



ITDiff: A Motion-Aware Diffusion Framework with Implicit Temperature Flow Fields for High-Frequency Temporal Reconstruction of Temperature

Ge Liao¹, Ju Wang^{1,2}, Qian Li^{1,2}, Liwen Wang¹, Tinger Hu¹, and Sheng Li¹

¹College of Meteorology and Oceanography, National University of Defense Technology, Changsha, China

²High Impact Weather Key Laboratory of CMA, Changsha, China

Correspondence: Ju Wang (wangjqy@vip.sina.com)

Abstract. High-frequency temporal reconstruction of temperature fields aims to recover sub-daily temperature evolution from low-frequency states, which is essential for representing rapidly changing thermal variations. However, most existing methods formulate this task as a direct estimation of intermediate temporal states without explicitly modeling the latent transition process connecting adjacent states, which limits their ability to capture underlying dynamics and may lead to temporally inconsistent predictions. To address this limitation, we propose ITDiff, a motion-aware diffusion framework for high-frequency temporal reconstruction of temperature fields, in which Implicit Temperature Flow Fields (ITFFs) are introduced to represent the latent transition process between boundary and intermediate states. First, a Flow Teacher estimates ITFFs from the boundary and target states; these ITFFs are then used to guide a Synthesis Model in reconstructing temporally consistent intermediate states. Building on this teacher-guided representation, a Conditional Diffusion Model is subsequently trained to infer plausible ITFFs from boundary conditions alone, thereby enabling the recovery of fine-grained temporal variations that are often oversmoothed by conventional approaches. Extensive experiments demonstrate that ITDiff consistently outperforms representative baselines, improving RMSE-based reconstruction accuracy by up to 48% while better preserving temporal structures across multiple scales.

1 Introduction

High-frequency temporal reconstruction of temperature aims to recover temporally continuous temperature fields from low-frequency states. Such reconstruction is important for representing rapid temperature variations and for supporting applications that require fine temporal resolution. For instance, hourly temperature information is needed to resolve diurnal variability in hydrological modeling, improve renewable energy prediction, and identify short-term heat-stress events in ecological assessments (Wang et al., 2020; Yang et al., 2023; Zhao et al., 2025; Loganathan et al., 2025). However, although hourly global reanalysis products such as ERA5 are available (Hersbach et al., 2020), many observation ally constrained datasets and regional climate simulations remain available only at 3- to 6-hour intervals (Qin et al., 2023). Moreover, generating outputs at higher temporal resolution with numerical weather prediction models is often computationally prohibitive, as finer time steps substantially



increase simulation costs. Therefore, there is a clear need for efficient and accurate data-driven methods to reconstruct hourly temperature fields from low-frequency states.(Wang et al., 2024, 2025b).

25 Existing methods for temporal reconstruction of meteorological variables can be broadly categorized into dynamical, statistical, and deep learning approaches. Among them, dynamical methods reconstruct high-frequency states by explicitly simulating the temporal evolution of atmospheric systems from coarse-resolution inputs. For example, Giorgi (2019) introduced regional climate models nested within global climate models as a dynamical framework for reconstructing high-frequency meteorological states. By explicitly modeling atmospheric evolution, dynamical methods generally produce reconstructed states with stronger physical consistency. However, as pointed out by Lopez-Gomez et al. (2025), its practical application is often constrained by high computational cost and by systematic biases inherited from the driving models. To reduce the computational burden of dynamical simulation, statistical methods reconstruct high-frequency states by learning empirical relationships from historical data. For instance, Cannon (2018) adopted linear regression and quantile mapping to establish mappings from low-frequency inputs to high-frequency states. These methods are computationally efficient and easy to implement, while their reliance on stationarity assumptions limits their ability to represent nonlinear temporal evolution under complex weather and climate conditions. To better reproduce the statistical properties of meteorological variables, Wilks (1992); Semenov and Barrow (1997) developed weather generators that simulate synthetic time series by reproducing selected statistical characteristics, such as the mean, variance, and temporal dependence. Although these stochastic simulation methods can better capture temporal statistics, they still often struggle to represent extreme events and complex regional variability.

