remote sensing

MDPI

Article

Less Is More: Online Spatio-Temporal Selective Learning for Multi-Variable Meteorological Forecasting

Pu Zhang [1,2], Deping Xiang [3], Chunlei Huo [2,4,*], Kun Ding [1] and Shiming Xiang [1,2]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; zhangpu2019@ia.ac.cn (P.Z.); kun.ding@ia.ac.cn (K.D.); smxiang@nlpr.ia.ac.cn (S.X.)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3] JD.com, Inc., Beijing 101111, China; xiangdeping15@mails.ucas.ac.cn
[4] Information Engineering College, Capital Normal University, Beijing 100048, China
* Correspondence: clhuo@nlpr.ia.ac.cn

Highlights

What are the main findings?

- VASTFormer, a variable-aware spatio-temporal Transformer with physics-enhanced Solar Positional Encoding, achieves the best accuracy among the evaluated lightweight spatio-temporal baselines on regional ERA5 multi-variable forecasting.
- Spatio-Temporal Selective Learning (ST-SL) further improves forecasting accuracy through residual-gap-based learnability estimation, yielding stronger meteorological skill and more accurate percentile-defined extreme-condition forecasts without adding inference-time cost.

What are the implications of the main findings?

- The results challenge the standard full-field supervision paradigm and show that selective supervision can improve generalization in heterogeneous atmospheric fields.
- The combined framework demonstrates that incorporating domain knowledge into both architecture design and training strategy offers a generalizable methodology for spatio-temporal scientific prediction.

Abstract

Accurate forecasting of high-dimensional meteorological fields remains challenging due to the complex spatio-temporal dynamics of atmospheric systems and the presence of heterogeneous training difficulty across space and lead time. Existing deep forecasting approaches usually optimize all prediction units uniformly, which may overemphasize low-benefit or weakly generalizable supervision signals. To address this issue, we propose Spatio-Temporal Selective Learning (ST-SL), an online training framework that estimates the learnability of each prediction unit by comparing the main model with a frozen reference model and computes the loss only over selected high-benefit spatio-temporal units. To provide an effective forecasting backbone, we further introduce VASTFormer, a variable-aware spatio-temporal Transformer that models cross-variable dependencies, incorporates physics-enhanced Solar Positional Encoding, and captures atmospheric trajectories with an efficient temporal translator. Experiments on the ERA5 reanalysis dataset show that VASTFormer outperforms representative spatio-temporal baselines, while ST-SL further improves accuracy without adding inference-time parameters or computational cost. Compared with the strongest baseline, VASTFormer+ST-SL reduces MSE, MAE, and RMSE by 8.84%, 6.70%, and 4.54%, respectively. Meteorological skill evaluation further shows

Check for updates

Academic Editor: Ismail Gultepe

Received: 15 May 2026
Revised: 24 June 2026
Accepted: 2 July 2026
Published: 5 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

an average ACC of 0.9801 and RMSESS of 0.8104, and percentile-based extreme-condition evaluations confirm consistent improvements across standard and high-impact forecasting scenarios. These results indicate that selective supervision can improve generalization in dense meteorological forecasting.

Keywords: meteorological forecasting; selective learning; spatio-temporal forecasting

1. Introduction

Accurate meteorological forecasting is fundamental to a wide range of societal and scientific applications, including extreme weather warning, agricultural planning, renewable energy management, transportation safety, and disaster mitigation [1–5]. Traditionally, Numerical Weather Prediction (NWP) systems have dominated this field by simulating atmospheric dynamics through complex partial differential equations [6]. However, the high computational cost and inherent physical approximations of NWP limit its real-time deployment. Recently, deep learning has emerged as a promising data-driven paradigm for spatio-temporal meteorological forecasting [7]. By directly learning complex atmospheric trajectories from vast amounts of historical reanalysis data, data-driven models have demonstrated remarkable inference speed and competitive accuracy.

Recent advances in deep learning methods have significantly improved the ability of neural networks to model complex dynamical systems. Recurrent architectures, such as ConvLSTM [8] and PredRNN [9], capture temporal dependencies through hidden-state propagation, while convolutional predictive models, such as SimVP [10] and TAU [11], improve efficiency by decoupling spatial encoding and temporal transformation. More recently, Transformer-based models have further expanded the receptive field and enhanced long-range dependency modeling for spatio-temporal prediction [12,13]. Despite these advances, meteorological forecasting remains challenging because atmospheric states are high-dimensional, multi-variable, spatially heterogeneous, and governed by nonlinear interactions across both space and time.

A common limitation across most existing spatio-temporal forecasting models is that they typically model meteorological fields as generic spatio-temporal tensors, with insufficient consideration of variable-specific semantics, structured cross-variable interactions, and physically meaningful spatio-temporal priors. Meteorological data are inherently multivariable and highly entangled across spatial, temporal, and variable dimensions. In practice, atmospheric variables such as temperature, humidity, and wind components have distinct physical meanings and evolution patterns. However, many representative forecasting models process these variables by concatenating them along the channel dimension and applying generic spatio-temporal operators, thereby treating heterogeneous meteorological variables as homogeneous feature channels [8–11,13]. This variable-agnostic formulation may obscure the structured dependencies among meteorological variables, such as the coupling between temperature gradients, humidity transport, and wind dynamics. Moreover, different architectural families suffer from their own limitations when applied to large-scale meteorological forecasting. Recurrent models are often difficult to parallelize and incur high computational costs when modeling long temporal dependencies. Convolutional models are computationally efficient but are constrained by local receptive fields, making it difficult to capture long-range spatial interactions and large-scale atmospheric circulation patterns. Transformer-based models provide stronger global dependency modeling, but directly extending full self-attention to high-resolution spatio-temporal meteorological fields leads to rapidly increasing computational complexity [14,15]. Furthermore, the posi-

tional encodings commonly used in generic Transformer architectures are usually defined according to discrete spatial or temporal indices, rather than physically meaningful atmospheric drivers [16,17]. As a result, important regularities such as geographically dependent diurnal and seasonal variations must be inferred primarily from data, which may reduce data efficiency and limit generalization.

Another overlooked issue lies in the training objective. Current deterministic data-driven forecasting models are commonly trained by minimizing a global regression loss, such as mean squared error, over all variables, lead times, and spatial locations [13,18,19]. This objective implicitly assumes that all spatio-temporal supervision signals are equally reliable and equally beneficial for optimization. However, this assumption is often violated in meteorological datasets. Observational noise, data assimilation artifacts, unresolved physical processes, local stochasticity, and intrinsically chaotic atmospheric dynamics can produce supervision signals with highly heterogeneous reliability and learning value [20–23]. Uniformly fitting all prediction units may therefore amplify gradients from noisy or weakly transferable regions, causing the model to waste capacity on low-benefit patterns and potentially degrading generalization. This observation raises a natural question: can meteorological forecasting models learn better by using less but more informative supervision?

To address these challenges, we propose VASTFormer, a Variable-Aware Spatio-Temporal Transformer backbone, together with Spatio-Temporal Selective Learning, an online selective training framework. VASTFormer incorporates a variable-aware Multi-Variable Fusion (MV-Fusion) module to dynamically capture cross-variable dependencies. To inject crucial physical priors, we introduce a physics-enhanced Solar Positional Encoding mechanism. Furthermore, a highly efficient trajectory-aware Temporal Translator equipped with Spatial Reduction Attention (SRA) [24] and SwiGLU-based [25] feed forward network (FFN) is designed to capture long-range interactions among atmospheric trajectories while alleviating the computational bottleneck of standard spatio-temporal attention.

Built upon this architectural design, we further introduce Spatio-Temporal Selective Learning (ST-SL), a highly adaptive online training framework for multi-variable meteorological forecasting. Unlike conventional training paradigms that uniformly penalize all prediction errors, ST-SL dynamically evaluates the learning benefit of dense spatio-temporal units during training and selectively constructs supervision signals. Specifically, ST-SL compares the main model with a frozen reference model and derives a residual-gap-based learnability-aware score for each prediction unit. An instance-aware thresholding strategy is then used to retain units with higher expected optimization potential, and the regression loss is computed only over these selected units. In this way, ST-SL encourages the model to focus on informative and learnable atmospheric transitions while reducing the influence of noisy or low-benefit supervision signals. Importantly, ST-SL is used only during training; the reference model and mask generation branch are discarded during inference, introducing no additional inference-time parameters or computational cost.

We evaluate the proposed approach on the ERA5 reanalysis dataset [20] over China and Southeast Asia under a unified regional short-range forecasting protocol. Extensive experiments demonstrate that the integration of VASTFormer with ST-SL consistently improves forecasting accuracy across multiple atmospheric variables while maintaining favorable computational efficiency.

Our contributions are summarized as follows:

1. We propose VASTFormer, a variable-aware spatio-temporal Transformer backbone for multi-variable meteorological forecasting, which explicitly models cross-variable interactions and trajectory-aware spatio-temporal dependencies while significantly reducing computational complexity.

2. We introduce Solar Positional Encoding, a physically meaningful spatio-temporal prior based on solar altitude, to help the model capture geographically dependent diurnal and seasonal variations.
3. We develop Spatio-Temporal Selective Learning, an online data selection framework that challenges the uniform supervision assumption by dynamically filtering unreliable spatio-temporal training signals through residual-gap learnability estimation.
4. Extensive experiments on the ERA5 dataset demonstrate state-of-the-art performance among the evaluated lightweight spatio-temporal baselines, showing that selective optimization combined with efficient architecture design significantly improves robustness and generalization in multi-variable meteorological forecasting.

2. Related Work

2.1. Deep Models for Spatio-Temporal Forecasting

Deep spatio-temporal forecasting models aim to predict future states from historical observations evolving across both spatial and temporal dimensions. Early methods mainly rely on recurrent architectures to model temporal dynamics. ConvLSTM [8] replaces fully connected transitions in LSTM with convolutional operations, enabling recurrent networks to capture spatial correlations and temporal evolution for precipitation nowcasting. TrajGRU [26] further learns location-variant recurrent connections to better adapt to motion patterns. PredRNN [9] introduces spatio-temporal memory flows across layers and time steps, and its variants improve recurrent predictive learning through gradient highway units, higher-order non-stationarity modeling, and decoupled spatio-temporal memory representations [27–29]. E3D-LSTM [30] incorporates 3D convolutions into recurrent units to jointly capture short-term motion and long-term dependencies. Despite their effectiveness, recurrent models rely on step-by-step hidden-state propagation, which limits parallelization and often leads to high computational cost.

To improve efficiency, recurrent-free and hybrid predictive models have been widely explored. CrevNet [31] introduces conditionally reversible architectures to preserve information while reducing memory consumption in video prediction. PhyDNet [32] disentangles PDE-inspired physical dynamics from unknown residual factors for unsupervised video prediction. SimVP [10] shows that a simple CNN-based encoder-translator-decoder architecture trained with MSE can achieve competitive performance without recurrent units or auxiliary tricks. SimVPv2 [33] further simplifies spatio-temporal predictive learning by replacing heavy U-Net-like modules with plain convolutional stacks enhanced by gated spatio-temporal attention. MAU [34] broadens the temporal receptive field through a motion-aware unit to capture reliable inter-frame motion information. TAU [11] proposes a Temporal Attention Unit that decomposes temporal modeling into intra-frame static attention and inter-frame dynamic attention. DMVFN [35] models video prediction from a flow-based perspective by using a dynamic multi-scale voxel flow network with adaptive routing for different motion scales. These models improve computational efficiency and predictive accuracy, but most of them are originally designed for generic video prediction and do not explicitly consider the physical structure of meteorological variables.

Transformer-based architectures have recently been introduced to enhance long-range dependency modeling in spatio-temporal forecasting. Earthformer [12] proposes Cuboid Attention to decompose space-time attention into scalable cuboid-level operations for Earth system forecasting. SwinLSTM [36] integrates Swin Transformer blocks into recurrent temporal modeling, combining window-based attention with sequential prediction. PredFormer [13] explores pure Transformer architectures for spatio-temporal predictive learning and introduces gated Transformer blocks to improve scalability and performance. However, most existing predictors treat different meteorological variables as homogeneous chan-

nels, leaving variable-aware fusion, physically meaningful positional priors, and efficient trajectory-aware modeling underexplored. In contrast, VASTFormer is designed to explicitly model cross-variable interactions and physics-enhanced spatio-temporal variation while improving computational efficiency for multi-variable meteorological forecasting.

2.2. Data-Driven Meteorological Forecasting

Data-driven meteorological forecasting has recently attracted increasing attention with the availability of large-scale reanalysis datasets and advances in deep neural networks. WeatherBench [37] establishes an ERA5-based benchmark and evaluation protocol for data-driven medium-range weather forecasting. FourCastNet [38] introduces a global high-resolution weather forecasting model based on Adaptive Fourier Neural Operators. Pangu-Weather [39] proposes a 3D Earth-Specific Transformer for accurate and efficient global medium-range forecasting. GraphCast [18] formulates global weather forecasting as message passing on a multi-scale graph and achieves skillful medium-range prediction. FuXi [40] develops a cascaded machine-learning system for 15-day global forecasting. FengWu [41] extends skillful global medium-range forecasting beyond 10-day lead times. ClimaX [19] explores foundation modeling for weather and climate by pre-training a flexible Transformer on heterogeneous climate datasets.

These studies demonstrate the strong potential of deep learning for meteorological forecasting, particularly in global medium-range weather forecasting and large-scale weather and climate modeling. Their success underscores the value of architectures tailored to atmospheric dynamics and large-scale contextual information. The present work addresses a complementary setting of regional short-range multi-variable forecasting, with an emphasis on computational efficiency and dense spatio-temporal learning. Rather than developing a large global forecasting system, our goal is to design an efficient variable-aware backbone and an online selective learning framework that selectively optimizes dense supervision signals and improves generalization for high-resolution regional meteorological fields.

2.3. Training Strategies for Spatio-Temporal Forecasting

Besides architectural design, training strategies play an important role in improving the stability and generalization of spatio-temporal forecasting models. For autoregressive predictors, a major challenge is the discrepancy between training and inference, commonly known as exposure bias. Teacher forcing [42] trains recurrent models by feeding ground-truth observations as inputs during training, whereas inference requires the model to condition on its own predictions. Scheduled sampling [43] gradually replaces ground-truth inputs with model-generated predictions to mitigate this mismatch. Reverse scheduled sampling [27] further improves long-term predictive learning by encouraging models to rely more on generated frames in early training stages. However, these strategies mainly target autoregressive rollout errors and do not directly address the heterogeneity of dense supervision signals in spatio-temporal fields.

Another line of research focuses on curriculum learning, hard example mining, and data selection. Curriculum learning [44] organizes training examples from easy to hard to improve optimization. Online hard example mining [45] prioritizes informative difficult examples during training and has been widely used in dense vision tasks. Data pruning methods aim to identify important training examples and remove redundant ones. For example, EL2N and GraNd scores [46] can identify important examples early in training, enabling large fractions of data to be pruned with limited accuracy degradation. RHO-Loss selection [47] further argues that useful samples should be learnable, worth learning, and not yet learned, and selects training points expected to reduce generalization loss. These studies suggest that not all training signals contribute equally to generalization.

However, most existing training strategies operate at the sample level or rely on predefined schedules. Such designs are suboptimal for dense meteorological forecasting, where informative and low-benefit regions coexist within the same spatio-temporal field. ST-SL addresses this limitation by performing online selection at the spatio-temporal-unit level, using a frozen reference model to construct learnability-aware scores and optimize only selected high-benefit units.

3. Methods

In this section, we present the proposed framework for multi-variable meteorological forecasting. We first formulate the spatio-temporal forecasting problem and introduce the standard full-field regression objective. We then describe Spatio-Temporal Selective Learning (ST-SL), an online selective training framework that extends data selection from the sample level to the spatio-temporal-unit level. Specifically, we introduce the selective loss formulation, the residual-gap-based mask generation mechanism, the online training algorithm, and the computational complexity. Finally, we present VASTFormer, the variable-aware spatio-temporal Transformer backbone used in our experiments.

3.1. Problem Formulation

Multi-variable meteorological forecasting. We consider the task of multi-variable meteorological forecasting, where the goal is to predict the future evolution of atmospheric variables given historical observations. Let $\mathbf{X}_{1:T} \in \mathbb{R}^{T\times C\times H\times W}$ denote the input sequence of meteorological fields over the past T time steps, where C represents the number of meteorological variables, such as temperature, humidity, and the horizontal wind components, and $H \times W$ denotes the spatial grid resolution. Given $\mathbf{X}_{1:T}$, the forecasting model $f(\cdot;\Theta)$ aims to predict the future sequence:

$$\hat{\mathbf{Y}}_{T+1:T+T'} = f(\mathbf{X}_{1:T};\Theta), \quad \hat{\mathbf{Y}}_{T+1:T+T'} \in \mathbb{R}^{T'\times C\times H\times W}, \tag{1}$$

where T' denotes the forecasting horizon and Θ represents the model parameters. The corresponding ground-truth future sequence is denoted as $\mathbf{Y}_{T+1:T+T'}$.

