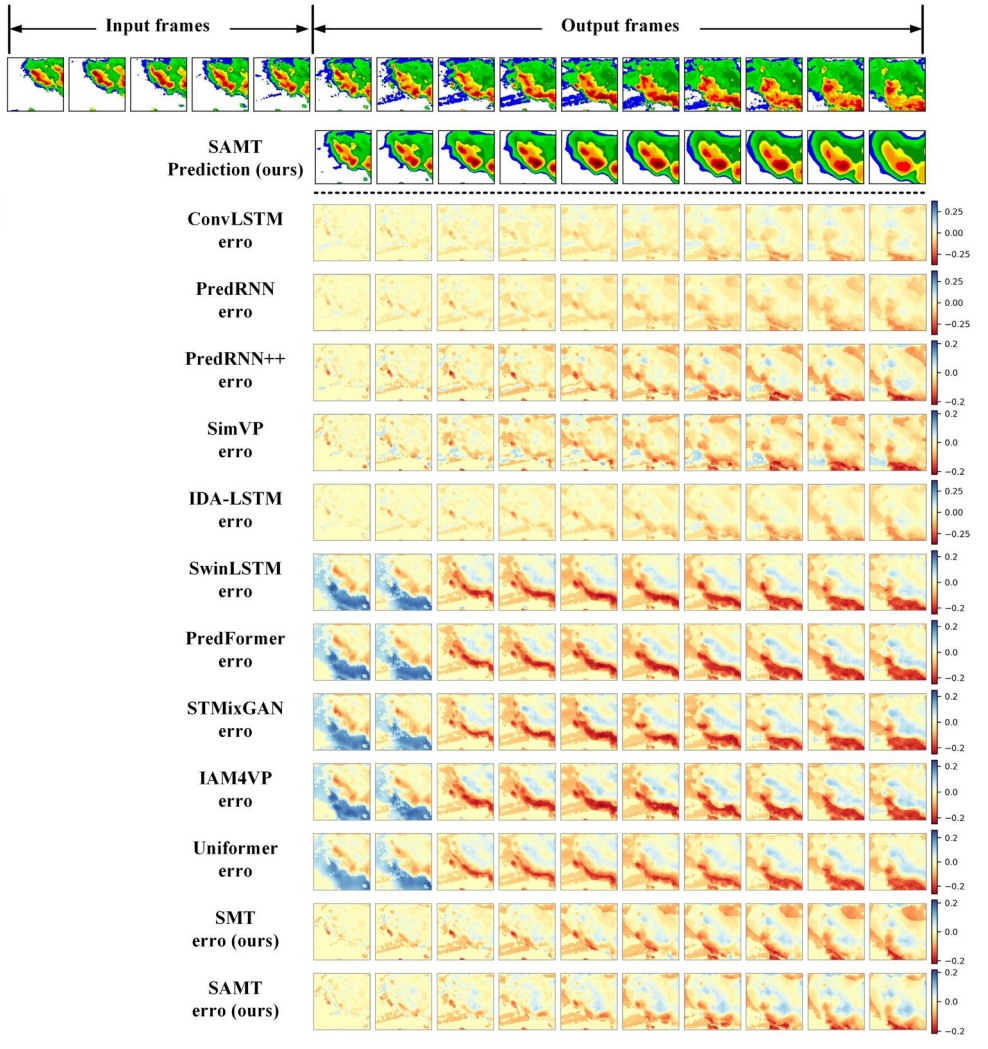


**Figure 7.** An example of error comparison visualization on the Shenzhen radar echo rainfall dataset. The top row displays five input images along with their corresponding ground truth outputs. The second row shows the predictions generated by SAMT, while the remaining rows present the normalized errors between the predictions of different models and the ground truth.

both short- and long-term dependencies. This advantage stems from its architecture, which leverages the short-range modelling capabilities of convolution operations and the long-range modelling strength of the transformer.

Similarly, the SAMT model was evaluated on the WeatherBench dataset. As presented in Table 3, our model achieves the highest scores across all metrics. Specifically, the proposed SAMT model improves upon the best benchmark by 22.27% in MAE, 41.93% in MSE, and 0.76% in SSIM. Notably, all state-of-the-art baseline models demonstrate a decline in performance compared to the proposed SAMT model, with SimVP model showing the most significant decline, highlighting the superior generalization capability of our model. Figures 8 and 9 present the visualization of prediction