40 More recently, deep learning, which can be divided into discriminative models and generative models, has emerged as a promising approach for temporal reconstruction by learning nonlinear relationships directly from data. Regression-based methods usually reconstruct target states through deterministic mappings from low-frequency inputs to high-frequency results. For example, Weyn et al. (2019) proposed a framework for forecasting gridded 500-hPa geopotential height from historical data, using neural networks to model the evolution of large-scale atmospheric patterns. Although it demonstrated the promise of deep learning for data-driven weather prediction, the framework remained focused on direct deterministic prediction of large-scale fields. Generative models further extended deep learning by introducing stochastic reconstruction and uncertainty modeling into meteorological applications. For example, Harris et al. (2022) developed a generative downscaling approach for precipitation forecasts based on GAN and VAE-GAN architectures, which improved the recovery of spatial variability and enabled uncertainty-aware generation. Diffusion-based methods subsequently provided a more flexible probabilistic framework for high-resolution reconstruction. Ling et al. (2024) introduced a diffusion probabilistic downscaling model for East Asian climate reconstruction, improving the representation of local details and uncertainty through ensemble sampling, while Mardani et al. (2025) proposed CorrDiff, which combines a deterministic UNet predictor with diffusion-based residual correction to better recover sharp gradients and coherent structures in km-scale atmospheric downscaling. Cachay et al. (2025) further developed rolling diffusion models for probabilistic forecasting of complex dynamics, enhancing the modeling of uncertainty growth across lead times, whereas Larsson et al. (2025) introduced Diffusion-LAM for limited-area probabilistic weather forecasting from surrounding boundary conditions. In addition, Lopez-Gomez et al. (2025) proposed a dynamical-generative downscaling framework that incorporates intermediate dynamical simulations into diffusion-based refinement, thereby improv-



ing physical plausibility. Despite these advances in uncertainty representation, structural recovery, and physical fidelity, most existing generative and hybrid frameworks still reconstruct or predict target states directly in data space, without explicitly modeling the latent transition process between adjacent states.

This limitation is particularly critical for temperature reconstruction, whose temporal evolution is shaped by the combined effects of radiative forcing, land-atmosphere exchange, boundary-layer adjustment, synoptic processes, and regional geographic contrasts (Wang et al., 2025a; Li et al., 2025). As a result, the evolution between adjacent states is often highly nonlinear, especially during strong daytime warming, nighttime cooling, frontal passages, and extreme temperature events. When such evolution is treated as a point-wise interpolation or direct regression problem, the reconstructed intermediate states tend to be overly smooth, temporally misaligned, and unable to preserve short-term variability. More importantly, these formulations focus primarily on estimating the target state itself, without explicitly modeling the latent transition process through which the boundary states evolve toward the intermediate state.

To address this issue, we propose ITDiff, a motion-aware diffusion framework for high-frequency temporal reconstruction of temperature. Instead of directly interpolating temperature values in data space, ITDiff represents the evolution between boundary data in a latent transition space through Implicit Temperature Flow Fields (ITFFs). Here, the term “flow” does not refer to an observed physical wind field or an explicitly diagnosed advection field. Rather, ITFFs serve as learned latent transition fields that encode how temperature patterns at the boundary times correspond to the intermediate target state. Based on this representation, ITDiff decomposes temporal reconstruction into two coupled components: latent transition estimation and state synthesis. Specifically, a Flow Teacher is first introduced to learn teacher-guided ITFFs from the boundary and target states, after which a Synthesis Model reconstructs the intermediate temperature state using the learned transition guidance. Building on these teacher-guided ITFFs, a Conditional Diffusion Model is then trained to generate plausible ITFFs from boundary data alone, thereby enabling the recovery of temporally consistent intermediate states while accounting for uncertainty in the underlying evolution. Compared with conventional point-wise interpolation or direct regression frameworks, the proposed formulation offers two main advantages. First, it shifts the modeling target from direct intermediate-state estimation to latent transition modeling, which is better suited to capturing nonlinear temporal evolution. Second, by combining teacher-guided latent representation learning with conditional diffusion-based generation, the framework can reconstruct intermediate states that are not only more accurate in a numerical sense but also more faithful to the temporal variation patterns implied by the boundary data.