Full-field regression objective. Most data-driven spatio-temporal forecasting models are trained by minimizing a regression loss over all predicted variables, lead times, and spatial locations [8,9,12,13,48]. A commonly used objective is the mean squared error (MSE):

$$\mathcal{L}_{\mathrm{MSE}}(\Theta) = \frac{1}{T'CHW}\sum_{t,c,i,j}\left(Y_{t,c,i,j} - \hat{Y}_{t,c,i,j}\right)^2. \tag{2}$$

Although this full-field objective is simple and effective, it assigns equal optimization weight to all spatio-temporal supervision signals. This treatment may be suboptimal for meteorological forecasting, where the reliability and learning value of supervision signals can vary substantially across variables, lead times, and spatial locations due to observational noise, data assimilation artifacts, unresolved physical processes, and local stochasticity [20–22,49]. This motivates a more fine-grained training strategy that can adaptively focus optimization on informative spatio-temporal units.

3.2. ST-SL: Spatio-Temporal Online Selective Learning Framework

Building on the above motivation, we propose Spatio-Temporal Selective Learning (ST-SL), an online selective training framework tailored for dense spatio-temporal prediction. The overall framework of ST-SL is illustrated in Figure 1. Inspired by recent online data selection methods [47,50–53], ST-SL extends the selection granularity from the sample level to the spatio-temporal-unit level.

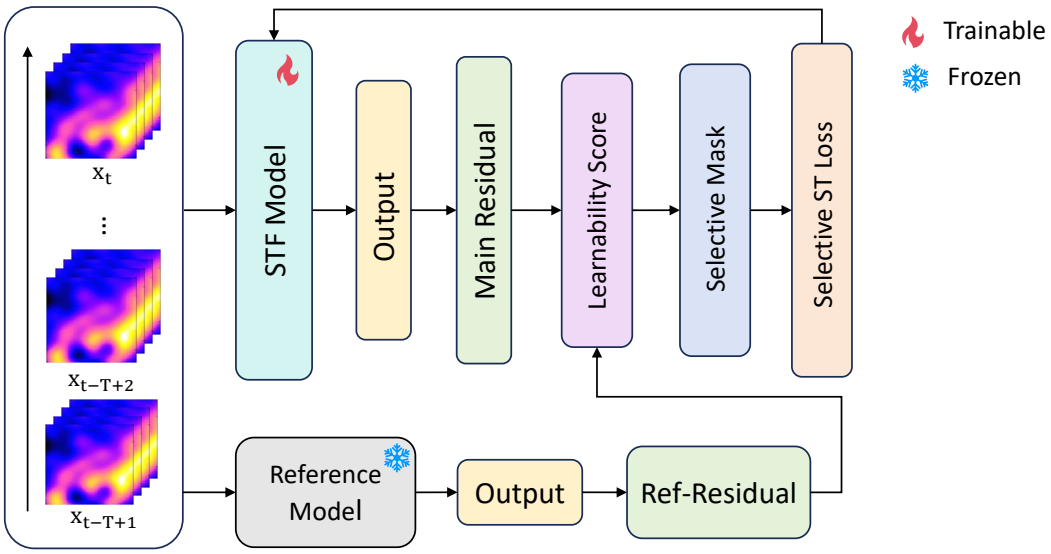


Figure 1. The overall framework of ST-SL. The main Spatio-Temporal Forecasting (STF) model outputs future fields, after which main residuals are computed. A pre-trained frozen reference model provides a lower-bound estimate of the residuals to generate a learnability-aware selective mask. This mask filters out regions with limited generalization potential, confining the selective loss to credible and highly informative areas, thereby enhancing prediction robustness and generalization capabilities in complex spatio-temporal scenarios.

This fine-grained design is particularly suitable for dense meteorological forecasting, where informative and less reliable supervision signals may coexist within the same spatio-temporal field. For example, large-scale atmospheric structures may exhibit relatively stable and learnable evolution patterns, whereas localized regions can be affected by observational noise, data assimilation artifacts, unresolved physical processes, or intrinsically stochastic dynamics. Therefore, discarding or reweighting entire training samples is too coarse for meteorological forecasting. Instead, ST-SL performs online unit-level selection, allowing the model to preserve informative regions while reducing the influence of low-benefit or weakly generalizable supervision signals.

Specifically, for the k-th training iteration, ST-SL constructs a binary spatio-temporal mask $\mathcal{M}^{(k)} \in \{0,1\}^{T' \times C \times H \times W}$ and computes the regression loss only over the selected prediction units:

$$\mathcal{L}_{\text{ST-SL}}^{(k)}(\Theta) = \frac{1}{|\mathcal{M}^{(k)}|} \sum_{t,c,i,j} \mathcal{M}_{t,c,i,j}^{(k)} \left(Y_{t,c,i,j} - \hat{Y}_{t,c,i,j}^{(k)} \right)^2, \tag{3}$$

where $|\mathcal{M}^{(k)}| = \sum_{t,c,i,j} \mathcal{M}_{t,c,i,j}^{(k)}$ denotes the number of selected spatio-temporal units. When $\mathcal{M}_{t,c,i,j}^{(k)} = 1$, the corresponding prediction unit contributes to the loss and gradient computation; otherwise, it is excluded from the current parameter update. In implementation, a small constant can be added to the denominator for numerical stability.

Compared with conventional training paradigms, ST-SL offers three significant benefits: (1) noise suppression, as unreliable or weakly informative units can be actively filtered during optimization; (2) instance-aware adaptivity, as the selective mask is dynamically generated online and tailored to the specific context of each input sample; and (3) fine-grained control, where the selection is performed at the highly granular spatio-temporal unit level rather than the coarse sample level. The key question is how to construct $\mathcal{M}^{(k)}$ in an online and instance-adaptive manner, and encourage the model to focus on informative and learnable spatio-temporal units. ST-SL addresses this question through residual-gap-

based learnability estimation, as detailed below. In our experiments, the main trainable STF model is instantiated by the proposed VASTFormer backbone.

3.2.1. Selective Mask via Residual-Gap Estimation

To identify which spatio-temporal locations still provide meaningful learning signals, we introduce a learnability-aware selective mechanism based on residual lower-bound estimation. The key idea is to utilize a frozen reference model $g(\cdot;\Phi^*)$ as a conservative estimator of the attainable residual level.

Specifically, the reference model is pre-trained independently on the complete training set $\mathcal{D}_{\text{train}}$ by minimizing the standard global regression objective:

$$\Phi^* = \arg\min_{\Phi} \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{X},\mathbf{Y})\in\mathcal{D}_{\text{train}}} \|\mathbf{Y} - g(\mathbf{X};\Phi)\|^2. \tag{4}$$

Once the pre-training is finalized, the parameters Φ^* are kept frozen. Consequently, for any given input instance $\mathbf{X}$, both the trainable main model $f(\cdot;\Theta)$ and the frozen reference model generate predictions. Let $\hat{Y}^{(k)}$ denote the prediction of the main model at the k-th training iteration. For a given spatial location (i,j), variable channel c, and prediction horizon t, the absolute residuals of the main model and the reference model are defined respectively as:

$$\epsilon^{(k)}_{t,c,i,j} = \left|Y_{t,c,i,j} - \hat{Y}^{(k)}_{t,c,i,j}\right|, \quad \epsilon^{\mathrm{LB}(k)}_{t,c,i,j} = \left|Y_{t,c,i,j} - g(X;\Phi^*)^{(k)}_{t,c,i,j}\right|. \tag{5}$$

Since the reference model is frozen throughout training, its residual $\epsilon^{\mathrm{LB}(k)}$ provides a stable baseline against which the current learning state of the main model on the k-th batch can be continuously measured.

Based on these residuals, we evaluate the optimization potential of each unit by defining a learnability-aware score at the k-th iteration as:

$$\mathcal{A}^{(k)}_{t,c,i,j} = \frac{\epsilon^{(k)}_{t,c,i,j} - \epsilon^{\mathrm{LB}(k)}_{t,c,i,j}}{\epsilon^{\mathrm{LB}(k)}_{t,c,i,j} + \delta}, \tag{6}$$

where $\delta > 0$ is a small constant to prevent zero division. This score acts as a learnability-oriented metric with an explicit physical interpretation: when $\mathcal{A} > 0$, the main model's residual is significantly higher than the baseline lower bound, indicating that the model still has ample learning space corresponding to learnable, normal patterns that have not yet been fully mastered; when $\mathcal{A} \approx 0$, the residual of the main model approaches the baseline performance, which likely corresponds to ungeneralizable anomalous patterns or inherent system noise since even the baseline model cannot achieve a better prediction; and when $\mathcal{A} < 0$, the main model outperforms the reference baseline, implying excellent local performance or an overly conservative baseline estimate. Through this continuous numerical characterization, ST-SL precisely distinguishes locations with remaining optimization potential from those with low expected training benefit.

Consequently, rather than penalizing all errors uniformly, ST-SL preferentially retains locations with the largest learnability-aware scores. To maintain instance-aware adaptivity, for each sample in the k-th training batch, we compute a specific dynamic threshold $\tau^{(k)}$ corresponding to the ρ-th empirical quantile of its own learnability-aware scores, where $\rho \in [0,1)$ is the predefined selective ratio. The learnability-aware selective mask is then concisely formulated as:

$$\mathcal{M}_{t,c,i,j}^{(k)} = \begin{cases} 1, & \text{if } \mathcal{A}_{t,c,i,j}^{(k)} \geq \tau^{(k)}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Finally, the selective training loss is computed exclusively over the retained credible elements by Equation (3). Through this mechanism, ST-SL encourages the model to focus its optimization on informative and learnable spatio-temporal transitions while reducing the influence of noisy or low-benefit regions.

3.2.2. Online Adaptive Training Algorithm

Building upon the residual-gap estimation mechanism described above, we present the streamlined online training algorithm for Spatio-Temporal Selective Learning (ST-SL) in Algorithm 1. We assume that an estimation reference model $g(\cdot;\Phi^*)$ has already been pre-trained offline. In our experiments, this reference model is implemented as a frozen version of the VASTFormer backbone pre-trained with standard global regression.

The core workflow operates dynamically at the batch level. For each training iteration, the trainable main network performs a forward pass to generate predictions and compute empirical residuals. Simultaneously, the pre-trained reference network computes the conservative residual lower bounds. These two residual fields are directly compared to calculate the spatio-temporal learnability scores. An instance-aware selective mask is then generated via empirical quantile thresholding. Finally, the selective loss is computed exclusively over the masked informative regions, and the parameters of the main network are updated via gradient backpropagation.

Algorithm 1 ST-SL: Spatio-Temporal Online Selective Training

Input: The main model $f(\cdot;\Theta)$, the training set $\mathcal{D}_{\text{train}}$, the pre-trained reference model $g(\cdot;\Phi^*)$, selective ratio $\rho \in [0,1)$, a small constant $\delta > 0$, and total iterations $K \in \mathbb{N}_+$.
Output: Optimized model parameters Θ_k.

Initialize $f(\cdot;\Theta_0)$
for $k \leftarrow 0, 1, \ldots, K-1$ **do**
  Sample a mini-batch $\mathcal{B} \leftarrow \{(\mathbf{X}, \mathbf{Y})\}$ from $\mathcal{D}_{\text{train}}$
  $\hat{\mathbf{Y}}^{(k)} \leftarrow f(\mathbf{X};\Theta_k)$ *// Forward pass*
  $\hat{\mathbf{Y}}_{\text{LB}} \leftarrow g(\mathbf{X};\Phi^*)$ *// Lower-bound estimation*
  // Calculate absolute residuals
  $\epsilon^{(k)} \leftarrow |\mathbf{Y} - \hat{\mathbf{Y}}^{(k)}|$
  $\epsilon^{\text{LB}(k)} \leftarrow |\mathbf{Y} - \hat{\mathbf{Y}}_{\text{LB}}|$
  Calculate learnability score $\mathcal{A}^{(k)}$ by Equation (6) *// Evaluate learnability*
  // Dynamic learnability masking
  for each sample $b \in \mathcal{B}$ **do**
    $\tau_b^{(k)} \leftarrow$ Top-ρ% $\mathcal{A}_b^{(k)}$ *// Calculate dynamic threshold*
    $\mathcal{M}_b^{(k)} \leftarrow \mathbb{I}(\mathcal{A}_b^{(k)} \geq \tau_b^{(k)})$ *// Generate selective mask*
  end for
  Calculate $\mathcal{L}_{\text{ST-SL}}^{(k)}$ over $\mathcal{M}^{(k)}$ by Equation (3) *// Selective loss*
  $\Theta_{k+1} \leftarrow \Theta_k - \eta \nabla_\Theta \mathcal{L}_{\text{ST-SL}}^{(k)}$ *// Gradient descent optimization*
end for

3.2.3. Computational Complexity

In this subsection, we explicitly analyze the computational overhead introduced by the proposed ST-SL framework compared with conventional training paradigms.

During the training phase, the additional computational cost primarily stems from the forward propagation of the reference network $g(\cdot;\Phi^*)$ and the generation of the dynamic mask. Assuming the reference model shares the same architecture as the main model,

its forward pass complexity is identical to that of the main forecasting model, denoted as $\mathcal{O}(F_{\text{backbone}})$. In deep learning frameworks, standard training requires both forward and backward passes, where the backpropagation typically consumes roughly twice the computational resources of the forward pass. Because the reference model is frozen and does not track gradients, it only introduces a single forward pass overhead. Therefore, the relative training time overhead is moderate, and is approximately consistent with the empirical cost observed in our experiments.

Regarding the masking logic, the learnability score computation purely involves element-wise operations with $\mathcal{O}(BTCHW)$ complexity. The instance-aware thresholding employs a parallel top-k sorting algorithm along the flattened spatial-temporal dimension. Let $N_{st} = T \times C \times H \times W$ denote the number of elements per sample; the sorting complexity is bounded by $\mathcal{O}(B \cdot N_{st} \log N_{st})$. Compared to the $\mathcal{O}(B \cdot N_{\text{patch}}^2 \cdot D)$ self-attention complexity within the backbone, this cost is mathematically trivial.

In terms of memory requirements, ST-SL only requires additional parameters of the frozen reference model, the transient storage of the reference predictions $\hat{\mathbf{Y}}_{\text{LB}}$, and the boolean mask $\mathcal{M}^{(k)}$. More importantly, the selective masking mechanism is exclusively utilized to regulate the gradient flow during the training phase. Once the model converges, both the reference model and the mask generation branch can be safely discarded. Consequently, ST-SL achieves enhanced performance with zero additional computational complexity or latency during inference.

Overall, the proposed ST-SL framework introduces only moderate computational and memory overhead while providing significant improvements in robustness and generalization for spatio-temporal forecasting tasks. In our experiments on ERA5 dataset, the additional training time introduced by ST-SL is approximately 1.3× compared with standard training, while yielding consistent improvements in forecasting accuracy and generalization.

3.3. VASTFormer: Variable-Aware Spatio-Temporal Transformer Backbone

Conventional spatio-temporal models predominantly treat multi-variable meteorological data (e.g., temperature, humidity, wind components) uniformly by concatenating them along the channel dimension. This variable-agnostic paradigm neglects the distinct physical meanings, varying evolutionary rates, and complex interaction dynamics among different meteorological variables. To effectively capture the complex dynamics of atmospheric systems, we propose a transformer-based backbone termed VASTFormer. The architecture is designed to model both spatial correlations across geographical locations and temporal evolution of atmospheric trajectories, and explicitly decomposes the modeling process into four stages: (1) variable-aware multi-variable fusion, (2) physics-enhanced encoder enriched with solar altitude priors, (3) temporal dependency modeling via a transformer-based translator, and (4) spatial reconstruction with skip connections. The overall pipeline is illustrated in Figure 2.

3.3.1. Variable-Aware Multi-Variable Fusion

The lightweight variable-aware Multi-Variable Fusion (MV-Fusion) block is specifically designed to decouple heterogeneous meteorological variables and capture their physical correlations without introducing excessive computational overhead.

Given the high-dimensional input features $\mathbf{X} \in \mathbb{R}^{B \times T \times (V \times L) \times H \times W}$, where V and L denote the number of meteorological variables and pressure levels, respectively, we first reshape and group $\mathbf{X}$ into $\mathbf{X}_{\text{grp}} \in \mathbb{R}^{B \times T \times V \times L \times H \times W}$. Each variable-level pair is then treated as an individual token source.