**Table 3.** Quantitative comparison results between our model and SOTA models on the WeatherBench dataset.

| Model | MAE ↓ | MSE ↓ | SSIM ↑ |
|---|---|---|---|
| ConvLSTM | 109.46 | 10.40 | 0.7654 |
| PredRNN | 148.25 | 18.45 | 0.6487 |
| PredRNN++ | 121.44 | 14.13 | 0.7917 |
| SimVP | 1808.38 | 296.04 | 0.2707 |
| IDA-LSTM | 135.02 | 15.47 | 0.7135 |
| SwinLSTM | 52.34 | 2.34 | 0.9025 |
| PredFormer | 47.12 | 1.78 | 0.9260 |
| IAM4VP | 43.00 | 1.55 | 0.9262 |
| STMixGAN | 54.38 | 6.91 | 0.9364 |
| Uniformer | 46.74 | 1.84 | 0.8908 |
| SMT (ours) | 31.86 | 0.93 | 0.9421 |
| SAMT (ours) | **31.27** | **0.90** | **0.9440** |

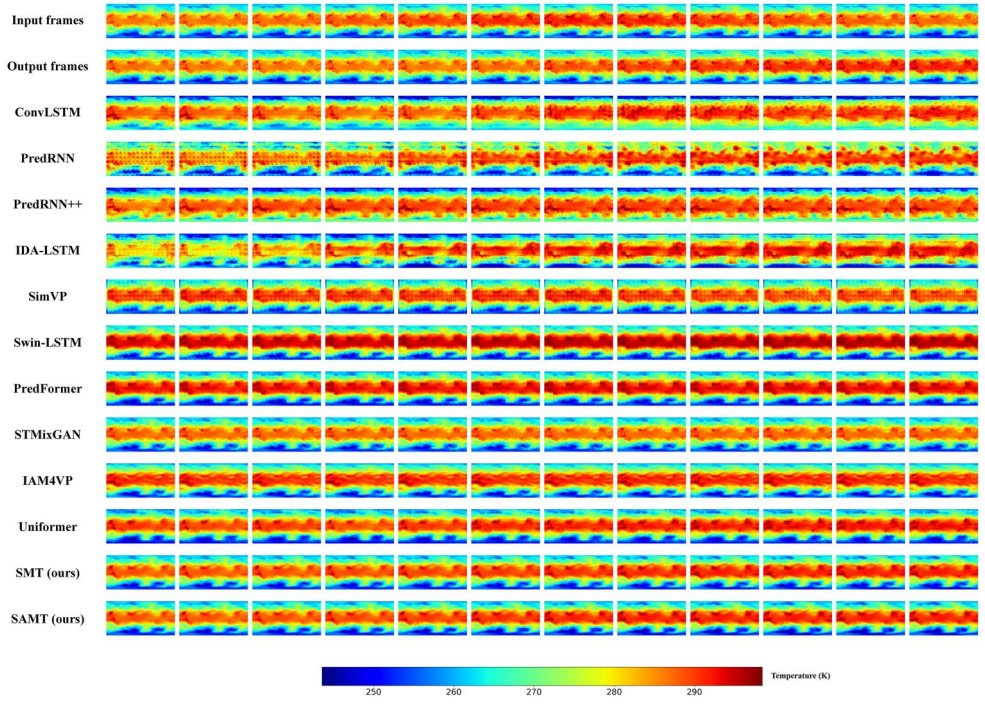


**Figure 8.** An example of visualization results on the WeatherBench dataset. The first and second rows display the ground truth temperature maps, with the first 12 hours serving as input and the subsequent 12 hours as output. The remaining rows show the 12-hour predictions from various models.

global temperature predictions results and the corresponding errors between the predicted and observed values for various models. The error is calculated as the absolute difference between the predicted and ground truth frames, i.e., $|\text{predicted} - \text{ground truth}|$.

As observed, the predictions generated by SAMT exhibit a high degree of consistency with the actual observations. Compared with the true temperature distribution, SAMT produces smaller error maps and effectively captures the global temperature