In summary, our main contributions are as follows: 1) We propose a motion-guided framework for high-frequency temperature reconstruction, in which Implicit Temperature Flow Fields (ITFFs) are introduced to characterize the latent transitions underlying temperature evolution and to move beyond conventional point-wise reconstruction. 2) We develop a teacher-guided reconstruction scheme, in which a conditional diffusion model learns to generate ITFFs from boundary states and a synthesis model reconstructs intermediate temperature states under the guidance of the learned transition representations. 3) We establish a comprehensive evaluation framework for temperature reconstruction that examines not only reconstruction accuracy but also structural fidelity, with particular attention to spatial gradients, temporal curvature, and power spectral density. Extensive ex-



periments demonstrate that the proposed method consistently outperforms representative baselines in reconstruction accuracy and better preserves multi-scale structural characteristics.

2 Method

95 2.1 Problem Formulation

The goal of high-frequency temporal reconstruction in this study is to recover the intermediate temperature state at time t_1 , which is represented in this work by the 2-m temperature (t2m) field $T_{t_1} \in \mathbb{R}^{H \times W}$, from two low-temporal-resolution boundary states, T_{t_0} and T_{t_2} , where $t_0 < t_1 < t_2$. Here, H and W denote the spatial height and width of the temperature field, respectively.

100 Rather than directly predicting T_{t_1} in a point-wise manner, we introduce ITFFs to represent the latent transition structure that connects the boundary states to the intermediate state. An ITFF is defined as a four-channel tensor

$$\mathbf{F} = [\Delta x_{t_0 \rightarrow t_1}, \Delta y_{t_0 \rightarrow t_1}, \Delta x_{t_2 \rightarrow t_1}, \Delta y_{t_2 \rightarrow t_1}] \in \mathbb{R}^{H \times W \times 4}, \quad (1)$$

where $(\Delta x_{t_0 \rightarrow t_1}, \Delta y_{t_0 \rightarrow t_1})$ denotes the two-dimensional displacement field that maps the earlier boundary state T_{t_0} toward the target state T_{t_1} , whereas $(\Delta x_{t_2 \rightarrow t_1}, \Delta y_{t_2 \rightarrow t_1})$ denotes the corresponding displacement field that maps the later boundary
 105 state T_{t_2} toward T_{t_1} .

Δx and Δy represent displacement offsets along the horizontal and vertical spatial axes, respectively, such that the first two channels describe how the temperature pattern in T_{t_0} is aligned with the intermediate state, while the last two channels describe how the temperature pattern in T_{t_2} is aligned with the same target. In this work, these displacement fields should not be regarded as observed physical wind vectors; instead, they are learned latent transition fields that encode how the two
 110 boundary temperature patterns correspond to the intermediate state.

With this formulation, the reconstruction problem is decomposed into two steps. First, the model estimates or generates the latent transition field $\mathbf{F}$ from the available boundary information. Second, the model reconstructs the intermediate temperature state T_{t_1} by synthesizing information from T_{t_0} and T_{t_2} under the guidance of $\mathbf{F}$.

2.2 Overall Framework

115 To explicitly model the latent transition process underlying temperature evolution, we propose ITDiff, a motion-aware diffusion framework which is designed to represent temporal evolution via ITFFs, and its overall architecture is illustrated in Figure 1. The framework consists of two training stages and one inference process: the Flow Teacher and the Synthesis Model are jointly trained in the first stage, the Conditional Diffusion Model is trained in the second stage, and the Conditional Diffusion Model together with the Synthesis Model is used during inference.