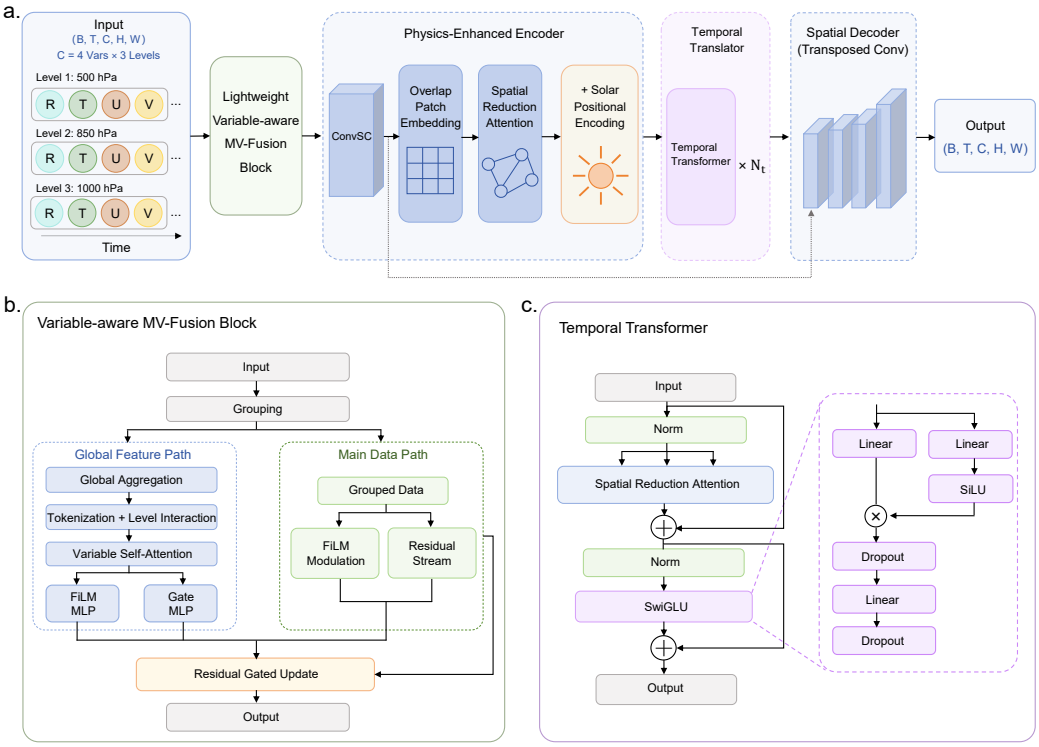


Figure 2. Overall architecture of the proposed VASTFormer. (**a**) The complete network pipeline, which processes multi-variable input across different atmospheric levels through a lightweight Variable-aware Multi-Variable Fusion (MV-Fusion) Block, a Spatial Encoder augmented with Solar Positional Encoding, a Temporal Translator, and a Spatial Decoder. (**b**) Detailed structure of the Variable-aware MV-Fusion Block, illustrating the feature interaction between the Global Feature Path and the Main Data Path prior to the Residual Gated Update. (**c**) Internal architecture of the Temporal Transformer block, featuring Spatial Reduction Attention and a SwiGLU-based feed forward network. The symbols "+" and "×" denote element-wise addition and multiplication, respectively.

To construct compact descriptors for these variable-level tokens, we apply global aggregation to each grouped variable-level field. In practice, this aggregation can preserve multi-scale spatial context by combining adaptive average pooling at multiple spatial scales with global average pooling. The resulting descriptors are projected into a latent embedding space through a lightweight MLP. During tokenization, a learnable variable embedding $\mathbf{E}_{\text{var}} \in \mathbb{R}^{V \times d_{\text{var}}}$ is added to the corresponding level tokens of each variable to provide explicit variable identity priors, yielding token features $\mathbf{Z} \in \mathbb{R}^{B \times T \times V \times L \times d_{\text{var}}}$.

A lightweight level-wise interaction is then incorporated within each variable to allow information exchange across different pressure levels. After reshaping $\mathbf{Z}$ into $\mathbf{Z}_{\text{seq}} \in \mathbb{R}^{(BT) \times (VL) \times d_{\text{var}}}$, the resulting $V \times L$ variable-level tokens are processed by a self-attention module:

$$\mathbf{Z}_{\text{attn}} = \text{LayerNorm}(\mathbf{Z}_{\text{seq}} + \text{MSA}(\mathbf{Z}_{\text{seq}})), \tag{8}$$

$$\mathbf{Z}_{\text{inter}} = \text{LayerNorm}(\mathbf{Z}_{\text{attn}} + \text{FFN}(\mathbf{Z}_{\text{attn}})). \tag{9}$$

The interacted tokens $\mathbf{Z}_{\text{inter}}$ are then reshaped back to the variable-level format.

Instead of directly concatenating $\mathbf{Z}_{\text{inter}}$ back to the spatial features, we employ a Feature-wise Linear Modulation (FiLM) and residual gating mechanism. The interacted tokens are fed into separate MLPs to predict the affine modulation parameters γ, β and a dynamic gating map $\mathbf{G}$ for each variable-level pair. These parameters are broadcast over the spatial dimensions and used to modulate the grouped input features:

$$\mathbf{X}_{\text{FiLM}} = \mathbf{X}_{\text{grp}} \odot (1 + \boldsymbol{\gamma}) + \boldsymbol{\beta}, \tag{10}$$

where $\odot$ denotes the Hadamard product. The final variable-aware fused feature $\tilde{\mathbf{X}}$ is obtained through a scaled residual gated update:

$$\tilde{\mathbf{X}} = \mathbf{X}_{\text{grp}} + \alpha \mathbf{G} \odot \left(\mathbf{X}_{\text{FiLM}} - \mathbf{X}_{\text{grp}}\right), \tag{11}$$

where α is a learnable scaling scalar that controls the strength of the modulation. Finally, $\tilde{\mathbf{X}}$ is reshaped back to $\mathbb{R}^{B\times T\times (V\times L)\times H\times W}$ and passed to the subsequent physics-enhanced encoder. This architecture adaptively modulates high-dimensional spatio-temporal representations according to variable-level interactions, while preserving fine-grained local details through the residual connection.

3.3.2. Physics-Enhanced Encoder

Following the MV-Fusion, the modulated feature $\tilde{\mathbf{X}} \in \mathbb{R}^{B\times T\times C\times H\times W}$ is processed by an initial convolutional layer (ConvSC) and an overlap patch embedding layer. This operation spatially downsamples the input and projects the channels into a latent space, producing tokenized features. These tokens are then fed into a spatial attention block with Spatial Reduction Attention to efficiently extract spatial structural dependencies. The encoded tokens are reshaped back into a spatial grid representation, yielding the intermediate spatio-temporal features $\mathcal{H}_{\text{spat}} \in \mathbb{R}^{B\times T\times D\times H'\times W'}$, where H' and W' denote the reduced spatial dimensions and D is the embedding dimension.

A critical limitation of standard spatio-temporal Transformers is their reliance on absolute or relative mathematical positional encodings, which are usually defined by discrete spatial or temporal indices and are not explicitly aligned with physical atmospheric drivers. In meteorological forecasting, however, many near-surface and lower-tropospheric variations are strongly related to geographically dependent solar forcing. The solar altitude angle, i.e., the angle of the sun relative to the Earth's horizon, provides an informative physical proxy for diurnal and seasonal radiative effects [54].

To inject this physical prior, we design a Solar Positional Encoding (Solar PE). Unlike conventional positional encodings that separately represent spatial and temporal indices, Solar PE provides a unified spatio-temporal prior because the solar geometry is jointly determined by latitude ϕ, longitude λ, and UTC time t. For each spatial grid at a specific time, the solar altitude angle θ_s is determined by:

$$\sin(\theta_s) = \sin(\phi)\sin(\delta) + \cos(\phi)\cos(\delta)\cos(\omega), \tag{12}$$

where δ is the solar declination determined by the day of the year, and ω is the hour angle derived from the local solar time.

Instead of using a purely index-based temporal embedding or a single scalar solar-altitude encoding, we construct a solar-geometry feature vector for each spatio-temporal location. This vector combines the normalized solar altitude angle, angular functions of the solar altitude, the local diurnal phase, and the seasonal phase:

$$\mathbf{s}(x,y,t) = \left[\tilde{\theta}_s,\ \sin(\theta_s),\ \cos(\theta_s),\ \sin\left(\frac{2\pi h_{\text{local}}}{24}\right),\ \cos\left(\frac{2\pi h_{\text{local}}}{24}\right),\ \sin\left(\frac{2\pi d}{365}\right),\ \cos\left(\frac{2\pi d}{365}\right)\right], \tag{13}$$

where h_{local} denotes the longitude-corrected local solar hour and d denotes the day of year. The vector $\mathbf{s}(x,y,t)$ jointly encodes solar altitude, local diurnal phase, and seasonal phase, thereby providing a physically meaningful spatio-temporal prior for geographically

dependent atmospheric variations. It is then projected into the hidden dimension through a lightweight learnable projection:

$$\mathcal{P}_{\text{solar}} = \psi_{\text{solar}}(\mathbf{s}), \quad \mathcal{P}_{\text{solar}} \in \mathbb{R}^{B\times T\times D\times H'\times W'}. \tag{14}$$

The spatio-temporal features are subsequently enriched with this physical prior:

$$\mathcal{H}_{\text{phys}} = \mathcal{H}_{\text{spat}} + \gamma\,\mathcal{P}_{\text{solar}}, \tag{15}$$

where γ is a learnable scaling parameter controlling the contribution of Solar PE. By injecting this solar-geometry prior, VASTFormer reduces the burden of implicitly learning geographically dependent diurnal and seasonal regularities purely from data. The physically enriched representation $\mathcal{H}_{\text{phys}}$ is then passed to the Temporal Translator for long-range atmospheric trajectory modeling.

3.3.3. Temporal Translator

The temporal translator is designed to model the non-linear evolutionary trajectories of meteorological states across the temporal horizon. Unlike standard sequence models that treat time steps in isolation, our translator leverages the global receptive field of the Transformer to capture long-range dependencies, such as large-scale atmospheric circulation patterns.

After the injection of Solar PE, the physically enriched spatio-temporal features $\mathcal{H}_{\text{phys}} \in \mathbb{R}^{B\times T\times D\times H'\times W'}$ are fed into a sequence of N_t Temporal Transformer blocks. To overcome the computational overhead and fragmentation of local physical states in conventional temporal attention, VASTFormer collapses the tensor along the temporal and channel dimensions. It is reshaped and transposed into a sequence of spatial tokens $\mathbf{Z}_{\text{in}} \in \mathbb{R}^{B\times(H'W')\times(T\times D)}$. In this paradigm, the tokens correspond to spatial locations, while their embedded features encapsulate the full multi-variable temporal trajectories. A linear projection $\mathbf{W}_{\text{in}}$ maps this high-dimensional temporal feature into a latent space of size d_{model}.

Subsequently, these spatial tokens are processed by the Temporal Transformer blocks equipped with Spatial Reduction Attention (SRA) and a SwiGLU-based feed forward network (FFN). Let $\mathbf{Z}^{(0)} = \mathbf{Z}_{\text{in}}\mathbf{W}_{\text{in}} \in \mathbb{R}^{B\times N\times d_{\text{model}}}$ denote the projected tokens, where $N = H'W'$ is the sequence length. To alleviate the $\mathcal{O}(N^2)$ quadratic complexity of standard self-attention, SRA reduces the spatial sequence length of keys and values. The core attention mechanism is formulated as:

$$\text{SRAttention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{16}$$

where $\mathbf{Q} \in \mathbb{R}^{B\times N\times d_k}$ represents the queries derived from the input tokens, while $\mathbf{K},\mathbf{V} \in \mathbb{R}^{B\times N_r\times d_k}$ represent the keys and values derived from a spatially reduced projection of the input tokens, with $N_r < N$. $\sqrt{d_k}$ is the scaling factor. This design preserves long-range spatial interaction while reducing the attention cost.

In the ℓ-th Transformer Block, the attention output is further refined by the gated FFN:

$$\tilde{\mathbf{Z}}^{(\ell)} = \mathbf{Z}^{(\ell-1)} + \text{SRAttention}(\text{LN}(\mathbf{Z}^{(\ell-1)})), \tag{17}$$

$$\mathbf{Z}^{(\ell)} = \tilde{\mathbf{Z}}^{(\ell)} + \text{FFN}_{\text{SwiGLU}}(\text{LN}(\tilde{\mathbf{Z}}^{(\ell)})), \tag{18}$$

where $\ell = 1,\ldots,N_t$. The SwiGLU-based FFN is defined as:

$$\text{FFN}_{\text{SwiGLU}}(\mathbf{X}) = [\text{SiLU}(\mathbf{X}\mathbf{W}_1)\odot(\mathbf{X}\mathbf{W}_2)]\mathbf{W}_3, \tag{19}$$

where $\odot$ denotes element-wise multiplication, $\mathrm{SiLU}(x) = x\sigma(x)$ is the Swish activation, and $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$ and $\mathbf{W}_3 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$ are learnable projections. Compared with a conventional FFN, this gated formulation adaptively modulates the activated hidden representation, improving the expressiveness of temporal trajectory modeling.

By computing attention across these spatial tokens that contain complete multi-variable temporal dynamics, this mechanism efficiently models how the weather trajectory at one geographical location dynamically influences the future trajectories of distant regions. Finally, a linear projection $\mathbf{W}_{\text{out}}$ maps the latent representations from size d_{model} back to the original $(T \times D)$ dimension, and the output is reshaped back to the spatial grid format $\mathcal{H}_{\text{temp}} \in \mathbb{R}^{B \times (T \times D) \times H' \times W'}$.

3.3.4. Decoder with Skip Connection

The final stage of VASTFormer aims to reconstruct the high-resolution future meteorological fields $\hat{\mathbf{Y}} \in \mathbb{R}^{B \times T \times C \times H \times W}$ from the deep latent representations $\mathcal{H}_{\text{temp}}$. We employ a time-separable spatial decoder enriched with a late skip connection.

Rather than naïvely concatenating frame-by-frame features at the latent resolution, we extract the shallow high-resolution spatial feature $\mathbf{F}_{\text{low}} \in \mathbb{R}^{B \times T \times C_{\text{low}} \times H \times W}$ from the initial encoder layer. To provide a compact high-resolution skip prior, the temporal and channel dimensions of $\mathbf{F}_{\text{low}}$ are first merged and then projected by a lightweight 1×1 convolution:

$$\mathbf{F}_{\text{skip}} = \Pi_{\text{skip}}(\text{Reshape}(\mathbf{F}_{\text{low}})), \tag{20}$$

where $\text{Reshape}(\mathbf{F}_{\text{low}}) \in \mathbb{R}^{B \times (TC_{\text{low}}) \times H \times W}$ and $\Pi_{\text{skip}}(\cdot)$ denotes the skip projection. This learned projection preserves high-resolution spatial details from shallow encoder features while avoiding direct frame-wise concatenation of redundant temporal information.

The latent representation $\mathcal{H}_{\text{temp}} \in \mathbb{R}^{B \times (T \times D) \times H' \times W'}$ is decoded in a time-separable manner. Specifically, $\mathcal{H}_{\text{temp}}$ is reshaped into timestep-specific latent features, and each timestep is upsampled through transposed convolutional blocks:

$$\mathbf{U}^{(t)} = \text{ConvT}_{N_s}\Big(\ldots \text{ConvT}_1(\mathcal{H}_{\text{temp}}^{(t)})\Big), \tag{21}$$

where $\mathbf{U}^{(t)} \in \mathbb{R}^{B \times D \times H \times W}$ denotes the upsampled feature of the t-th timestep. After upsampling, the high-resolution skip feature is fused with the decoded feature at the full spatial resolution:

$$\mathbf{H}_{\text{fuse}}^{(t)} = \text{Compress}\Big([\mathbf{U}^{(t)}; \mathbf{F}_{\text{skip}}]\Big), \tag{22}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation and $\text{Compress}(\cdot)$ is implemented by a 1×1 convolution.

Finally, a refinement convolution and a readout convolution map the fused hidden features back to the original physical channel dimension, yielding the predicted future sequence $\hat{\mathbf{Y}} \in \mathbb{R}^{B \times T \times C \times H \times W}$. This reconstruction strategy helps preserve fine-grained spatial details from shallow features while allowing the deep temporal representation to control the predicted meteorological evolution.

The proposed VASTFormer explicitly disentangles the complex meteorological modeling process into variable interaction, spatial representation, and temporal dynamics. By integrating variable-aware multi-variable fusion and physics-enhanced solar positional encoding, alongside a highly efficient spatial-token temporal translator, the architecture effectively overcomes the computational bottlenecks and physical fragmentation inherent in standard Transformers. Consequently, VASTFormer emerges as a highly scalable and physically grounded backbone, particularly well-suited for high-resolution multi-variable meteorological forecasting tasks.

4. Results

In this section, we systematically evaluate the proposed VASTFormer and ST-SL framework. We first introduce the datasets, evaluation metrics, baselines, and implementation details. We then compare VASTFormer with representative spatio-temporal predictive models under the same regional short-range forecasting protocol, and examine whether ST-SL further improves accuracy and generalization through adaptive unit filtering. To strengthen the meteorological interpretation, we additionally report ACC/RMSESS, percentile-based extreme-condition verification, and terrain-related error analyses. Ablation studies are then used to evaluate the individual contributions of ST-SL, variable-aware fusion, Solar Positional Encoding, decoder skip connection, and the selective ratio. We further provide qualitative visualization analysis of prediction results and spatial error distributions across different methods. Finally, we analyze the practical computational overhead introduced by the ST-SL framework during both the training and inference phases.