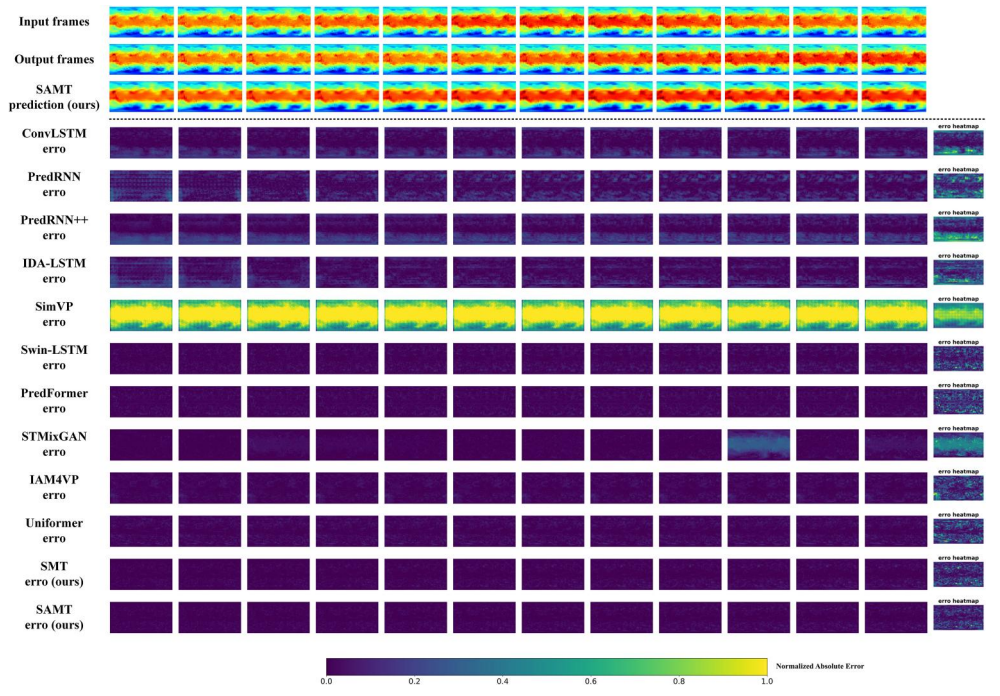


**Figure 9.** An example of error comparison visualization on the WeatherBench dataset. The first and second rows show the ground truth temperature maps, with the first 12 hours used as inputs and the following 12 hours as outputs. The third row presents the predictions from our proposed model, while the remaining rows display the error maps between the predicted and true values for different models over the next 12 hours.

distribution patterns across low-, medium-, and high-heterogeneity regions. SwinLSTM and PredFormer generally align with the observed temperature distribution but tend to overestimate high-value regions (red), leading to larger errors. The predictions from Uniformer, IAM4VP, and STMixGAN models are mostly consistent with the ground truth, but their error distributions are larger than that of SAMT, with STMixGAN exhibiting relatively larger errors in certain cases (e.g., the third-to-last predicted frame). In contrast, models such as ConvLSTM, PredRNN++, and IDA-LSTM fail to accurately capture temperature dynamics, especially in medium-to-high heterogeneity regions, resulting in deviations from the ground. Moreover, PredRNN and SimVP exhibit large prediction errors from the beginning, with accuracy deteriorating over time, particularly for SimVP, whose degradation is more pronounced. The results in Table 3, Figures 8 and 9 further demonstrate that the SAMT model, which incorporates multiscale spatial heterogeneity, has a distinct advantage in capturing complex spatial patterns, especially in heterogeneous regions, enabling accurate prediction of dynamic changes in global temperature.

Beyond evaluating the effectiveness of the proposed model on radar echo datasets, we further performed comparative experiments using the flood inundation dataset to demonstrate its applicability across diverse spatiotemporal processes. Table 4 summarizes the quantitative performance results on the flood inundation dataset. Figure 10

**Table 4.** Quantitative comparison results between our model and SOTA models on the flood inundation dataset.

| Model | MAE ↓ | MSE ↓ | SSIM ↑ |
|---|---|---|---|
| ConvLSTM | 77.49 | 36.75 | 0.9821 |
| PredRNN | 61.21 | 34.68 | 0.9861 |
| PredRNN++ | 74.8 | 185.07 | 0.9630 |
| SimVP | 58.75 | 70.00 | 0.9822 |
| IDA-LSTM | 85.64 | 35.97 | 0.9859 |
| SwinLSTM | 69.92 | 34.86 | 0.9831 |
| PredFormer | 108.92 | 35.28 | 0.9717 |
| IAM4VP | 99.07 | 36.84 | 0.9822 |
| STMixGAN | 102.84 | 45.45 | 0.9792 |
| Uniformer | 157.94 | 37.86 | 0.9621 |
| SMT (ours) | 53.26 | 33.89 | 0.9878 |
| SAMT (ours) | **49.21** | **33.86** | **0.9880** |

provides visualization examples of the predictions, where white represents "water" and black represents "no water." The first shows the ground truth inundation areas, with the first 12 frames as input and the subsequent 12 frames as the expected output. The remaining rows display the predicted inundation areas for different models. For clarity, one frame is visualized every three frames. Our model achieves superior performance across three commonly used evaluation metrics (see Table 4), namely MAE, MSE, and SSIM, outperforming the strongest baseline by 16.23, 2.36, and 0.19%, respectively. Among the compared models, PredRNN++ performers the worst in terms of MSE, while Uniformer exhibits the poorest performance in MAE. As for SSIM, the differences among models are relatively small, with only a 2.59% gap between the best and worst-performing models.

Figure 10 illustrates that all models can effectively capture the overall shape of the inundation areas. However, from the visualization results, ConvLSTM, PredRNN, PredRNN++, and IDA-LSTM produce blurry predictions, with IDA-LSTM being the most affected, followed by PredRNN++. In contrast, SwinLSTM, PredFormer, IAM4VP, STMixGAN, Uniformer, SimVP, SMT and SAMT generate clearer and more reliable predictions, indicating that these models have stronger generalization ability and can better adapt to different spatiotemporal process scenarios. Furthermore, compared to SimVP, the proposed model aligns more closely with the actual inundation extent dynamics, especially in complex regions such as boundaries. The results highlight the superior capability of our model to learn and capture the intricate dynamics inherent in the flood inundation process.