120 In the first training stage, the Flow Teacher and the Synthesis Model are trained. Given a training triplet $(T_{t_0}, T_{t_1}, T_{t_2})$, the Flow Teacher first takes the boundary-to-target pairs (T_{t_0}, T_{t_1}) and (T_{t_2}, T_{t_1}) as input and estimates two pairwise displacement fields, which are then concatenated to form the ITFF associated with the intermediate state T_{t_1} . Then, this teacher-generated

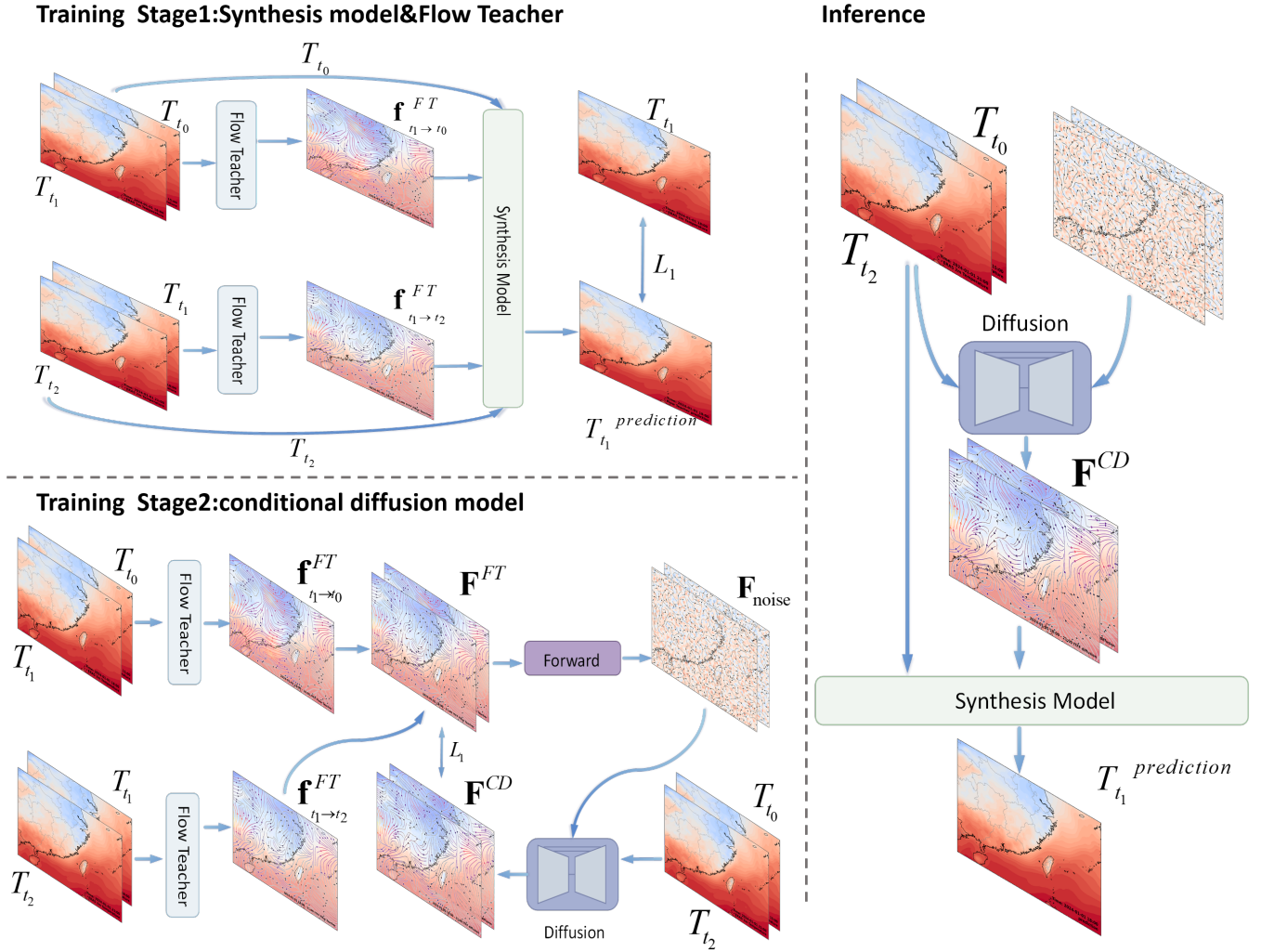


Figure 1. Overall framework of ITDiff. In Stage 1, the Flow Teacher estimates ITFFs for reconstruction. In Stage 2, the frozen Flow Teacher supervises the conditional diffusion model. During inference, the diffusion model generates ITFFs, which are used by the Synthesis Model to reconstruct the intermediate state.