4.1. Experimental Setup

4.1.1. Datasets and Preprocessing

We conduct extensive experiments on the ERA5 reanalysis dataset [20]. Produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), ERA5 is the fifth-generation global atmospheric reanalysis dataset, providing high-resolution global gridded atmospheric variables with a spatial resolution of $0.25^\circ \times 0.25^\circ$ and a temporal resolution of 1 h.

In this study, our research scope covers China and Southeast Asia, spanning from 70°E to 140°E in longitude and from the Equator (0°) to 55°N in latitude (Figure 3). All input fields are resized to a spatial resolution of 192×256 for model training and evaluation. We select four core meteorological variables across three distinct pressure levels (500 hPa, 850 hPa, and 1000 hPa), yielding a total of 12 physical channels. The variables include Relative Humidity (R), Temperature (T), the U-component of wind (U), and the V-component of wind (V).

We utilize five years of continuous ERA5 data from 2016 to 2020. Following chronological splitting practices commonly used in data-driven weather forecasting benchmarks [37], the dataset is divided as follows: the period from 2016 to 2018 is used for training, 2019 for validation, and 2020 for testing. To ensure numerical stability during optimization, all variables are standardized to zero mean and unit variance based on the training set statistics before being fed into the models.

4.1.2. Evaluation Metrics

We assess the forecasting performance using three standard meteorological evaluation metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Given the prediction $\hat{\mathbf{Y}}$ and the ground truth $\mathbf{Y}$, these metrics are computed over all predicted time steps, variables, and spatial locations. Lower values for all three metrics indicate higher grid-level accuracy and better forecasting performance. MSE emphasizes large errors, MAE reflects the average absolute deviation, and RMSE provides an error measure in the same scale as the target variables.

To better align the evaluation protocol with operational meteorological verification, we further compute the Anomaly Correlation Coefficient (ACC) and the Root Mean Square Error Skill Score (RMSESS). For each forecast lead time and variable group, ACC measures the spatial correlation between the predicted and observed anomaly fields:

$$\mathrm{ACC} = \frac{\sum_i (\hat{y}_i - c_i)(y_i - c_i)}{\sqrt{\sum_i (\hat{y}_i - c_i)^2}\sqrt{\sum_i (y_i - c_i)^2}}, \tag{23}$$

where $\hat{y}_i$, y_i, and c_i denote the predicted field, the ground-truth field, and the climatological mean at grid point i, respectively. RMSESS quantifies the relative improvement over the climatological forecast:

$$\mathrm{RMSESS} = 1 - \frac{\mathrm{RMSE}_{\mathrm{model}}}{\mathrm{RMSE}_{\mathrm{clim}}}. \tag{24}$$

The climatological baseline is constructed from the training years as a day-of-year mean field for each variable, pressure level, and grid point after applying the same center crop used in model evaluation. This protocol avoids using test-year information in the baseline while preserving the seasonal spatial structure of the regional ERA5 fields.

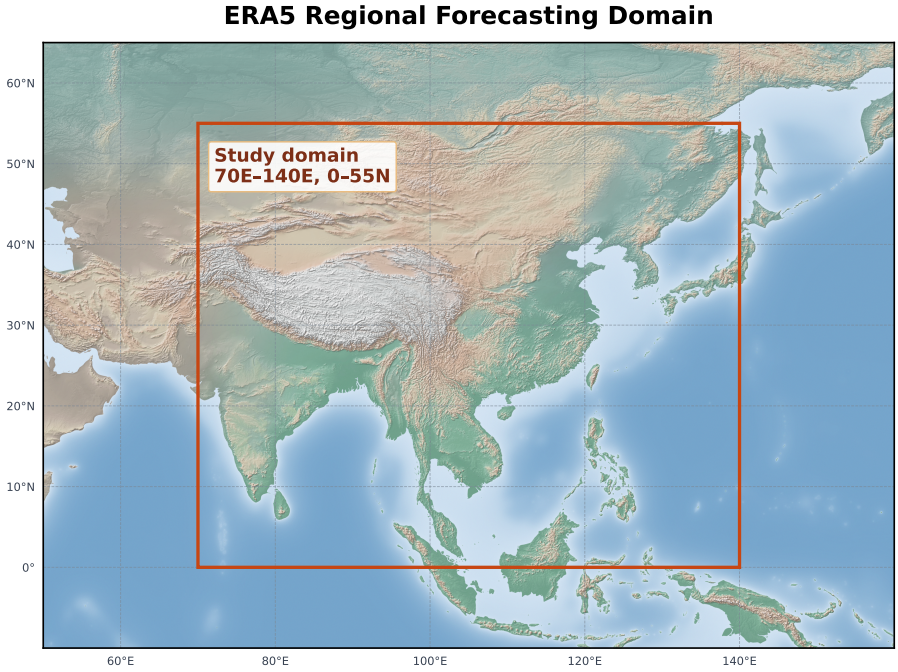


Figure 3. Study area used for regional ERA5 forecasting. The highlighted domain spans 70°E–140°E and 0°–55°N, covering China and Southeast Asia.

4.1.3. Baselines

We compare the proposed method with a broad range of representative and competitive spatio-temporal prediction models under the same regional ERA5 forecasting setting. The baselines include two non-learning reference forecasts and a set of trainable neural predictors.

(1) Non-learning Reference Forecasts:

1. **Persistence**: This reference forecast repeats the last observed input field for all future lead times. It tests whether a model can improve over simple short-range temporal persistence.
2. **Climatology**: This reference forecast uses the day-of-year climatological mean field constructed from the training years for each variable, pressure level, and grid point. It provides a seasonal background reference without using test-year information.

These references provide basic meteorological sanity checks for the short-range forecasting setting. The trainable baselines encompass various foundational architectures:

(2) RNN-based Models:

3. **ConvLSTM** [8]: The pioneering model that integrates convolutional operations into LSTMs for spatio-temporal sequence forecasting.

4. **PredRNN** [9]: Introduces spatio-temporal LSTM (ST-LSTM) cells with a zigzag memory flow to capture spatial and temporal dynamics simultaneously.
5. **PredRNN++** [27]: An enhancement of PredRNN that utilizes causal LSTMs and a gradient highway unit to capture short-term and long-term dependencies.
6. **MIM** [28]: Replaces the standard forget gates in ST-LSTM with non-stationary and stationary modules to model complex spatio-temporal dynamics.
7. **MAU** [34]: Employs a motion-aware unit to explicitly capture inter-frame motions alongside an attention mechanism for broader temporal receptive fields.
8. **PredRNNv2** [29]: Further improves the memory mechanism in PredRNN by decoupling spatial and temporal memories to alleviate predictive forgetting.

(3) CNN and Transformer-based Models:

9. **SimVP** [10]: A simple yet highly effective fully-convolutional baseline that disentangles spatio-temporal modeling into spatial encoding, temporal translation, and spatial decoding.
10. **SimVPv2** [33]: An evolution of SimVP that introduces an inception-like temporal translation module to capture multi-scale temporal dynamics.
11. **TAU** [11]: A temporal attention unit-based architecture that leverages self-attention for highly efficient inter-frame temporal evolution modeling.
12. **OpenSTL_ViT** [48]: A Vision Transformer (ViT)-based baseline within the OpenSTL framework that models spatio-temporal data purely through self-attention mechanisms.
13. **OpenSTL_Swin** [48]: Utilizes Swin Transformers to introduce shifted window-based local attention, enhancing computational efficiency for high-resolution predictions.
14. **PredFormer** [13]: A unified Transformer architecture specifically designed for video and spatio-temporal prediction, capturing multi-scale spatio-temporal dependencies.

All trainable baselines are trained and evaluated under the same input-output setting and data split for fair comparison. Large global weather forecasting systems are discussed in Section 2.2 as complementary meteorological context rather than direct baselines, because their standard settings differ substantially in terms of spatial domain, input configuration, forecast horizon, and training regime.

4.1.4. Implementation Details

All experiments are implemented using PyTorch version 2.9.1 with CUDA 12.8 and Python 3.10. The VASTFormer backbone and the ST-SL framework are trained end-to-end. Both the historical input sequence length and the forecasting horizon are set to 6 (i.e., utilizing the past 6 h to predict the future 6 h). The models are optimized using the AdamW [55] optimizer for 50 epochs, with an initial learning rate of 5×10^{-4}, a weight decay of 1×10^{-2}, and a batch size of 4. For the proposed ST-SL training strategy, the reference model is first trained using the standard full-field regression objective and then kept frozen during selective training. Unless otherwise specified, the same optimization settings are used for all compared methods. All trainable models are trained and evaluated on NVIDIA GeForce RTX 3090 GPUs.

4.2. Main Forecasting Results

Table 1 reports the comprehensive quantitative evaluation of our proposed models, the 12 competitive trainable baselines, and two non-learning reference forecasts on the ERA5 dataset. Overall, VASTFormer achieves the best performance among all backbone-only models, and VASTFormer equipped with ST-SL further improves the forecasting accuracy. Based on these empirical results, we draw the following key observations.

Superiority of the VASTFormer Backbone. The proposed VASTFormer achieves state-of-the-art performance among the evaluated lightweight spatio-temporal forecasting baselines under the regional ERA5 protocol, outperforming all compared methods by a clear margin. The persistence and climatology references are substantially weaker than the trainable neural models on the overall average, confirming that the benchmark requires learning non-trivial atmospheric evolution rather than simply repeating the latest state or seasonal mean. Among the baselines, TAU obtains the lowest MSE and RMSE, while PredRNN achieves the lowest MAE. Compared with these strongest baseline results, VASTFormer reduces MSE from 6.4211 to 5.9493, MAE from 1.1600 to 1.0896, and RMSE from 2.5306 to 2.4351, corresponding to relative improvements of 7.35%, 6.07%, and 3.77%, respectively. These results demonstrate that explicitly modeling variable interactions, physics-enhanced spatio-temporal representations, and trajectory-aware spatio-temporal dependencies leads to more accurate meteorological forecasting.

Enhancement via Selective Learning. The integration of the Spatio-Temporal Selective Learning framework yields the best performance across all metrics. VASTFormer+ST-SL achieves an MSE of 5.8532, an MAE of 1.0823, and an RMSE of 2.4156. Compared with VASTFormer trained using the standard full-field regression objective, ST-SL further reduces MSE, MAE, and RMSE by 1.62%, 0.67%, and 0.80%, respectively. Compared with the strongest baseline results, VASTFormer+ST-SL improves MSE, MAE, and RMSE by 8.84%, 6.70%, and 4.54%, respectively. This indicates that selectively optimizing high-benefit spatio-temporal units can improve generalization beyond what is achieved by architectural design alone.

Computational Efficiency Trade-offs. Furthermore, VASTFormer achieves a favorable complexity-accuracy trade-off. For example, recurrent models such as ConvLSTM, PredRNN, and PredRNN++ require substantially higher computational costs, with FLOPs ranging from 0.940 T to 1.634 T, but still produce larger forecasting errors. PredRNN++ contains 42.55 M parameters, more than twice that of VASTFormer, yet its MSE remains much higher. In contrast, VASTFormer remains remarkably lightweight with only 17.36 M parameters and 0.102T FLOPs, which is even smaller than TAU in both parameter count and computational cost, while achieving clearly better forecasting accuracy. These comparisons indicate that the proposed framework improves forecasting accuracy without relying on excessive model size or computational complexity. Most importantly, ST-SL introduces no additional inference-time parameters or FLOPs compared with VASTFormer. This is because the reference model and the selective mask generation branch are used only during training and discarded during inference. Therefore, ST-SL provides a pure training-time improvement while preserving the deployment efficiency of the original forecasting backbone.

Table 1. Overall forecasting performance on the ERA5 dataset. Lower MSE, MAE, and RMSE indicate better performance. The best results are highlighted in **bold**, and the second-best are underlined. Params and FLOPs are marked as "–" for non-learning reference forecasts.

Methods	Params (M)	FLOPs (T)	MSE ↓	MAE ↓	RMSE ↓
Persistence	–	–	38.4118	2.6376	6.1977
Climatology	–	–	148.9972	6.7274	12.2064
ConvLSTM	15.50	1.634	7.2739	1.2403	2.6931
PredRNN	27.80	0.940	6.5081	1.1600	2.5473
PredRNN++	42.55	1.327	6.8921	1.2055	2.6214
MIM	29.42	0.796	6.8661	1.1997	2.6159
MAU	26.60	0.279	8.1317	1.3349	2.8474
PredRNNv2	27.82	0.944	6.6480	1.1738	2.5745

Table 1. *Cont.*

Methods	Params (M)	FLOPs (T)	MSE ↓	MAE ↓	RMSE ↓
SimVP	32.34	0.134	8.4016	1.3705	2.8946
SimVPv2	24.62	0.110	6.6491	1.1920	2.5751
TAU	23.54	0.106	6.4211	1.1691	2.5306
OpenSTL_ViT	24.13	0.166	6.9815	1.2216	2.6382
OpenSTL_Swin	24.14	0.109	6.6275	1.1840	2.5705
PredFormer	24.26	0.134	6.8615	1.2072	2.6157
VASTFormer (Ours)	17.36	0.102	5.9493	1.0896	2.4351
VASTFormer+ST-SL (Ours)	17.36	0.102	**5.8532**	**1.0823**	**2.4156**

4.3. Meteorological Skill Evaluation

To further assess the operational meteorological skill of the proposed framework, Table 2 reports ACC and RMSESS across forecast lead times. Higher ACC values indicate better agreement with observed anomaly patterns, while higher RMSESS values indicate stronger predictive skill relative to the climatological reference. PredFormer and TAU are included as representative lightweight spatio-temporal baselines, with TAU being the strongest baseline according to the overall RMSE results in Table 1. Persistence and climatology are also reported as non-learning reference forecasts. ACC is not defined for the climatological forecast itself because its anomaly relative to the same climatology is zero.

Table 2. Meteorological skill evaluation on the ERA5 test set. ACC and RMSESS are computed with respect to a day-of-year climatology constructed from the training years. Higher values indicate better forecast skill. The best results are highlighted in **bold**, and the second-best are underlined. A dash (–) indicates that ACC is not defined for the climatological forecast because the climatological anomaly relative to itself is zero.

Metric	Method	T+1	T+2	T+3	T+4	T+5	T+6	Avg.
ACC	Persistence	0.9824	0.9411	0.8964	0.8494	0.8005	0.7568	0.8711
	Climatology	–	–	–	–	–	–	–
	PredFormer	0.9933	0.9853	0.9799	0.9744	0.9665	0.9605	0.9766
	TAU	0.9933	0.9859	0.9811	0.9760	0.9683	0.9626	0.9779
	VASTFormer	0.9960	0.9881	0.9830	0.9778	0.9699	0.9639	0.9798
	VASTFormer+ST-SL	**0.9960**	**0.9881**	**0.9832**	**0.9781**	**0.9705**	**0.9647**	**0.9801**
RMSESS	Persistence	0.8122	0.6568	0.5449	0.4511	0.3685	0.3027	0.5227
	Climatology	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	PredFormer	0.8843	0.8289	0.8006	0.7751	0.7434	0.7217	0.7923
	TAU	0.8842	0.8328	0.8065	0.7822	0.7502	0.7290	0.7975
	VASTFormer	0.9111	0.8459	0.8167	0.7905	0.7566	0.7336	0.8090
	VASTFormer+ST-SL	**0.9112**	**0.8463**	**0.8175**	**0.7919**	**0.7588**	**0.7366**	**0.8104**

Persistence is competitive only at the shortest forecast step, with ACC decreasing from 0.9824 at T+1 to 0.7568 at T+6 and RMSESS decreasing from 0.8122 to 0.3027. This rapid degradation shows that simple short-range extrapolation is insufficient for the full six-hour prediction task. PredFormer obtains an average ACC of 0.9766 and an average RMSESS of 0.7923, while TAU improves these values to 0.9779 and 0.7975, respectively. Compared with TAU, VASTFormer improves the average ACC from 0.9779 to 0.9798 and the average RMSESS from 0.7975 to 0.8090. VASTFormer+ST-SL further improves the average ACC to 0.9801 and the average RMSESS to 0.8104. The improvement introduced by ST-SL is more evident at longer lead times: at T+6, RMSESS increases from 0.7336 to 0.7366 and ACC increases from 0.9639 to 0.9647 compared with VASTFormer. These results

confirm that the proposed architecture and selective supervision strategy improve not only average regression errors but also standard meteorological forecast skill relative to a climatological baseline.

To further inspect the channel-level behavior at the longest evaluated lead time, Table 3 reports the T+6 ACC and RMSESS for all 12 variable-level channels. VASTFormer+ST-SL achieves the highest skill across all channels, indicating that the improvement is not driven by a single variable or pressure level.