#### *4.3.2. Comparison with variants of proposed method*

To further verify the effectiveness of our model, we performed comparative experiments to assess the contribution of its key components. Table 5 summarizes the ablation study results, where "w/o" indicates the exclusion of a specific module. Notably, CBAM + MSHM is equivalent to ESMHB as described in Section 3.2.1, but for consistency in comparison, we refer to it uniformly as CBAM + MSHM. As shown in Table 5, removing either the CBAM or ASSM component leads to a decline in all metrics. The results suggest that CBAM enables the SAMT model to prioritize crucial information while filtering out irrelevant details. Meanwhile, ASSM allows for adaptive selection of multiscale information, assigning varying weights to different scales, which significantly

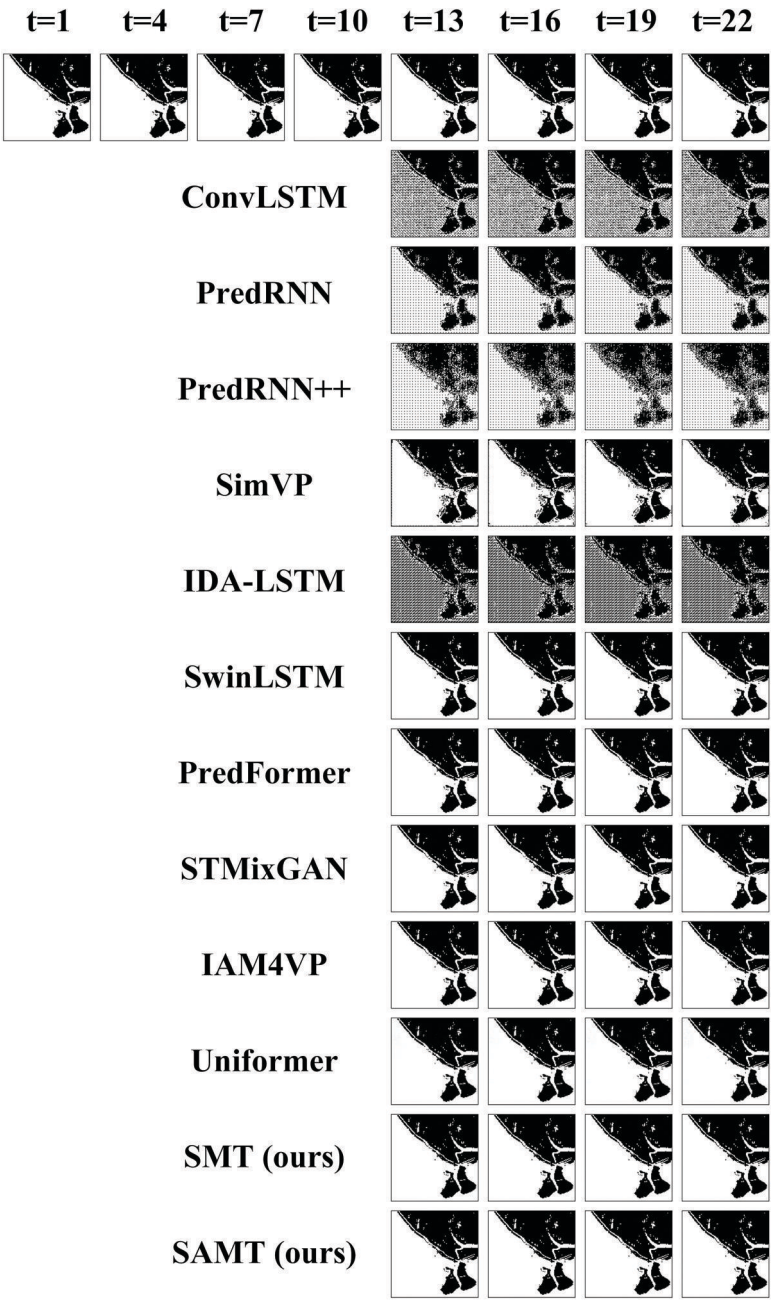


**Figure 10.** Visualization of prediction examples on the Dongting Lake flood inundation dataset.

enhances the model's performance. When the model lacks the MSHM component, its performance is significantly worse than all results obtained with MSHM included, emphasizing the critical role of multiscale features and spatial heterogeneity in spatio-temporal prediction.