ITFF is subsequently fed into the Synthesis Model, together with the two boundary temperature states T_{t_0} and T_{t_2} , to reconstruct the intermediate temperature state $T_{t_1}^{prediction}$. Finally, the reconstruction loss is computed as the difference between $T_{t_1}^{prediction}$ and the ground-truth state T_{t_1} , enabling the Flow Teacher to learn ITFFs that are useful for reconstruction and the Synthesis Model to learn how to synthesize the intermediate state under the guidance of these transition representations.

In the second training stage, the Flow Teacher is frozen and used to provide supervision for diffusion learning. For each training sample, the frozen Flow Teacher generates a teacher ITFF from the triplet $(T_{t_0}, T_{t_1}, T_{t_2})$, which is treated as the target latent representation, while the boundary states (T_{t_0}, T_{t_2}) are used as conditions for the conditional diffusion model. The



130 diffusion model is then trained to recover the teacher ITFF from noisy latent inputs, thereby learning to generate an ITFF from boundary states alone rather than reconstructing temperature directly.

During inference, the conditional diffusion model starts from Gaussian noise and progressively generates an ITFF conditioned on the two boundary states T_{t_0} and T_{t_2} , after which the generated ITFF is passed to the Synthesis Model together with the boundary temperature states to produce the final reconstruction of the intermediate temperature state. In this way, the diffusion model is responsible for generating the ITFF, whereas the Synthesis Model converts it into the reconstructed temperature field.

2.3 Flow Teacher

The Flow Teacher is introduced to estimate the ITFF associated with the intermediate temperature state. Given a triplet $(T_{t_0}, T_{t_1}, T_{t_2})$, the Flow Teacher provides a teacher-guided ITFFs. Rather than directly predicting the intermediate temperature state, it learns how the two boundary states are aligned with the intermediate state in a latent displacement space.

Specifically, the Flow Teacher estimates two displacement fields, one from the earlier boundary state T_{t_0} to the intermediate state T_{t_1} and the other from the later boundary state T_{t_2} to the same intermediate state, which are defined as

$$\mathbf{f}_{t_0 \rightarrow t_1} = G_\theta(T_{t_0}, T_{t_1}), \quad \mathbf{f}_{t_2 \rightarrow t_1} = G_\theta(T_{t_2}, T_{t_1}), \quad (2)$$

where $G_\theta(\cdot)$ denotes the Flow Teacher. Each displacement field contains two channels corresponding to horizontal and vertical offsets at each grid location, and the final ITFF is constructed by concatenating the two pairwise displacement fields:

$$\mathbf{F}_{t_1} = [\mathbf{f}_{t_0 \rightarrow t_1}, \mathbf{f}_{t_2 \rightarrow t_1}] \in \mathbb{R}^{H \times W \times 4}. \quad (3)$$

In implementation, the Flow Teacher is built on the RAFT backbone (Teed and Deng, 2020; Uusinoka et al., 2025; Wang et al., 2023), which is a dense correspondence model based on iterative correlation refinement. The model is applied separately to the two boundary-to-intermediate pairs, and the resulting pairwise displacement fields are concatenated to form the four-channel ITFF.

It should be noted that the Flow Teacher is not trained with ground-truth flow labels. Instead, it is optimized jointly with the Synthesis Model through the final reconstruction objective. Specifically, the ITFF generated by the Flow Teacher is passed to the Synthesis Model together with the two boundary temperature states, after which the reconstructed intermediate state is compared with the ground-truth state T_{t_1} . The resulting reconstruction loss is then back-propagated through the Synthesis Model to the Flow Teacher, so that the Flow Teacher learns transition representations that are useful for reconstructing the intermediate temperature state rather than matching a predefined displacement target.