Table 3. Channel-wise meteorological skill at T+6 with respect to the day-of-year climatology. Higher values indicate better forecast skill. The best results are highlighted in **bold**.

Channel	PredFormer		TAU		VASTFormer		VASTFormer+ST-SL	
	ACC	RMSESS	ACC	RMSESS	ACC	RMSESS	ACC	RMSESS
RH500	0.9580	0.7133	0.9603	0.7210	0.9608	0.7225	**0.9619**	**0.7265**
RH850	0.9596	0.7187	0.9621	0.7270	0.9636	0.7325	**0.9642**	**0.7347**
RH1000	0.9564	0.7078	0.9580	0.7133	0.9620	0.7268	**0.9624**	**0.7283**
T500	0.9869	0.8389	0.9869	0.8387	0.9877	0.8440	**0.9881**	**0.8461**
T850	0.9875	0.8426	0.9879	0.8446	0.9890	0.8519	**0.9892**	**0.8533**
T1000	0.9815	0.8083	0.9819	0.8102	0.9846	0.8250	**0.9847**	**0.8260**
U500	0.9820	0.8113	0.9834	0.8184	0.9833	0.8182	**0.9838**	**0.8207**
U850	0.9763	0.7835	0.9781	0.7916	0.9783	0.7928	**0.9789**	**0.7957**
U1000	0.9680	0.7492	0.9696	0.7550	0.9715	0.7631	**0.9722**	**0.7658**
V500	0.9753	0.7792	0.9767	0.7855	0.9763	0.7837	**0.9772**	**0.7875**
V850	0.9706	0.7592	0.9724	0.7667	0.9724	0.7669	**0.9733**	**0.7705**
V1000	0.9694	0.7543	0.9707	0.7597	0.9721	0.7653	**0.9728**	**0.7681**

4.4. Extreme-Condition Evaluation

To evaluate model behavior beyond average meteorological conditions, we further construct percentile-based proxy extreme subsets from the available ERA5 variables. Specifically, high-relative-humidity, high-temperature, low-temperature, and strong-wind cases are defined using level-wise percentile thresholds estimated from the observed 2020 test targets. These thresholds are used only for evaluation stratification and are not used during model training or inference. For the 500/850/1000 hPa levels, the resulting thresholds are 97.37/96.60/95.37% for high relative humidity, 271.11/291.88/304.39 K for high temperature, 243.05/262.45/271.51 K for low temperature, and 28.38/14.47/10.66 m s^{-1} for strong wind speed. For relative humidity and temperature, MAE, MSE, and RMSE are computed over the selected variable values. For strong wind, we evaluate wind-speed errors computed from the U and V components. In addition to intensity errors, we report event-detection metrics based on hits (H), misses (M), and false alarms (F): CSI $= H/(H+M+F)$, POD $= H/(H+M)$, and FAR $= F/(H+F)$.

Table 4 shows that the proposed VASTFormer backbone substantially improves forecasting accuracy under these challenging regimes. Compared with the strongest lightweight baseline for each metric, the best proposed model reduces RMSE by 5.73% for high relative humidity, 10.16% for high temperature, 16.26% for low temperature, and 3.29% for strong wind. For event detection, VASTFormer+ST-SL obtains the highest CSI and the lowest FAR across all four extreme-condition categories. Although some baselines have slightly higher POD in high-temperature and low-temperature cases, this is accompanied by higher false-alarm rates; therefore, CSI and FAR provide a more balanced indication of event-detection skill. VASTFormer achieves the best high-humidity intensity errors and the lowest high-temperature MAE, while VASTFormer+ST-SL obtains the best high-temperature RMSE, low-temperature performance, strong-wind performance, and the most balanced detection skill. These results indicate that the proposed framework remains effective under percentile-defined extreme states, while also clarifying that event-specific

verification for typhoons or heavy rainfall would require additional variables and event labels beyond the current setting.

Table 4. Extreme-condition evaluation on percentile-defined subsets. Thresholds are estimated separately for each pressure level from the observed 2020 test targets. For strong wind, errors are computed on wind speed derived from U and V components. Lower values are better for MAE, MSE, RMSE, and FAR; higher values are better for CSI and POD. The best results are highlighted in **bold**, and the second-best are underlined.

Event	Metric	PredFormer	TAU	VASTFormer	VASTFormer+ST-SL
High RH	MAE	3.5411	3.5249	**3.2019**	3.2231
	MSE	30.9109	29.4289	**26.1515**	26.3573
	RMSE	5.5598	5.4248	**5.1139**	5.1339
	CSI	0.5329	0.5286	0.5618	**0.5620**
	POD	0.6180	0.6144	**0.6479**	0.6460
	FAR	0.2054	0.2090	0.1913	**0.1878**
High T	MAE	0.3981	0.4046	**0.3523**	0.3527
	MSE	0.3137	0.3188	0.2532	**0.2531**
	RMSE	0.5600	0.5646	0.5032	**0.5031**
	CSI	0.8247	0.8224	0.8411	**0.8412**
	POD	**0.9094**	0.8887	0.9042	0.9029
	FAR	0.1015	0.0831	0.0767	**0.0750**
Low T	MAE	0.5042	0.5362	0.4124	**0.4023**
	MSE	0.4280	0.4847	0.3141	**0.3001**
	RMSE	0.6542	0.6962	0.5604	**0.5478**
	CSI	0.9329	0.9329	0.9466	**0.9476**
	POD	**0.9880**	0.9806	0.9824	0.9821
	FAR	0.0565	0.0496	0.0370	**0.0358**
Strong wind	MAE	0.8497	0.8441	0.8039	**0.7920**
	MSE	1.5873	1.5358	1.4873	**1.4364**
	RMSE	1.2599	1.2393	1.2195	**1.1985**
	CSI	0.8229	0.8258	0.8342	**0.8360**
	POD	0.9006	0.8979	0.9011	**0.9019**
	FAR	0.0949	0.0886	0.0817	**0.0805**

4.5. Terrain-Related Error Analysis

To further examine spatial error mechanisms, we evaluate model performance over different terrain classes. ERA5 single-level geopotential and land-sea mask fields are used to construct the terrain mask. Geopotential is converted to elevation by dividing by the standard gravitational acceleration, and the land-sea mask is used to separate ocean and land grids. For the 192×256 input domain, grid cells with land-sea-mask values below 0.5 are treated as ocean; land grid cells with elevation below 500 m are treated as plains; and land grid cells with elevation no lower than 500 m are treated as mountains. This produces 18,994 ocean grids, 11,117 plain grids, and 19,041 mountain grids.

Table 5 reports the average error over all forecast lead times and variable-level channels for each terrain class. All methods show larger errors over land than over ocean, and the mountain class has the largest RMSE, indicating that complex topography remains a challenging source of spatial heterogeneity. Nevertheless, VASTFormer+ST-SL achieves the lowest MAE, MSE, and RMSE in all three terrain classes. Compared with the strongest lightweight baseline TAU, VASTFormer+ST-SL reduces RMSE by 2.73%, 4.90%, and 6.87% over ocean, plains, and mountains, respectively. The larger relative gain over mountains suggests that complex terrain is an important source of regional meteorological forecasting error, and that variable-aware representation and selective supervision are particularly useful in spatially heterogeneous regions where atmospheric states are strongly affected by

land-surface and topographic forcing. Compared with the base VASTFormer, ST-SL further reduces RMSE by 0.95%, 0.91%, and 0.66% over the three terrain classes, respectively.

Table 5. Terrain-class error analysis. Rows report average metrics over all forecast lead times and variable-level channels. Lower values indicate better performance. The best results are highlighted in **bold**, and the second-best are underlined.

Terrain	Metric	PredFormer	TAU	VASTFormer	VASTFormer+ST-SL
Ocean	MAE	1.0530	1.0199	0.9785	**0.9708**
	MSE	5.2728	4.9368	4.7612	**4.6708**
	RMSE	2.2963	2.2219	2.1820	**2.1612**
Plain	MAE	1.2530	1.2230	1.1303	**1.1219**
	MSE	7.2412	6.8691	6.3266	**6.2123**
	RMSE	2.6909	2.6209	2.5153	**2.4924**
Mountain	MAE	1.3343	1.3120	1.1769	**1.1705**
	MSE	8.2261	7.8680	6.9158	**6.8247**
	RMSE	2.8681	2.8050	2.6298	**2.6124**

We further quantify the relationship between terrain elevation and spatial prediction error over land grid cells. For each model, spatial MAE and RMSE maps are first accumulated over all test samples, lead times, and channels, and Pearson and Spearman correlations are then computed between elevation and the corresponding error map. Pearson correlation measures the linear association between elevation and error, whereas Spearman rank correlation measures whether higher-elevation grids tend to have larger errors in a monotonic sense. In this analysis, a larger positive correlation is not desirable; lower absolute correlation indicates weaker dependence of prediction error on terrain elevation. As shown in Table 6, elevation has a positive correlation with prediction error for all models, confirming that higher-elevation regions are generally more difficult to forecast in this regional setting. VASTFormer substantially weakens this dependence compared with PredFormer and TAU, reducing the Pearson correlation between elevation and RMSE from 0.3858 for TAU to 0.3180. VASTFormer+ST-SL maintains a clearly lower elevation-error correlation than the lightweight baselines, although its correlation is slightly higher than that of the base VASTFormer. Thus, the terrain analysis suggests that the proposed architecture is the main factor reducing topography-related error sensitivity, while ST-SL further lowers the absolute error within each terrain class.

Table 6. Correlation between elevation and spatial prediction error over land grid cells. Correlations are computed from spatial MAE/RMSE maps accumulated over the 2020 test set. Lower absolute correlation indicates weaker terrain-related error dependence; the values are reported as diagnostic statistics rather than direct accuracy scores.

Method	MAE Pearson	MAE Spearman	RMSE Pearson	RMSE Spearman
PredFormer	0.4425	0.3562	0.4141	0.3257
TAU	0.4284	0.3794	0.3858	0.3271
VASTFormer	0.3498	0.2704	0.3180	0.2506
VASTFormer+ST-SL	0.3631	0.2787	0.3315	0.2597

We also clarify that the present model performs direct multi-step prediction for the future six hours rather than autoregressive rollout prediction. Therefore, the analysis here does not describe error propagation in the classical sense of repeatedly feeding previous predictions back into the model. Instead, we analyze how prediction error grows with forecast lead time under the direct multi-step setting. Table 7 compares the T+2 and T+6 RMSE of VASTFormer and VASTFormer+ST-SL across terrain classes. We use T+2

rather than T+1 as the starting point because the first forecast step is often dominated by short-term persistence and initialization effects, whereas the T+2–T+6 interval better reflects longer-horizon error growth within the evaluated six-hour window. Although all terrain classes show increased error at longer lead times, ST-SL reduces the absolute RMSE growth from T+2 to T+6 by 2.35%, 2.65%, and 2.22% over ocean, plains, and mountains, respectively. Mechanistically, ST-SL can be viewed as a fine-grained unit mining strategy that focuses on difficult but still learnable forecasting targets. Instead of emphasizing all regions with large errors, it selects units whose residuals remain larger than those of the frozen reference model and therefore still contain useful optimization potential. Since longer lead times usually involve more difficult atmospheric transitions and larger underfitting residuals, this selective supervision provides more effective learning signals for later lead times while suppressing noisy or weakly generalizable errors. This indicates that ST-SL helps moderate error growth at longer lead times while preserving the direct multi-step forecasting formulation.

Table 7. Terrain-wise lead-time RMSE growth under direct multi-step forecasting. The increase is computed as T+6 RMSE minus T+2 RMSE. The ST-SL reduction is measured relative to the corresponding increase of the base VASTFormer. Lower RMSE and RMSE Inc. indicate better performance, and the better values within each terrain are highlighted in **bold**. A dash (–) indicates that ST-SL reduction is not applicable to the base VASTFormer.

Terrain	Method	T+2 RMSE	T+6 RMSE	RMSE Inc.	ST-SL Reduction
Ocean	VASTFormer	1.7092	2.9088	1.1996	–
	VASTFormer+ST-SL	**1.7019**	**2.8733**	**1.1715**	2.35%
Plain	VASTFormer	1.9269	3.3727	1.4458	–
	VASTFormer+ST-SL	**1.9225**	**3.3299**	**1.4075**	2.65%
Mountain	VASTFormer	2.0149	3.4963	1.4814	–
	VASTFormer+ST-SL	**2.0130**	**3.4615**	**1.4485**	2.22%

4.6. Lead-Time Forecasting Performance

Meteorological forecasting becomes increasingly challenging as the prediction horizon grows. To further investigate the forecasting behavior across different prediction horizons, we report the lead-time performance from T+1 to T+6 in Table 8. For each forecast horizon, MSE and MAE are averaged over all meteorological variables. This evaluation provides a more detailed view of how different models behave as the forecasting difficulty increases over time.

An interesting observation is that recurrent models remain competitive at the shortest horizon among the baseline models. For example, PredRNN achieves the best baseline T+1 performance, with an MSE of 1.4320 and an MAE of 0.5490. However, VASTFormer already substantially outperforms all baseline models at T+1, and VASTFormer equipped with ST-SL further achieves the best result. This indicates that the proposed framework is effective not only for longer-horizon prediction but also for short-term extrapolation, where the future state is strongly constrained by the most recent observations. Across T+1 to T+6, VASTFormer consistently outperforms all baseline models, and VASTFormer equipped with ST-SL achieves the best results across all horizons and metrics. The non-learning references are substantially less accurate than the trainable models, and the persistence error increases rapidly with lead time, confirming that the task requires learning non-trivial atmospheric evolution beyond simple extrapolation from the latest input state.

Specifically, compared with the strongest baseline at each lead time, VASTFormer+ST-SL reduces MSE by 18.04%, 11.39%, 9.76%, 7.53%, 5.71%, and 4.43% from T+1 to T+6, respectively. The corresponding MAE reductions are 11.86%, 7.44%, 6.85%, 5.15%, 3.80%,

and 2.90%. These improvements demonstrate that the proposed framework provides consistent accuracy gains across different forecasting horizons, from short-term extrapolation to more challenging multi-step meteorological forecasting.

The benefit of ST-SL also becomes clearer at longer lead times. Compared with the base VASTFormer, ST-SL progressively reduces MSE by 0.47%, 1.07%, 1.57%, 1.87%, and 2.25% from T+2 to T+6, respectively. This trend indicates that selective supervision is more beneficial when forecasting uncertainty and spatio-temporal heterogeneity increase over longer horizons. By focusing optimization on high-benefit spatio-temporal units, ST-SL encourages the model to learn more transferable atmospheric evolution patterns rather than uniformly fitting all prediction errors.

Overall, the lead-time results show that VASTFormer provides stronger forecasting capability than existing spatio-temporal predictors across all prediction horizons, while ST-SL further improves the stability and accuracy of multi-step forecasts.

Table 8. Lead-time forecasting performance on the ERA5 dataset from T+1 to T+6. For each forecast horizon, the mean metric value over all meteorological variables is reported. The best results are highlighted in **bold**, and the second-best are underlined. The full results are in Appendix A.

Methods	T+1		T+2		T+3		T+4		T+5		T+6	
	MSE	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
Persistence	5.2487	0.9622	17.5695	1.8467	30.8668	2.4805	44.8417	3.0294	59.4795	3.5515	72.4644	3.9552
Climatology	148.9972	6.7274	148.9972	6.7274	148.9972	6.7274	148.9972	6.7274	148.9972	6.7274	148.9972	6.7274
ConvLSTM	1.4928	0.5776	4.2844	0.9695	6.1526	1.1811	8.1130	1.3692	10.8033	1.5999	12.7974	1.7448
PredRNN	1.4320	0.5490	3.9774	0.9291	5.6112	1.1173	7.2580	1.2779	9.5876	1.4850	11.1855	1.6021
PredRNN++	1.7375	0.6184	4.2802	0.9718	5.9577	1.1574	7.6530	1.3178	10.0381	1.5243	11.6888	1.6433
MIM	1.4810	0.5627	4.2163	0.9624	5.9858	1.1614	7.7203	1.3273	10.0897	1.5337	11.7034	1.6509
MAU	2.5018	0.7604	5.3160	1.1086	7.1867	1.2896	9.0416	1.4475	11.5069	1.6433	13.2372	1.7602
PredRNNv2	1.4991	0.5660	4.0615	0.9404	5.7277	1.1298	7.4137	1.2919	9.7769	1.4984	11.4094	1.6159
SimVP	2.9922	0.8476	5.3822	1.1219	7.1792	1.2954	9.1240	1.4628	11.8302	1.6777	13.9017	1.8174
SimVPv2	2.0014	0.6798	4.2069	0.9748	5.6635	1.1328	7.2079	1.2778	9.5498	1.4815	11.2654	1.6055
TAU	1.9372	0.6638	4.1025	0.9592	5.5021	1.1138	6.9665	1.2537	9.1995	1.4521	10.8189	1.5719
OpenSTL_ViT	1.9531	0.6677	4.1983	0.9711	5.7921	1.1469	7.5385	1.3130	10.1634	1.5402	12.2437	1.6907
OpenSTL_Swin	1.9668	0.6652	4.1203	0.9569	5.5921	1.1190	7.1691	1.2696	9.5594	1.4802	11.3570	1.6130
PredFormer	1.9913	0.6685	4.3641	0.9845	5.9251	1.1543	7.5254	1.3053	9.8203	1.5028	11.5426	1.6275
VASTFormer (Ours)	1.1737	0.4859	3.5408	0.8630	5.0192	1.0431	6.5452	1.1975	8.8399	1.4073	10.5770	1.5411
+ ST-SL (Ours)	**1.1736**	**0.4839**	**3.5242**	**0.8600**	**4.9653**	**1.0375**	**6.4422**	**1.1891**	**8.6746**	**1.3969**	**10.3392**	**1.5263**

4.7. Variable-Wise Analysis

Different meteorological variables exhibit distinct physical characteristics and prediction difficulties. To further examine the forecasting behavior across different meteorological variables, we report the variable-wise performance in Table 9. For each variable, the results are averaged over the corresponding pressure levels, forecast horizons, and spatial locations. This analysis helps reveal whether the proposed framework improves forecasting uniformly across heterogeneous atmospheric variables, rather than only benefiting a specific variable. Persistence and climatology are included as non-learning references to contextualize the difficulty of each variable group.