In addition, we investigated the impact of CBAM placement on prediction performance. The results indicate that applying feature enhancement first outperforms

**Table 5.** Quantitative results of the ablation study for our model on the Shenzhen radar echo rainfall dataset.

| | HSS ↑ | | | CSI ↑ | | | | |
|---|---|---|---|---|---|---|---|---|
| dBZ Threshold | 5 | 20 | 40 | 5 | 20 | 40 | MAE ↓ | MSE ↓ |
| w/o CBAM<br>w/o MSHM<br>w/o ASSM | 0.717 | 0.462 | 0.232 | 0.779 | 0.419 | 0.147 | 17.52 | 26.79 |
| w/o CBAM<br>MSHM + ASSM | 0.721 | 0.517 | 0.251 | 0.785 | 0.454 | 0.158 | 14.56 | 22.97 |
| w/o ASSM<br>CBAM + MSHM | 0.723 | 0.515 | 0.237 | 0.784 | 0.450 | 0.150 | 14.47 | 23.00 |
| w/o MSHM<br>w/o ASSM<br>CBAM | 0.722 | 0.512 | 0.247 | 0.783 | 0.452 | 0.153 | 14.92 | 23.66 |
| MSHM + ASSM + CBAM | 0.715 | 0.518 | 0.220 | 0.788 | 0.448 | 0.138 | 14.36 | 22.59 |
| CBAM + MSHM + ASSM | **0.723** | **0.533** | **0.255** | **0.788** | **0.463** | **0.161** | **14.31** | **22.56** |

**Table 6.** Performance comparison of single-scale and multi-scale spatial heterogeneity.

| | Scale | | | | HSS ↑ | | | CSI ↑ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | 1 | 3 | 5 | 7 | 5 | 20 | 40 | 5 | 20 | 40 |
| scale_1 | ✓ | | | | 0.702 | 0.471 | 0.247 | 0.787 | 0.432 | 0.155 |
| scale_3 | | ✓ | | | 0.708 | 0.489 | 0.224 | 0.772 | 0.417 | 0.138 |
| scale_5 | | | ✓ | | 0.711 | 0.497 | 0.205 | 0.776 | 0.440 | 0.130 |
| scale_7 | | | | ✓ | 0.713 | 0.502 | 0.240 | **0.789** | 0.446 | 0.150 |
| scale_1_3_5_7 | ✓ | ✓ | ✓ | ✓ | **0.723** | **0.533** | **0.255** | 0.788 | **0.463** | **0.161** |

applying it afterwards. Moreover, removing the designed components results in the worst performance, with significant declines across all metrics. Ultimately, the model incorporating all components achieves the best results, confirming the effectiveness of the proposed design through ablation experiments.

### *4.3.3. Multiscale spatial heterogeneity analysis*

The ablation study results in Section 4.3.2 demonstrate that ignoring the inherent spatial heterogeneity in spatiotemporal processes significantly degrades the SAMT model's prediction accuracy. To further emphasize the importance of incorporating multiscale features, we conducted a comparative analysis on the Shenzhen radar echo rainfall dataset by adjusting the convolution kernel size in the MSHM, evaluating the model's performance under both single-scale and multiscale settings. Specifically, we tested four individual scales corresponding to convolution kernel sizes of 1, 3, 5, and 7, as well as a multiscale scenario that integrates all four. Notably, different kernel sizes represent different receptive fields, with each kernel size capturing spatial dependencies at a distinct scale. The quantitative results are presented in Table 6, while Figure 11 provides a visual comparison of predictions under single-scale and multiscale settings.

As presented in Table 6, the multiscale model consistently outperforms all single-scale counterparts across nearly all threshold settings for both HSS and CSI metrics, with particularly pronounced improvements in the medium-to-high value ranges. The results clearly demonstrate that multiscale feature representations are crucial for capturing spatial heterogeneity, significantly enhancing the model's ability to forecast dynamic patterns in heterogeneous regions accurately. In detail, when the HSS and CSI

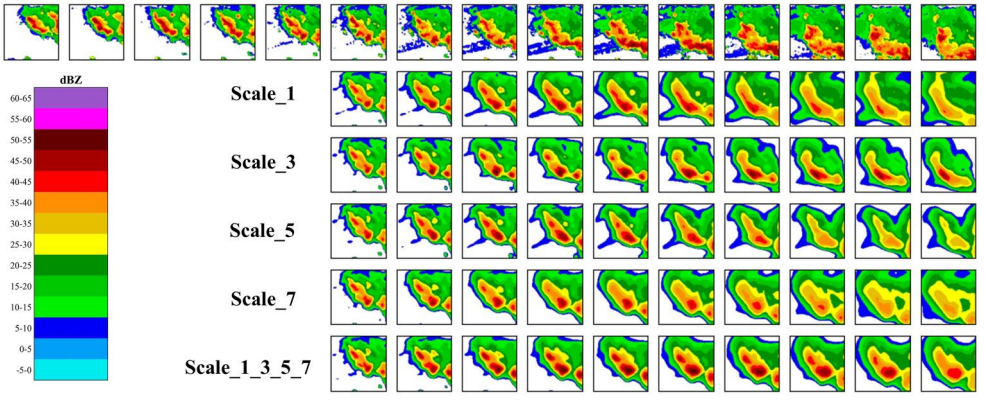


**Figure 11.** Visualization of prediction examples under single-scale and multiscale settings on the Shenzhen radar echo rainfall dataset.