2.4 Synthesis Model

The Synthesis Model is used to reconstruct the intermediate temperature state from the two boundary states and the ITFF. Its role is not to estimate the ITFF itself, but to convert the teacher-generated or diffusion-generated ITFF into the final reconstruction of the target state T_{t_1} .

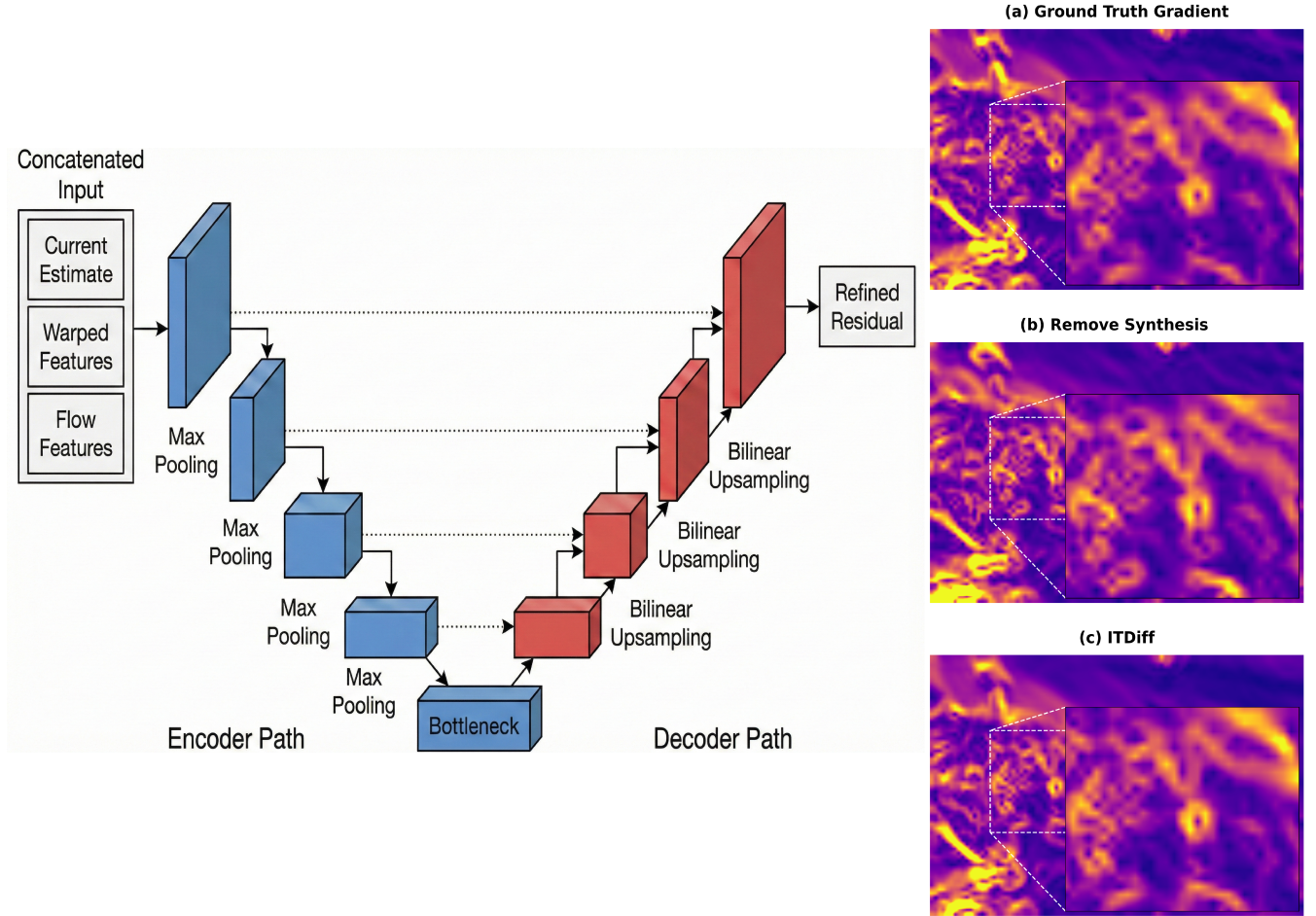


Figure 2. Architecture of the Synthesis Model. The model takes the current estimate, warped boundary features, and flow features as concatenated input, and reconstructs the intermediate temperature state through a multi-scale encoder-decoder structure with progressive downsampling, bilinear upsampling, and skip connections. The right panel presents an ablation comparison, showing that removing the synthesis module leads to weaker recovery of fine-scale structural details than the full ITDiff framework.