Overall, VASTFormer achieves the second-best performance across all variables and metrics, while VASTFormer+ST-SL achieves the best performance across all variables and metrics. Compared with the strongest baseline for each variable, the proposed framework brings clear improvements for relative humidity, temperature, and wind components. For relative humidity, VASTFormer+ST-SL reduces MSE from 24.0731 to 21.9518 and MAE

from 3.2066 to 2.9818, corresponding to relative improvements of 8.81% and 7.01%, respectively. For wind forecasting, VASTFormer+ST-SL also achieves consistent gains: it reduces U-wind MSE and MAE by 8.09% and 6.03%, and V-wind MSE and MAE by 6.24% and 4.99% compared with the best baseline results. These improvements suggest that variable-aware fusion is beneficial for modeling physically coupled meteorological variables, especially wind fields whose evolution depends strongly on surrounding atmospheric structures.

Table 9. Variable-wise forecasting performance on the ERA5 dataset. Metrics are averaged over pressure levels and lead times. Lower values indicate better performance. The best results are highlighted in **bold**, and the second-best are underlined. The full results are in Appendix B.

Methods	Hum.		Tem.		U-Wind		V-Wind	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Persistence	144.7158	7.3945	1.4160	0.7153	3.5571	1.2004	3.9581	1.2402
Climatology	528.8999	17.3913	11.8175	2.2867	31.5176	3.9036	23.7679	3.3283
ConvLSTM	27.2083	3.3953	0.2609	0.3525	0.8333	0.6147	0.7930	0.5988
PredRNN	24.4333	3.2066	0.2137	0.3195	0.7087	0.5636	0.6788	0.5506
PredRNN++	25.8664	3.3303	0.2333	0.3349	0.7540	0.5863	0.7165	0.5705
MIM	25.7686	3.3170	0.2370	0.3364	0.7500	0.5812	0.7088	0.5643
MAU	30.5289	3.6992	0.2816	0.3779	0.8921	0.6445	0.8241	0.6181
PredRNNv2	24.9463	3.2411	0.2189	0.3232	0.7305	0.5726	0.6964	0.5581
SimVP	31.4431	3.7582	0.3211	0.4038	0.9624	0.6752	0.8796	0.6447
SimVPv2	24.9147	3.2671	0.2480	0.3544	0.7340	0.5816	0.6997	0.5650
TAU	24.0731	3.2111	0.2311	0.3413	0.7071	0.5700	0.6732	0.5540
OpenSTL_ViT	26.0824	3.3304	0.2454	0.3500	0.8234	0.6129	0.7747	0.5932
OpenSTL_Swin	24.8190	3.2486	0.2233	0.3333	0.7544	0.5859	0.7132	0.5682
PredFormer	25.7109	3.3150	0.2361	0.3434	0.7708	0.5935	0.7280	0.5767
VASTFormer (Ours)	22.3046	2.9991	0.1823	0.2969	0.6627	0.5339	0.6476	0.5287
+ ST-SL (Ours)	**21.9518**	**2.9818**	**0.1798**	**0.2946**	**0.6499**	**0.5296**	**0.6312**	**0.5231**

For temperature forecasting, VASTFormer+ST-SL achieves the best performance, with an MSE of 0.1798 and an MAE of 0.2946. Compared with the strongest baseline, VASTFormer+ST-SL reduces temperature MSE and MAE by 15.86% and 7.79%, respectively. This indicates that the proposed framework is also effective for temperature fields, which usually exhibit stronger regularity and smoother spatio-temporal evolution. The improvement suggests that combining a physics-enhanced backbone with selective supervision can further refine the prediction of regular atmospheric variables, rather than only benefiting highly heterogeneous fields.

The effect of ST-SL is more evident for variables with stronger spatial heterogeneity. Compared with VASTFormer, ST-SL further reduces the MSE of relative humidity, temperature, U-wind, and V-wind by 1.58%, 1.37%, 1.93%, and 2.53%, respectively. The larger gains on relative humidity, U-wind and V-wind suggest that selective supervision is especially beneficial for variables with more complex spatial dynamics. This supports the motivation of selective learning: heterogeneous variables often contain regions with different learning benefits, and selectively optimizing high-benefit spatio-temporal units can improve robustness across variables. Overall, the variable-wise results demonstrate that VASTFormer provides a strong multi-variable forecasting backbone, while ST-SL further improves all variables without introducing inference-time overhead.

4.8. Ablation Study

In this subsection, we conduct ablation studies to evaluate the contribution of each component in the proposed framework. The results are reported in Table 10. All variants are evaluated under the same ERA5 forecasting setting, and the parameter counts and FLOPs are also provided to distinguish architectural gains from simple model-size effects.

Table 10. Ablation study of each key component in the proposed framework. Lower metric values indicate better performance. Best results are highlighted in **bold**.

Variant	Params (M)	FLOPs (T)	MSE ↓	MAE ↓	RMSE ↓
VASTFormer+ST-SL	17.364	0.102	**5.8532**	**1.0823**	**2.4156**
VASTFormer	17.364	0.102	5.9493	1.0896	2.4351
w/o Variable-Aware Fusion	17.293	0.102	6.0029	1.0984	2.4461
Solar PE → Learnable PE	17.356	0.102	6.0527	1.1011	2.4563
w/o Skip Connection	17.331	0.098	6.1140	1.1248	2.4688

Contribution of ST-SL. Compared with the base VASTFormer trained using the standard full-field regression objective, VASTFormer+ST-SL reduces MSE, MAE, and RMSE by 1.62%, 0.67%, and 0.80%, respectively. Since both models share the same inference-time parameter count and FLOPs, this improvement comes from the selective training strategy rather than additional inference complexity. This result supports the benefit of optimizing high-benefit spatio-temporal units instead of uniformly fitting all prediction errors.

Role of Variable-Aware Fusion. Removing the lightweight Variable-Aware Fusion module leads to consistent degradation, increasing MSE, MAE, and RMSE by 0.90%, 0.81%, and 0.45%, respectively, compared with the complete VASTFormer. This indicates that simply treating meteorological variables as homogeneous channels is suboptimal. The variable-aware fusion module provides useful cross-variable interaction modeling, which is important for atmospheric fields where humidity, temperature, and wind components are physically coupled.

Impact of Solar PE. The replacement of Solar Positional Encoding with a standard learnable positional encoding [17] also significantly degrades performance, with MSE, MAE, and RMSE increasing by 1.74%, 1.06%, and 0.87%, respectively. This confirms that the solar-geometry positional prior provides meaningful information for meteorological forecasting. By encoding geographically dependent diurnal and seasonal patterns, Solar PE helps the model better capture physically relevant spatio-temporal variations.

Benefit of Skip Connection. Removing the decoder skip connection leads to the largest degradation among the architectural ablations, increasing MSE, MAE, and RMSE by 2.77%, 3.23%, and 1.38%, respectively. This suggests that shallow high-resolution spatial features remain useful for reconstructing fine-grained meteorological fields. The skip connection helps preserve local spatial details that may be weakened during patch embedding and latent transformation.

Overall, the ablation results demonstrate that each component contributes positively to the final performance. The improvements are not caused by larger model capacity, since all variants have similar parameter counts and FLOPs. Instead, the results indicate that VASTFormer benefits from variable-aware interaction modeling, physics-enhanced positional priors, and spatial detail preservation, while ST-SL further improves optimization through online selective supervision.

4.9. Effect of Selective Ratio

The selective ratio ρ controls the proportion of spatio-temporal units retained for loss computation. To verify the "less is more" hypothesis, we evaluate different fixed selection ratios and report the results in Table 11. Here, $\rho = 1.0$ corresponds to the base VASTFormer trained with full-field supervision, where all spatio-temporal units are used for loss computation. As ρ decreases from 0.9 to 0.6, the forecasting performance generally improves, indicating that removing a moderate proportion of low-benefit supervision signals can improve optimization. The best performance is achieved at $\rho = 0.6$, where MSE, MAE, and RMSE are reduced by 1.62%, 0.67%, and 0.80%, respectively, compared with full-field supervision.

However, a small selection ratio discards too many useful units and leads to performance degradation. For example, when ρ decreases from 0.6 to 0.3, MSE increases from 5.8532 to 5.9848, corresponding to relative degradations of 2.25%. This suggests that overly aggressive selection discards too many useful supervision signals, weakening the model's ability to learn diverse atmospheric patterns. Therefore, selective learning requires a balance between supervision quality and supervision coverage: retaining a properly selected subset of high-benefit spatio-temporal units is more effective than either using all units or discarding too many units.

Table 11. Effect of different selective ratios. Lower metric values indicate better performance. Best results are highlighted in **bold**.

Selective Ratio	MSE ↓	MAE ↓	RMSE ↓
1.0	5.9493	1.0896	2.4351
0.9	5.9785	1.0928	2.4411
0.8	5.9342	1.0890	2.4321
0.7	5.8700	1.0831	2.4189
0.6	**5.8532**	**1.0823**	**2.4156**
0.5	5.8842	1.0850	2.4220
0.4	5.9402	1.0905	2.4335
0.3	5.9848	1.0947	2.4426

4.10. Computational Efficiency

Finally, we analyze the computational efficiency of the proposed ST-SL framework. Table 12 reports the number of parameters, training time, and inference cost. Compared with standard VASTFormer training, ST-SL introduces an additional forward pass through the frozen reference model and mask construction during training. In our implementation, this increases the training time by approximately 1.3×. However, after training, the reference model and mask generation branch are discarded. Therefore, VASTFormer+ST-SL has exactly the same inference architecture, parameter count, and latency as VASTFormer.

Table 12. Computational efficiency comparison. ST-SL increases training cost but introduces no additional inference overhead.

Method	Params	Training Time	Inference Time
VASTFormer	17.36 M	1.0×	1.0×
VASTFormer+ST-SL	17.36 M	1.3×	1.0 × ϕ

These results demonstrate that ST-SL is a practical training framework: it improves forecasting accuracy and robustness while preserving the deployment efficiency of the original backbone.

4.11. Visualization

Prediction visualization. To provide an intuitive understanding of the forecasting behavior, we visualize representative prediction results of the proposed model in Figure 4. The figure compares the predicted fields with the corresponding ground-truth fields from T+1 to T+6 for four meteorological variables, including relative humidity, temperature, U-wind, and V-wind. Overall, the proposed model accurately preserves the large-scale spatial structures and temporal evolution patterns across different variables. For relatively smooth variables such as temperature, the predicted fields remain highly consistent with the targets over the entire forecasting horizon. For relative humidity and wind components, which exhibit stronger spatial heterogeneity, the model also captures the major atmospheric patterns and their gradual evolution from short-term to longer-term forecasts.

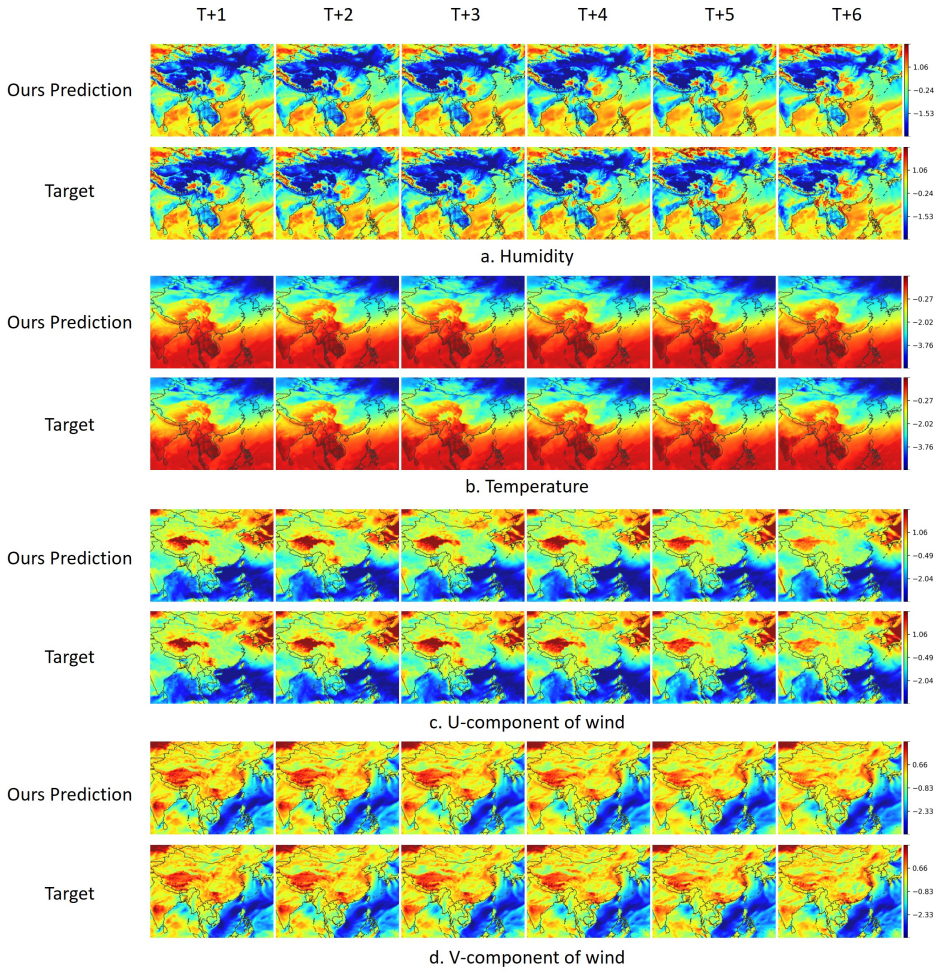


Figure 4. Qualitative prediction results of the proposed model on the ERA5 dataset. Four meteorological variables are visualized from T+1 to T+6, including (**a**) relative humidity, (**b**) temperature, (**c**) U-component of wind, and (**d**) V-component of wind. For each variable, the predicted fields are compared with the corresponding ground-truth fields.

Error comparison. To further compare spatial error distributions across different models, we visualize spatial MSE maps for different meteorological variables. Temperature, shown in Figure 5, is selected as the representative variable in the main text because it clearly reflects the model's ability to preserve smooth large-scale atmospheric structures while controlling error accumulation over lead times. Compared with TAU [11], PredFormer [13], and PredRNN [9], the proposed model exhibits consistently lower spatial errors from T+1 to T+6. The advantage remains visible at longer horizons, where baseline models show more pronounced error accumulation over complex land regions. This qualitative comparison is

consistent with the quantitative lead-time and variable-wise results, confirming that the proposed framework improves both numerical accuracy and spatial forecast quality.

Additional MSE-map comparisons for other meteorological variables, including relative humidity in Figure A1, the U-component of wind in Figure A2, and the V-component of wind in Figure A3, are provided in Appendix C. These supplementary visualizations show similar trends across heterogeneous meteorological variables and further support the robustness of the proposed framework.