thresholds are set to 40, the scale_5 configuration yields the weakest performance, followed by scale_3, scale_7, and scale_1, respectively. Under thresholds of 5 and 20, scale_1 exhibits the poorest performance, while scale_7 performs best. These results indicate that different convolutional scales capture complementary aspects of spatial heterogeneity that are beneficial for spatiotemporal prediction. They also highlight the limitations of single-scale modelling in representing complex dynamic processes.

Furthermore, as illustrated in Figure 11, although the prediction quality of all models degrades over time, the multiscale model consistently stays closer to the ground truth. These findings indicate that the multiscale model better captures long-term dependencies and more accurately represents the direction and morphological changes of complex spatiotemporal dynamics.

## 5. Conclusions and future work

In this study, we propose a novel model, SAMT, for improving the prediction of spatiotemporal processes. The model effectively captures the multiscale characteristics and spatial heterogeneity of spatiotemporal dynamics, which are commonly ignored by existing approaches. To enhance feature representation, we introduce an Adaptive Scale Selection Module (ASSM) to assign dynamic weights to features at different scales based on their individual contributions and thus reducing information redundancy and improving utilization efficiency. Then, a Spatiotemporal Transformer Block (STTB) is incorporated to simultaneously model short-term fluctuations and long-term dependencies. The results of three representative spatiotemporal datasets show that SAMT consistently surpasses state-of-the-art models across different evaluation metrics, with HSS increasing by 0.8–25.2%, CSI by 1–15.9%, MAE by 16.23–98.26%, MSE by 2.36–99.69%, and SSIM by 0.19–67.25%. The proposed model contributes to the advancement of geographic information science through more accurate and effective dynamic spatiotemporal prediction. The SAMT can be implemented in the dynamic spatial analysis and spatiotemporal process evolution in broader fields, such as ecology, hydrology, and meteorology. Even promising progress has been made, the study

is constrained to regions with abundant data, which requires future research to enhance spatiotemporal prediction in data-scarce or data-absent areas through methods such as transfer learning and physics-informed modelling.

## Acknowledgements

We thank the editor and anonymous reviewers for their constructive comments.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported in part by the National Natural Science Foundation of China (42371428), and China Scholarship Council (202406410052).

## Notes on contributors

***Lai Chen*** is currently a PhD student in the National Engineering Research Center of Geographic Information System at China University of Geosciences (Wuhan). He is also a Research Associate at Curtin University, Perth, Australia. His research interests include GeoAI, Spatiotemporal process prediction, deep learning, and flood disaster. He contributed to the idea, study design, methodology, implementation, and manuscript writing of this paper.

***Zeqiang Chen*** is a Professor in the National Engineering Research Center of Geographic Information System at China University of Geosciences (Wuhan). His research interests include sensor Web, spatiotemporal big data intelligence and its application in smart watersheds. He contributed to the conceptualization of the research idea, supervision and project administration.

***Yongze Song*** is an Associate Professor at Curtin University, Australia, and a Harvard Spatial Data Lab (SDL) Fellow. His current research interests include geospatial analysis methods, spatial statistics, sustainable development and infrastructure management. He contributed to the conceptualization of the research idea, supervision and project administration.

***Chao Yang*** is an Associate Professor in the National Engineering Research Center of Geographic Information System at China University of Geosciences (Wuhan). His research interests include geospatial sensor networks, data mining, urban sensing, and 3D reconstruction. He contributed to review, editing and formal analysis.

***Sijia He*** is currently a PhD student in the National Engineering Research Center of Geographic Information System at China University of Geosciences (Wuhan). Her research interests focus on flood monitoring methods and knowledge graph construction. She contributed to data collection and collation.

***Zhiqing Li*** is currently a PhD student in the National Engineering Research Center of Geographic Information System at China University of Geosciences (Wuhan). His research interests include remote sensing image processing, object detection, deep learning. He contributed to data collection and collation.

***Wenfeng Guo*** is a scientific research personnel member in Hubei Institute of Land Surveying and Mapping, Wuhan, China. His research interests include spatiotemporal big data analysis and trajectory data mining. He contributed to data collection and collation.

***Nengcheng Chen*** is a Professor in the National Engineering Research Center of Geographic Information System at China University of Geosciences (Wuhan). His research interests focus on Earth observation sensor Web, spatiotemporal big data, Web GIS and smart city. He contributed to review and editing.

## ORCID

Zeqiang Chen http://orcid.org/0000-0002-3521-9972
Yongze Song http://orcid.org/0000-0003-3420-9622

## Data and codes availability statement

The data and codes that support the findings of this study are openly available in Figshare at this public link (https://doi.org/10.6084/m9.figshare.28943921).