Given the ITFF $\mathbf{F}_{t_1}$, the two boundary states are first aligned toward the intermediate time through differentiable backward warping. Let $\mathbf{f}_{t_0 \rightarrow t_1}$ and $\mathbf{f}_{t_2 \rightarrow t_1}$ denote the two pairwise displacement fields contained in the ITFF. The warped boundary states are written as

$$\tilde{T}_{t_0 \rightarrow t_1} = W(T_{t_0}, \mathbf{f}_{t_0 \rightarrow t_1}), \quad \tilde{T}_{t_2 \rightarrow t_1} = W(T_{t_2}, \mathbf{f}_{t_2 \rightarrow t_1}), \quad (4)$$

165 where $W(\cdot)$ denotes the differentiable backward warping operator.

In addition to the raw boundary states, the model extracts feature maps from the two boundary inputs and warps these features using the same displacement fields. As shown in Figure 2, the Synthesis Model takes three types of information as



input, namely the current estimate of the intermediate state, the warped boundary features, and the flow features derived from the ITFF, which are concatenated channel-wise and then fed into a multi-scale encoder-decoder network.

170 The encoder progressively aggregates contextual information through several downsampling stages, whereas the decoder gradually restores spatial resolution through bilinear upsampling and skip connections, so that the model can combine large-scale contextual structure with fine-scale local details. At the coarsest scale, the current estimate is initialized as the average of the two warped boundary states,

$$\hat{T}_{t_1}^{(L)} = \frac{1}{2} \left(\tilde{T}_{t_0 \rightarrow t_1}^{(L)} + \tilde{T}_{t_2 \rightarrow t_1}^{(L)} \right), \quad (5)$$

175 where L denotes the coarsest scale level. Starting from this initialization, the network progressively refines the estimate from coarse to fine scales until the final reconstruction $T_{t_1}^{\text{prediction}}$ is obtained.

In implementation, the Synthesis Model outputs a refined estimate of the intermediate state at each scale and is supervised by the reconstruction error between the final prediction and the ground-truth state T_{t_1} . In this study, the main training objective is defined as the L1 reconstruction loss. As further illustrated by the ablation examples in the right panel of Figure 2, removing
 180 the Synthesis Model weakens the recovery of fine-scale structural details.

2.5 Conditional Diffusion Model

After the first-stage training, the Flow Teacher is frozen and used to provide teacher-guided ITFFs for diffusion learning. To enable ITFF generation from the two boundary states alone, we propose a conditional diffusion model that learns the conditional distribution of ITFFs associated with the intermediate state T_{t_1} .

185 Let $\mathbf{F}_{t_1}$ denote the ITFF generated by the Flow Teacher corresponding to the training triplet $(T_{t_0}, T_{t_1}, T_{t_2})$. During training, Gaussian noise is progressively added to this ITFF according to the forward diffusion process,

$$q(\mathbf{z}_k | \mathbf{F}_{t_1}) = \mathcal{N}(\mathbf{z}_k; \sqrt{\bar{\alpha}_k} \mathbf{F}_{t_1}, (1 - \bar{\alpha}_k) \mathbf{I}), \quad (6)$$

where $\mathbf{z}_k$ denotes the noisy latent variable at diffusion step k , and $\bar{\alpha}_k$ is determined by the predefined noise schedule.