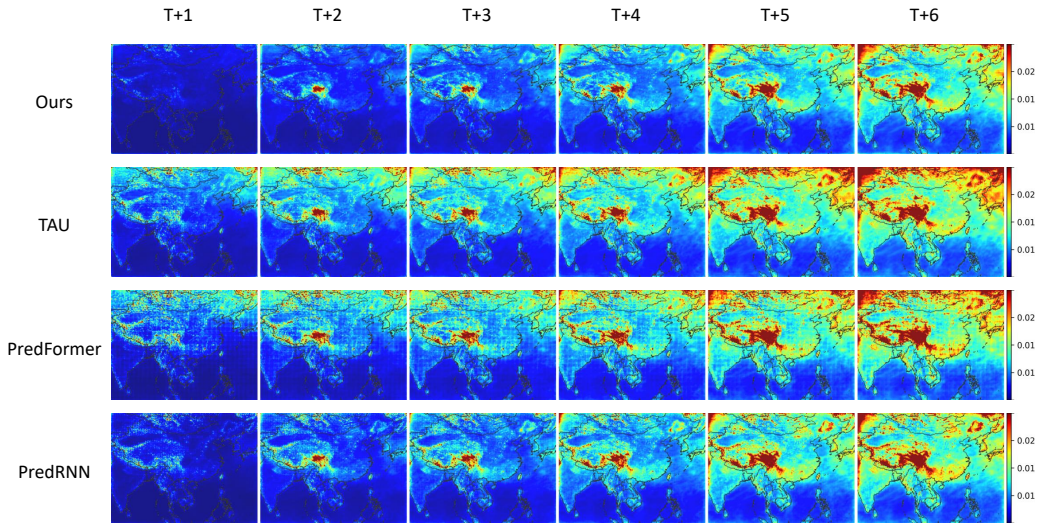


Figure 5. MSE-map comparison of different models for temperature from T+1 to T+6. Warmer colors indicate larger prediction errors, and cooler colors indicate smaller errors. The proposed model shows consistently lower spatial errors than representative baselines across all forecasting horizons.

4.12. Summary of Results

Overall, the experimental results demonstrate the effectiveness of both components of the proposed framework. VASTFormer provides a strong variable-aware and physics-enhanced spatio-temporal backbone for meteorological forecasting. ST-SL further improves generalization by selectively optimizing high-benefit spatio-temporal units instead of uniformly fitting all prediction errors. The consistent improvements across overall metrics, lead times, variables, ablations, and qualitative cases support the central claim of this work: in dense meteorological forecasting, less but more informative supervision can lead to better generalization.

5. Discussion

5.1. Interpretation and Implications

The results of this study suggest that improving meteorological forecasting does not necessarily require uniformly increasing model capacity. Instead, better performance can be obtained by jointly improving how atmospheric fields are represented and how dense supervision signals are used during training. VASTFormer addresses the representation problem by explicitly considering the structured nature of meteorological data, including variable-specific semantics, cross-variable interactions, physically meaningful spatio-temporal priors, and high-resolution spatial reconstruction. Ablation studies validate the effectiveness of each component. These components provide useful inductive biases for regional meteorological forecasting, where different variables and pressure levels are physically coupled but exhibit distinct evolution patterns.

The proposed ST-SL framework further highlights the importance of supervision quality in dense forecasting tasks. Conventional regression training assigns equal weight

to all prediction units, although meteorological fields often contain regions with different levels of predictability, uncertainty, and learning value. This observation is consistent with prior studies on curriculum learning, data pruning, and online data selection, which suggest that not all training signals contribute equally to generalization [44,46,47,56]. By performing online selection at the spatio-temporal-unit level, ST-SL extends the idea of data selection from sample-level learning to dense atmospheric prediction. This distinction is important because informative and low-benefit regions may coexist within the same meteorological field. Rather than discarding entire samples, ST-SL selectively emphasizes units with higher expected learning benefit, allowing the model to focus on more useful supervision signals while preserving the dense forecasting formulation.

The selective ratio analysis also clarifies the meaning of the "less is more" hypothesis. The goal is not to use as little supervision as possible, but to balance supervision quality and supervision coverage. Moderate selection can reduce the influence of noisy or weakly beneficial units, whereas overly aggressive selection may remove useful atmospheric information and weaken generalization. This observation suggests that selective supervision should be viewed as a training-time regularization strategy rather than a simple data reduction technique.

The added ACC and RMSESS evaluation further indicates that the improvement is not limited to scale-dependent regression errors. Because ACC measures anomaly-pattern agreement and RMSESS compares the model with a climatological reference, the gains suggest that VASTFormer+ST-SL better preserves dynamically meaningful departures from the seasonal background state. This is important for meteorological applications, where accurate anomaly structure is often as relevant as the absolute magnitude of pointwise errors. The percentile-based extreme-condition evaluation further examines whether these improvements persist under extreme conditions. Although the present task does not include precipitation, tropical cyclone tracks, or event labels, the available ERA5 variables allow us to examine high-humidity, high-temperature, low-temperature, and strong-wind regimes. The proposed framework improves both intensity-oriented metrics and categorical event-detection scores on these proxy extremes, indicating that selective supervision does not merely improve easy background states but also benefits meteorologically important tail conditions. From a spatial perspective, the terrain-related analysis reveals that prediction errors increase from ocean to plains and mountains, consistent with the stronger spatial heterogeneity induced by land surface and topographic effects. The proposed VASTFormer backbone reduces the correlation between elevation and spatial error, suggesting that the variable-aware and physics-enhanced representation weakens topography-related error dependence. ST-SL then further lowers the absolute errors within each terrain class, which is consistent with its role as a training-time mechanism for improving supervision quality rather than explicitly encoding topography.

Another practical implication is that ST-SL improves training without increasing deployment cost. The frozen reference model and mask generation process are only used during optimization and are discarded during inference. Therefore, the final forecasting model retains the same parameter count, FLOPs, and latency as the original VASTFormer backbone. This property is valuable for operational meteorological applications, where inference efficiency is often as important as forecasting accuracy.

5.2. Limitations and Future Work

Despite these advantages, several limitations remain. First, this study focuses on a regional ERA5 forecasting setting, which enables consistent comparison across lightweight spatio-temporal baselines. Further validation on additional geographic regions or meteorological datasets would help assess the broader applicability of the proposed frame-

work. Second, the current framework is formulated as a deterministic forecasting method with regression losses and therefore does not directly provide calibrated uncertainty estimates [57]. Such uncertainty information is important for risk-sensitive meteorological decision-making, especially under high-impact weather conditions. Finally, the physics-enhanced prior used in VASTFormer is based on solar altitude. Incorporating additional physical information, such as topography, land-sea distribution, or conservation constraints, may further strengthen the physical grounding of the model [58].

Future research can proceed in several directions. One promising direction is to apply ST-SL to broader meteorological settings, including longer forecast lead times, global forecasting, precipitation nowcasting, and multi-source forecasting with satellite, radar, station, and reanalysis data. Another direction is to explore adaptive or variable-specific selection criteria that account for different predictability characteristics across meteorological variables and pressure levels. It would also be valuable to combine selective learning with probabilistic prediction and physical constraint modeling, enabling models not only to forecast accurately but also to quantify uncertainty and maintain physical consistency.

6. Conclusions

In this paper, we present VASTFormer+ST-SL, a framework for multi-variable meteorological forecasting that combines efficient spatio-temporal representation learning with online selective supervision. VASTFormer serves as a variable-aware spatio-temporal Transformer backbone, incorporating multi-variable fusion, physics-enhanced Solar Positional Encoding, and trajectory-aware temporal translation to capture structured atmospheric dynamics. Built upon this backbone, ST-SL introduces an online selective learning strategy that constructs learnability-aware scores with a frozen reference model and optimizes only high-benefit spatio-temporal units. Extensive experiments on regional ERA5 forecasting demonstrate that the proposed framework outperforms representative lightweight spatio-temporal baselines, improves ACC and RMSESS relative to climatology, and strengthens prediction under percentile-defined extreme conditions, while adding no inference-time parameters or computational cost. These results support the central claim of this work: in dense meteorological forecasting, less but more informative supervision can lead to better generalization. The proposed framework offers a practical and efficient direction for improving data-driven meteorological forecasting, and it opens up future opportunities for selective learning in broader spatio-temporal scientific prediction tasks.

Author Contributions: Conceptualization, P.Z., D.X. and K.D.; methodology, P.Z.; software, P.Z.; validation, P.Z. and S.X.; formal analysis, P.Z.; investigation, P.Z. and D.X.; resources, K.D., C.H. and S.X.; data curation, P.Z.; writing—original draft preparation, P.Z.; writing—review and editing, D.X., K.D., C.H. and S.X.; visualization, P.Z.; supervision, K.D., C.H. and S.X.; project administration, S.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant No. 62071466).

Data Availability Statement: The data that support the findings of this study are openly available. The ERA5 reanalysis dataset was downloaded from the Copernicus Climate Data Store at https://cds.climate.copernicus.eu/ (accessed on 15 May 2026), and its DOI is: 10.24381/cds.bd0915c6. No new data were generated in this study.

Conflicts of Interest: Author Deping Xiang was employed by the company JD.com, Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Full Lead-Time Forecasting Results

The full results of lead-time forecasting are summarized in Table A1. The best results are obtained for the overall performance of our proposed framework.

Table A1. Lead-time forecasting performance on the ERA5 dataset from T+1 to T+6. For each forecast horizon, the mean metric value over all meteorological variables is reported. Lower metric values indicate better performance. The best results are highlighted in **bold**, and the second-best are underlined.

Metric	Method	T+1	T+2	T+3	T+4	T+5	T+6
MSE ↓	Persistence	5.2487	17.5695	30.8668	44.8417	59.4795	72.4644
	Climatology	148.9972	148.9972	148.9972	148.9972	148.9972	148.9972
	ConvLSTM	1.4928	4.2844	6.1526	8.1130	10.8033	12.7974
	PredRNN	1.4320	3.9774	5.6112	7.2580	9.5876	11.1855
	PredRNN++	1.7375	4.2802	5.9577	7.6530	10.0381	11.6888
	MIM	1.4810	4.2163	5.9858	7.7203	10.0897	11.7034
	MAU	2.5018	5.3160	7.1867	9.0416	11.5069	13.2372
	PredRNNv2	1.4991	4.0615	5.7277	7.4137	9.7769	11.4094
	SimVP	2.9922	5.3822	7.1792	9.1240	11.8302	13.9017
	SimVPv2	2.0014	4.2069	5.6635	7.2079	9.5498	11.2654
	TAU	1.9372	4.1025	5.5021	6.9665	9.1995	10.8189
	OpenSTL_ViT	1.9531	4.1983	5.7921	7.5385	10.1634	12.2437
	OpenSTL_Swin	1.9668	4.1203	5.5921	7.1691	9.5594	11.3570
	PredFormer	1.9913	4.3641	5.9251	7.5254	9.8203	11.5426
	VASTFormer (Ours)	1.1737	3.5408	5.0192	6.5452	8.8399	10.5770
	VASTFormer+ST-SL (Ours)	**1.1736**	**3.5242**	**4.9653**	**6.4422**	**8.6746**	**10.3392**
MAE ↓	Persistence	0.9622	1.8467	2.4805	3.0294	3.5515	3.9552
	Climatology	6.7274	6.7274	6.7274	6.7274	6.7274	6.7274
	ConvLSTM	0.5776	0.9695	1.1811	1.3692	1.5999	1.7448
	PredRNN	0.5490	0.9291	1.1173	1.2779	1.4850	1.6021
	PredRNN++	0.6184	0.9718	1.1574	1.3178	1.5243	1.6433
	MIM	0.5627	0.9624	1.1614	1.3273	1.5337	1.6509
	MAU	0.7604	1.1086	1.2896	1.4475	1.6433	1.7602
	PredRNNv2	0.5660	0.9404	1.1298	1.2919	1.4984	1.6159
	SimVP	0.8476	1.1219	1.2954	1.4628	1.6777	1.8174
	SimVPv2	0.6798	0.9748	1.1328	1.2778	1.4815	1.6055
	TAU	0.6638	0.9592	1.1138	1.2537	1.4521	1.5719
	OpenSTL_ViT	0.6677	0.9711	1.1469	1.3130	1.5402	1.6907
	OpenSTL_Swin	0.6652	0.9569	1.1190	1.2696	1.4802	1.6130
	PredFormer	0.6685	0.9845	1.1543	1.3053	1.5028	1.6275
	VASTFormer (Ours)	0.4859	0.8630	1.0431	1.1975	1.4073	1.5411
	VASTFormer+ST-SL (Ours)	**0.4839**	**0.8600**	**1.0375**	**1.1891**	**1.3969**	**1.5263**
RMSE ↓	Persistence	2.2910	4.1916	5.5558	6.6964	7.7123	8.5126
	Climatology	12.2064	12.2064	12.2064	12.2064	12.2064	12.2064
	ConvLSTM	1.2195	2.0666	2.4768	2.8438	3.2817	3.5709
	PredRNN	1.1942	1.9910	2.3654	2.6900	3.0920	3.3391
	PredRNN++	1.3161	2.0659	2.4375	2.7623	3.1638	3.4133
	MIM	1.2126	2.0491	2.4423	2.7735	3.1709	3.4143
	MAU	1.5798	2.3029	2.6772	3.0021	3.3868	3.6316
	PredRNNv2	1.2219	2.0119	2.3896	2.7184	3.1220	3.3718
	SimVP	1.7278	2.3174	2.6760	3.0161	3.4345	3.7221
	SimVPv2	1.4130	2.0484	2.3767	2.6808	3.0858	3.3507
	TAU	1.3902	2.0227	2.3425	2.6355	3.0288	3.2839
	OpenSTL_ViT	1.3957	2.0460	2.4032	2.7411	3.1828	3.4924
	OpenSTL_Swin	1.4006	2.0268	2.3613	2.6732	3.0869	3.3639
	PredFormer	1.4092	2.0858	2.4307	2.7391	3.1291	3.3916
	VASTFormer (Ours)	1.0801	1.8777	2.2366	2.5539	2.9644	3.2459
	VASTFormer+ST-SL (Ours)	**1.0799**	**1.8734**	**2.2247**	**2.5340**	**2.9406**	**3.2097**

Appendix B. Full Variable-Wise Forecasting Results

The full results of variable-wise forecasting are summarized in Table A2. The best results are obtained for the overall performance of our proposed framework.

Table A2. Variable-wise forecasting performance on the ERA5 dataset. Metrics are averaged over pressure levels and lead times. Lower metric values indicate better performance. The best results are highlighted in **bold**, and the second-best are underlined.

Metric	Method	Hum.	Tem.	U-Wind	V-Wind
MSE ↓	Persistence	144.7158	1.4160	3.5571	3.9581
	Climatology	528.8999	11.8175	31.5176	23.7679
	ConvLSTM	27.2083	0.2609	0.8333	0.7930
	PredRNN	24.4333	0.2137	0.7087	0.6788
	PredRNN++	25.8664	0.2333	0.7540	0.7165
	MIM	25.7686	0.2370	0.7500	0.7088
	MAU	30.5289	0.2816	0.8921	0.8241
	PredRNNv2	24.9463	0.2189	0.7305	0.6964
	SimVP	31.4431	0.3211	0.9624	0.8796
	SimVPv2	24.9147	0.2480	0.7340	0.6997
	TAU	24.0731	0.2311	0.7071	0.6732
	OpenSTL_ViT	26.0824	0.2454	0.8234	0.7747
	OpenSTL_Swin	24.8190	0.2233	0.7544	0.7132
	PredFormer	25.7109	0.2361	0.7708	0.7280
	VASTFormer (Ours)	22.3046	0.1823	0.6627	0.6476
	VASTFormer+ST-SL (Ours)	**21.9518**	**0.1798**	**0.6499**	**0.6312**
MAE ↓	Persistence	7.3945	0.7153	1.2004	1.2402
	Climatology	17.3913	2.2867	3.9036	3.3283
	ConvLSTM	3.3953	0.3525	0.6147	0.5988
	PredRNN	3.2066	0.3195	0.5636	0.5506
	PredRNN++	3.3303	0.3349	0.5863	0.5705
	MIM	3.3170	0.3364	0.5812	0.5643
	MAU	3.6992	0.3779	0.6445	0.6181
	PredRNNv2	3.2411	0.3232	0.5726	0.5581
	SimVP	3.7582	0.4038	0.6752	0.6447
	SimVPv2	3.2671	0.3544	0.5816	0.5650
	TAU	3.2111	0.3413	0.5700	0.5540
	OpenSTL_ViT	3.3304	0.3500	0.6129	0.5932
	OpenSTL_Swin	3.2486	0.3333	0.5859	0.5682
	PredFormer	3.3150	0.3434	0.5935	0.5767
	VASTFormer (Ours)	2.9991	0.2969	0.5339	0.5287
	VASTFormer+ST-SL (Ours)	**2.9818**	**0.2946**	**0.5296**	**0.5231**
RMSE ↓	Persistence	12.0298	1.1900	1.8860	1.9895
	Climatology	22.9978	3.4377	5.6141	4.8752
	ConvLSTM	5.2083	0.5095	0.9104	0.8879
	PredRNN	4.9359	0.4613	0.8395	0.8215
	PredRNN++	5.0787	0.4818	0.8661	0.8440
	MIM	5.0675	0.4852	0.8633	0.8390
	MAU	5.5168	0.5291	0.9423	0.9053
	PredRNNv2	4.9869	0.4667	0.8521	0.8317
	SimVP	5.5995	0.5647	0.9790	0.9355
	SimVPv2	4.9845	0.4946	0.8548	0.8343
	TAU	4.8996	0.4781	0.8388	0.8182
	OpenSTL_ViT	5.0991	0.4932	0.9049	0.8774
	OpenSTL_Swin	4.9742	0.4711	0.8660	0.8417
	PredFormer	5.0632	0.4843	0.8755	0.8506
	VASTFormer (Ours)	4.7149	0.4254	0.8111	0.8017
	VASTFormer+ST-SL (Ours)	**4.6779**	**0.4226**	**0.8033**	**0.7916**

Appendix C. Full Visualization Results

This section presents the full visualization results of the spatial distribution of MSE loss for the predictions of different models across meteorological variables, including relative humidity in Figure A1, the U-component of wind in Figure A2, and the V-component of wind in Figure A3. The experiments are conducted on the ERA5 dataset. These results show consistent trends across meteorological variables and further support the robustness of the proposed framework.