## References

Adnan, R.M., *et al.*, 2019. Daily streamflow prediction using optimally pruned extreme learning machine. *Journal of Hydrology*, 577, 123981.

Bližňák, V., Sokol, Z., and Zacharov, P., 2017. Nowcasting of deep convective clouds and heavy precipitation: Comparison study between NWP model simulation and extrapolation. *Atmospheric Research*, 184, 24–34.

Brunner, M.I., *et al.*, 2021. Challenges in modeling and predicting floods and droughts: A review. *WIREs Water*, 8 (3), e1520.

Bui, D.T., *et al.*, 2019. Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection technique and tree based ensemble methods. *The Science of the Total Environment*, 668, 1038–1054.

Cai, X., *et al.*, 2024. Poly Kernel inception network for remote sensing detection. *In*: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, 27706–27716.

Campolo, M., Andreussi, P., and Soldati, A., 1999. River flood forecasting with a neural network model. *Water Resources Research*, 35 (4), 1191–1197.

Chen, L., Chen, Z., and Chen, N., 2024. PDFID: A high-resolution flood inundation dataset with a long time series. *Journal of Hydrology: Regional Studies*, 52, 101715.

Fotheringham, A.S., and Sachdeva, M., 2022. Modelling spatial processes in quantitative human geography. *Annals of GIS*, 28 (1), 5–14.

Gao, H., *et al.*, 2023. AMSSE-Net: adaptive multiscale spatial–spectral enhancement network for classification of hyperspectral and LiDAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–17.

Gao, Z., *et al.*, 2022. SimVP: simpler yet better video prediction. *In*: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3170–3180.

Ge, Y., *et al.*, 2019. Principles and methods of scaling geospatial Earth science data. *Earth-Science Reviews*, 197, 102897.

Graves, A., 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Gu, J., *et al.*, 2018. Recent advances in convolutional neural networks. *Pattern Recognition*, 77, 354–377.

Gu, J., *et al.*, 2023. ConvFormer: combining CNN and transformer for medical image segmentation. *In: 2023 IEEE 20th international symposium on biomedical imaging (ISBI)*. IEEE, 1–5.

He, L., *et al.*, 2025. A spatiotemporal mixed-enhanced generative adversarial network for radar-based precipitation nowcasting. *Computers & Geosciences*, 200, 105919.

Hogan, R.J., *et al.*, 2010. Equitability revisited: Why the "Equitable Threat Score" is not equitable. *Weather and Forecasting*, 25 (2), 710–726.
Hu, X., *et al.*, 2024. A local indicator of stratified power. *International Journal of Geographical Information Science*, 39 (4), 925–943.
Jin, S., *et al.*, 2025. SASTGCN: semantic-augmented spatio-temporal graph convolutional network for subway flow prediction. *International Journal of Applied Earth Observation and Geoinformation*, 139, 104530.
Leskens, J.G., *et al.*, 2014. Why are decisions in flood disaster management so poorly supported by information from flood models? *Environmental Modelling & Software*, 53, 53–61.
Li, B., *et al.*, 2016. Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China. *Hydrology Research*, 47 (S1), 69–83.
Li, H., and Reynolds, J.F., 1995. On definition and quantification of heterogeneity. *Oikos*, 73 (2), 280.
Li, K., *et al.*, 2023. UniFormer: unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (10), 12581–12600.
Li, Q., Han, Z., and Wu, X., 2018. Deeper insights into graph convolutional networks for semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (1), 3538–3545.
Liu, J., Xu, L., and Chen, N., 2022. A spatiotemporal deep learning model ST-LSTM-SA for hourly rainfall forecasting using radar echo images. *Journal of Hydrology*, 609, 127748.
Lü, H., *et al.*, 2013. The streamflow estimation using the Xinanjiang rainfall runoff model and dual state-parameter estimation method. *Journal of Hydrology*, 480, 102–114.
Luo, C., *et al.*, 2021. A Novel LSTM model with interaction dual attention for radar echo extrapolation. *Remote Sensing*, 13 (2), 164.
Noori, N., and Kalin, L., 2016. Coupling SWAT and ANN models for enhanced daily streamflow prediction. *Journal of Hydrology*, 533, 141–151.
Panahi, M., *et al.*, 2021. Deep learning neural networks for spatially explicit prediction of flash flood probability. *Geoscience Frontiers*, 12 (3), 101076.
Pulukuri, S., Keesara, V.R., and Deva, P., 2018. Flow forecasting in a watershed using autoregressive updating model. *Water Resources Management*, 32 (8), 2701–2716.
Rasp, S., *et al.*, 2020. WeatherBench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12 (11), e2020MS002203.
Ren, K., Song, Y., and Yu, Q., 2025. Second-dimension outliers for spatial prediction. *International Journal of Geographical Information Science*, 1–28.
Riera, J.L., and Magnuson, J.J., 1998. Analysis of large-scale spatial heterogeneity in vegetation indices among north american landscapes. *Ecosystems*, 1 (3), 268–282.
Sadeghi Tabas, S., *et al.*, 2023. FlowDyn: a daily streamflow prediction pipeline for dynamical deep neural network applications. *Environmental Modelling & Software*, 170, 105854.
Seo, M., *et al.*, 2023. Implicit stacked autoregressive model for video prediction.
Shi, X., *et al.*, 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28, 802–810.
Song, Y., 2022. The second dimension of spatial association. *International Journal of Applied Earth Observation and Geoinformation*, 111, 102834.
Song, Y., *et al.*, 2020. An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: cases with different types of spatial data. *GIScience & Remote Sensing*, 57 (5), 593–610.
Song, Y., *et al.*, 2025. Advancing geospatial methods for addressing global resource and sustainability challenges. *Resources, Conservation and Recycling*, 223, 108517.
Speight, L.J., *et al.*, 2021. Operational and emerging capabilities for surface water flood forecasting. *WIREs Water*, 8 (3), e1517.
Su, L., *et al.*, 2024. Improving runoff simulation in the Western United States with Noah-MP and VIC models. *Hydrology and Earth System Sciences*, 28 (13), 3079–3097.