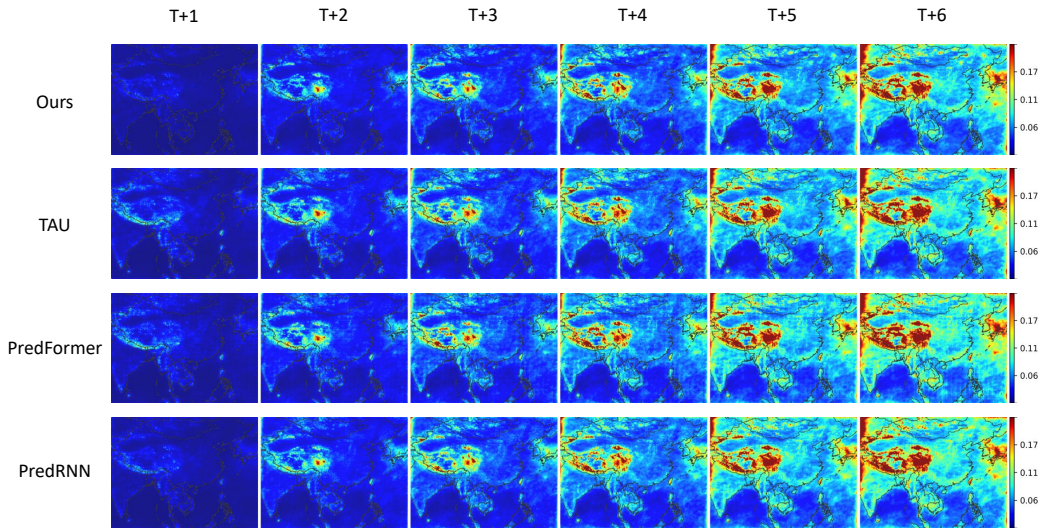


Figure A1. MSE-map comparison of different models for relative humidity from T+1 to T+6. Warmer colors indicate larger prediction errors, and cooler colors indicate smaller errors. Relative humidity exhibits strong spatial heterogeneity, and the proposed model produces lower and more stable error distributions than representative baselines across all forecasting horizons.

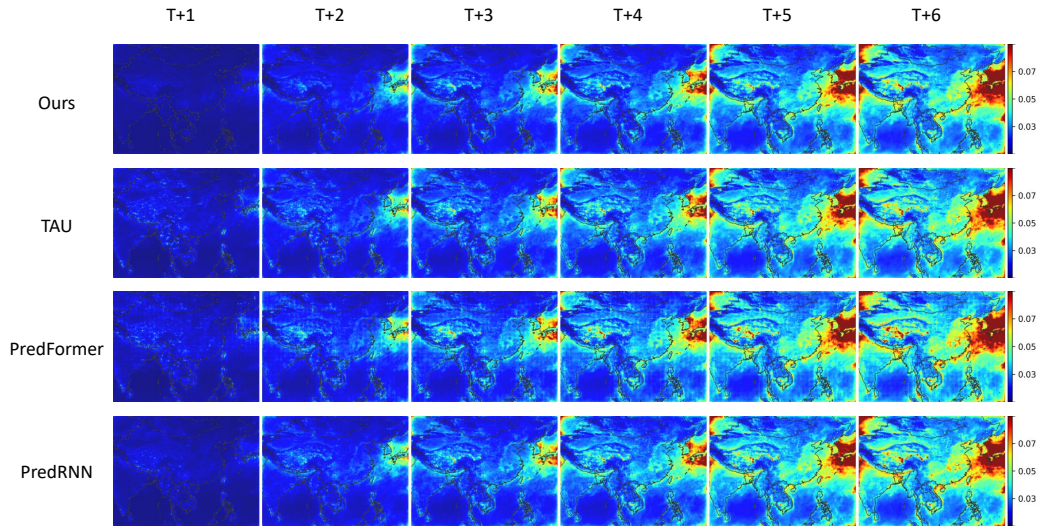


Figure A2. MSE-map comparison of different models for the U-component of wind from T+1 to T+6. Warmer colors indicate larger prediction errors, and cooler colors indicate smaller errors. The proposed model shows lower and more stable spatial error distributions than representative baselines across all forecasting horizons.

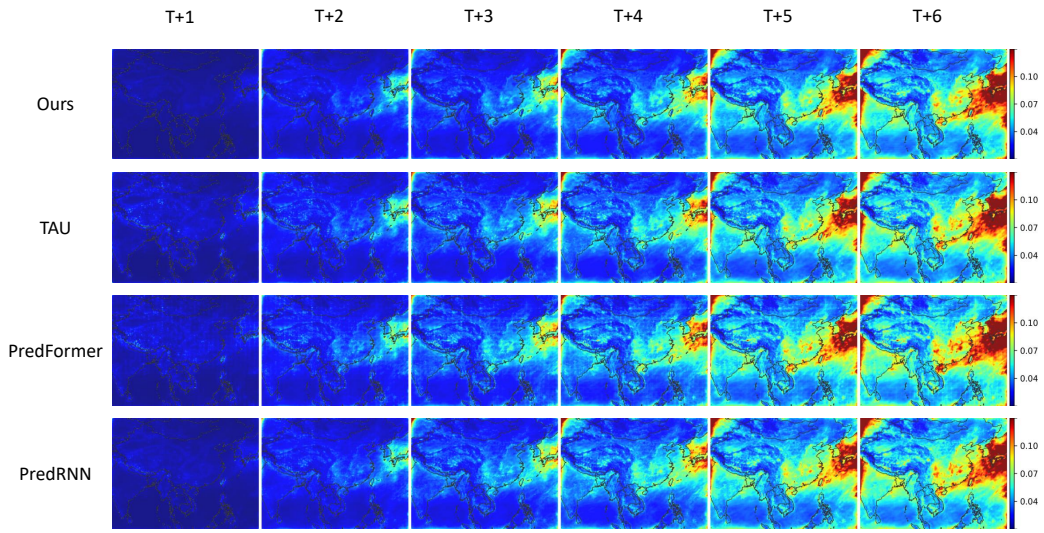


Figure A3. MSE-map comparison of different models for the V-component of wind from T+1 to T+6. Warmer colors indicate larger prediction errors, and cooler colors indicate smaller errors. The proposed model shows lower and more stable spatial error distributions than representative baselines across all forecasting horizons.

References

1. Alley, R.B.; Emanuel, K.A.; Zhang, F. Advances in weather prediction. *Science* **2019**, *363*, 342–344. [CrossRef] [PubMed]
2. Potter, S.; Harrison, S.; Kreft, P. The benefits and challenges of implementing impact-based severe weather warning systems: Perspectives of weather, flood, and emergency management personnel. *Weather Clim. Soc.* **2021**, *13*, 303–314. [CrossRef]
3. Born, L.; Prager, S.; Ramirez-Villegas, J.; Imbach, P. A global meta-analysis of climate services and decision-making in agriculture. *Clim. Serv.* **2021**, *22*, 100231. [CrossRef]
4. Sweeney, C.; Bessa, R.J.; Browell, J.; Pinson, P. The future of forecasting for renewable energy. *Wiley Interdiscip. Rev. Energy Environ.* **2020**, *9*, e365.
5. Peng, Y.; Jiang, Y.; Lu, J.; Zou, Y. Examining the effect of adverse weather on road transportation using weather and traffic sensors. *PLoS ONE* **2018**, *13*, e0205409. [CrossRef] [PubMed]
6. Bauer, P.; Thorpe, A.; Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **2015**, *525*, 47–55. [CrossRef] [PubMed]
7. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat, F. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef] [PubMed]
8. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28, pp. 802–810.
9. Wang, Y.; Long, M.; Wang, J.; Gao, Z.; Yu, P.S. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 879–888.
10. Gao, Z.; Tan, C.; Wu, L.; Li, S.Z. SimVP: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2022; pp. 3170–3180.
11. Tan, C.; Gao, Z.; Wu, L.; Xu, Y.; Xia, J.; Li, S.; Li, S.Z. Temporal Attention Unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 18770–18782.
12. Gao, Z.; Shi, X.; Wang, H.; Zhu, Y.; Wang, Y.B.; Li, M.; Yeung, D.Y. EarthFormer: Exploring space-time transformers for earth system forecasting. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 25390–25403.
13. Tang, Y.; Qi, L.; Xie, F.; Li, X.; Ma, C.; Yang, M.H. PredFormer: Transformers Are Effective Spatial-Temporal Predictive Learners. *arXiv* **2024**, arXiv:2410.04733.
14. Bertasius, G.; Wang, H.; Torresani, L. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the International Conference on Machine Learning*; Proceedings of Machine Learning Research; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 813–824.

15. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A Video Vision Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 6836–6846.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5998–6008.
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*; OpenReview.net: Amherst, MA, USA, 2021.
18. Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. Learning skillful medium-range global weather forecasting. *Science* **2023**, *382*, 1416–1421. [CrossRef] [PubMed]
19. Nguyen, T.; Brandstetter, J.; Kapoor, A.; Gupta, J.K.; Grover, A. ClimaX: A foundation model for weather and climate. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 25904–25938.
20. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [CrossRef]
21. Dee, D.P.; Uppala, S.M.; Simmons, A.J.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M.; Balsamo, G.; Bauer, D.P.; et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **2011**, *137*, 553–597. [CrossRef]
22. Berner, J.; Achatz, U.; Batte, L.; Bengtsson, L.; Cámara, A.d.l.; Christensen, H.M.; Colangeli, M.; Coleman, D.R.; Crommelin, D.; Dolaptchiev, S.I.; et al. Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 565–588. [CrossRef]
23. Lorenz, E.N. Deterministic nonperiodic flow. In *Universality in Chaos*, 2nd ed.; Routledge: Abingdon, UK, 2017; pp. 367–378.
24. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 568–578.
25. Shazeer, N. Glu variants improve transformer. *arXiv* **2020**, arXiv:2002.05202.
26. Shi, X.; Gao, Z.; Lausen, L.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Deep Learning for Precipitation Nowcasting: A Benchmark and a New Model. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5617–5627.
27. Wang, Y.; Gao, Z.; Long, M.; Wang, J.; Yu, P.S. PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2018; Volume 80, pp. 5123–5132.
28. Wang, Y.; Zhang, J.; Zhu, H.; Long, M.; Wang, J.; Yu, P.S. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2019; pp. 9154–9162.
29. Wang, Y.; Wu, H.; Zhang, J.; Gao, Z.; Wang, J.; Yu, P.S.; Long, M. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2208–2225. [CrossRef]
30. Wang, Y.; Jiang, L.; Yang, M.H.; Li, L.J.; Long, M.; Fei-Fei, L. Eidetic 3D LSTM: A Model for Video Prediction and Beyond. In *Proceedings of the International Conference on Learning Representations*; OpenReview.net: Amherst, MA, USA, 2019.
31. Yu, W.; Lu, Y.; Easterbrook, S.; Fidler, S. Crevnet: Conditionally reversible video prediction. *arXiv* **2019**, arXiv:1910.11577.
32. Guen, V.L.; Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2020; pp. 11474–11484.
33. Tan, C.; Gao, Z.; Li, S.; Li, S.Z. SimVPv2: Towards simple yet powerful spatiotemporal predictive learning. *IEEE Trans. Multimed.* **2025**, *27*, 5170–5184. [CrossRef]
34. Chang, Z.; Zhang, X.; Wang, S.; Ma, S.; Ye, Y.; Xiang, X.; Gao, W. MAU: A motion-aware unit for video prediction and beyond. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 26950–26962.
35. Hu, X.; Huang, Z.; Huang, A.; Xu, J.; Zhou, S. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 6121–6131.
36. Tang, S.; Li, C.; Zhang, P.; Tang, R. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2023; pp. 13470–13479.
37. Rasp, S.; Dueben, P.D.; Scher, S.; Weyn, J.A.; Mouatadid, S.; Thuerey, N. WeatherBench: A benchmark data set for data-driven weather forecasting. *J. Adv. Model. Earth Syst.* **2020**, *12*, e2020MS002203. [CrossRef]
38. Pathak, J.; Subramanian, S.; Harrington, P.; Raja, S.; Chattopadhyay, A.; Mardani, M.; Kurth, T.; Hall, D.; Li, Z.; Azizzadenesheli, K.; et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv* **2022**, arXiv:2202.11214.

39. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533–538. [CrossRef] [PubMed]
40. Chen, L.; Zhong, X.; Zhang, F.; Cheng, Y.; Xu, Y.; Qi, Y.; Li, H. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *arXiv* **2023**, arXiv:2306.12873.
41. Chen, K.; Han, T.; Gong, J.; Bai, L.; Ling, F.; Luo, J.J.; Chen, X.; Ma, L.; Zhang, T.; Su, R.; et al. FengWu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv* **2023**, arXiv:2304.02948.
42. Williams, R.J.; Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1989**, *1*, 270–280. [CrossRef]
43. Bengio, S.; Vinyals, O.; Jaitly, N.; Shazeer, N. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28, pp. 1171–1179.
44. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. In *Proceedings of the International Conference on Machine Learning*; Association for Computing Machinery: New York, NY, USA, 2009; pp. 41–48.
45. Shrivastava, A.; Gupta, A.; Girshick, R. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2016; pp. 761–769.
46. Paul, M.; Ganguli, S.; Dziugaite, G.K. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 20596–20607.
47. Mindermann, S.; Brauner, J.M.; Razzak, M.T.; Sharma, M.; Kirsch, A.; Xu, W.; Höltgen, B.; Gomez, A.N.; Morisot, A.; Farquhar, S.; et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2022; Volume 162, pp. 15630–15649.
48. Tan, C.; Li, S.; Gao, Z.; Guan, W.; Wang, Z.; Liu, Z.; Wu, L.; Li, S.Z. OpenSTL: A comprehensive benchmark of spatio-temporal predictive learning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 69819–69831.
49. Bengtsson, L.; Hagemann, S.; Hodges, K.I. Can climate trends be calculated from reanalysis data? *J. Geophys. Res. Atmos.* **2004**, *109*, D11111. [CrossRef]
50. Sujit, S.; Nath, S.; Braga, P.H.; Kahou, S.E. Prioritizing samples in reinforcement learning with reducible loss. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023; Volume 36, pp. 23237–23258.
51. Fu, Y.; Shao, Z.; Yu, C.; Li, Y.; An, Z.; Wang, Q.; Xu, Y.; Wang, F. Selective learning for deep time series forecasting. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2025; Volume 38, pp. 98084–98115.
52. Lin, Z.; Gou, Z.; Gong, Y.; Liu, X.; Shen, Y.; Xu, R.; Lin, C.; Yang, Y.; Jiao, J.; Duan, N.; et al. Not all tokens are what you need for pretraining. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 29029–29063.
53. Tong, R.; Liu, Y.; Shi, J.Q.; Gong, D. Coreset selection via reducible loss in continual learning. In *Proceedings of the International Conference on Learning Representations*; OpenReview.net: Amherst, MA, USA, 2025.
54. Zhang, X.; Jin, Q.; Yu, T.; Xiang, S.; Kuang, Q.; Prinet, V.; Pan, C. Multi-modal spatio-temporal meteorological forecasting with deep neural network. *ISPRS J. Photogramm. Remote Sens.* **2022**, *188*, 380–393. [CrossRef]
55. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
56. Yang, S.; Xie, Z.; Peng, H.; Xu, M.; Sun, M.; Li, P. Dataset pruning: Reducing training data by examining generalization influence. In *Proceedings of the International Conference on Learning Representations*; OpenReview.net: Amherst, MA, USA, 2023.
57. Price, I.; Sanchez-Gonzalez, A.; Alet, F.; Andersson, T.R.; El-Kadi, A.; Masters, D.; Ewalds, T.; Stott, J.; Mohamed, S.; Battaglia, P.; et al. Probabilistic weather forecasting with machine learning. *Nature* **2025**, *637*, 84–90. [PubMed]
58. Kashinath, K.; Mustafa, M.; Albert, A.; Wu, J.L.; Jiang, C.; Esmaeilzadeh, S.; Azizzadenesheli, K.; Wang, R.; Chattopadhyay, A.; Singh, A.; et al. Physics-informed machine learning: Case studies for weather and climate modelling. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2021**, *379*, 20200093. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.