Szegedy, C., *et al.*, 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31 (1), 4278–4284.

Tang, S., *et al.*, 2023. SwinLSTM: improving spatiotemporal prediction accuracy using swin transformer and LSTM. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13470–13479.

Tang, X., *et al.*, 2021. A novel index to evaluate discretization methods: A case study of flood susceptibility assessment based on random forest. *Geoscience Frontiers*, 12 (6), 101253.

Tang, Y., *et al.*, 2024. PredFormer: transformers are effective spatial-temporal predictive learners.

Vaswani, A., *et al.*, 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

Wang, Y., *et al.*, 2017. PredRNN: recurrent neural networks for predictive learning using spatiotemporal LSTMs. *Advances in Neural Information Processing Systems*, 30, 879–888.

Wang, Y., *et al.*, 2018. PredRNN++: towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. *International Conference on Machine Learning. PMLR*, 5123–5132.

Wang, Y., *et al.*, 2019. Memory in memory: a predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, 9146–9154.

Woo, S., *et al.*, 2018. CBAM: convolutional block attention module. *In*: V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds. *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 3–19.

Wu, D., *et al.*, 2022. Short-term rainfall prediction based on radar echo using an improved self-attention PredRNN deep learning model. *Atmosphere*, 13 (12), 1963.

Xiong, T., *et al.*, 2021. Contextual Sa-attention convolutional LSTM for precipitation nowcasting: a spatiotemporal sequence forecasting view. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 12479–12491.

Xu, L., *et al.*, 2024. Incorporating spatial autocorrelation into deformable ConvLSTM for hourly precipitation forecasting. *Computers & Geosciences*, 184, 105536.

Yan, J., *et al.*, 2018. Urban flash flood forecast using support vector machine and numerical simulation. *Journal of Hydroinformatics*, 20 (1), 221–231.

Yang, Z., Yang, X., and Lin, Q., 2022. TCTN: a 3D-temporal convolutional transformer network for spatiotemporal predictive learning.

Yao, S., *et al.*, 2023. An integrated process-based framework for flood phase segmentation and assessment. *International Journal of Geographical Information Science*, 37 (6), 1315–1337.

Yin, M., Chen, Z., and Zhang, C., 2023. A CNN-transformer network combining cbam for change detection in high-resolution remote sensing images. *Remote Sensing*, 15 (9), 2406.

Zhang, J., *et al.*, 2021. Daily runoff forecasting by deep recursive neural network. *Journal of Hydrology*, 596, 126067.

Zhang, Y., *et al.*, 2018. Predicting runoff signatures using regression and hydrological modeling approaches. *Water Resources Research*, 54 (10), 7859–7878.

Zhang, Z., *et al.*, 2024. Robust interaction detector: a case of road life expectancy analysis. *Spatial Statistics*, 59, 100814.

Zhao, J., *et al.*, 2024. SWAT model applications: from hydrological processes to ecosystem services. *The Science of the Total Environment*, 931, 172605.

Zheng, K., *et al.*, 2022. A knowledge representation model based on the geographic spatiotemporal process. *International Journal of Geographical Information Science*, 36 (4), 674–691.

Zhu, J., *et al.*, 2021. Attention-based parallel networks (APNet) for PM2.5 spatiotemporal prediction. *The Science of the Total Environment*, 769, 145082.

Zhu, J., Fang, L., and Ghamisi, P., 2018. Deformable convolutional neural networks for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 15 (8), 1254–1258.