The denoising network takes the noisy ITFF and the two boundary temperature states T_{t_0} and T_{t_2} as input, and is trained to
 190 recover the clean ITFF conditioned on the boundary states. Accordingly, the diffusion objective is defined in the ITFF space as

$$\mathcal{L}_{\text{diff}} = \mathbb{E}[\|D_{\theta}(\mathbf{z}_k, k, T_{t_0}, T_{t_2}) - \mathbf{F}_{t_1}\|_1], \quad (7)$$

where $D_{\theta}(\cdot)$ denotes the denoising network.

A key component of this stage is the customized U-Net backbone used for conditional denoising. As shown in Figure 3,
 195 the two boundary temperature states and the noisy ITFF are first compressed and encoded before entering the main denoising backbone, after which denoising is performed in a compressed feature space through a residual encoder-decoder architecture with downsampling and upsampling blocks connected by skip connections. This design reduces the computational cost of iterative diffusion while preserving the large-scale transition structure required for ITFF generation. Another component is the

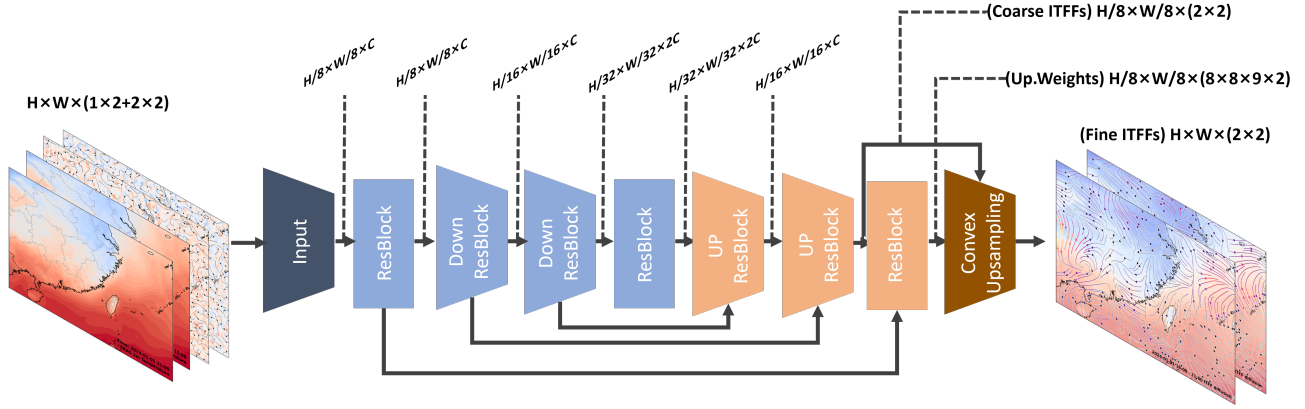


Figure 3. Architecture of the conditional diffusion model. The model takes the two boundary temperature states together with a noisy latent ITFF as input, performs denoising in a compressed feature space through residual downsampling and upsampling blocks, and finally restores the full-resolution latent transition field by a learned convex upsampling module.

final upsampling module used to restore the coarse ITFF to the original spatial resolution. Because denoising is performed in a compressed latent space, the network first predicts a coarse ITFF together with the corresponding upsampling weights, after which the final upsampling module reconstructs the high-resolution ITFF for subsequent synthesis.

During inference, only the two boundary temperature states T_{t_0} and T_{t_2} are available. Starting from Gaussian noise, the conditional diffusion model progressively denoises the latent variable into a generated ITFF conditioned on the boundary states, after which the resulting high-resolution ITFF is passed to the Synthesis Model to reconstruct the final intermediate temperature state $T_{t_1}^{\text{prediction}}$.

Compared with direct temperature generation, this formulation shifts the generative task from temperature reconstruction to ITFF generation and thereby aligns the diffusion process with the transition representation used in the overall framework. In addition, the customized denoising backbone combines compressed-space conditional encoding with learned upsampling, which makes it more suitable for generating structurally coherent ITFFs for subsequent temperature-state synthesis.

2.6 Implementation and Training Details

The proposed ITDiff is implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU with 24 GB memory. In the final model configuration, the hidden dimensions of the conditional diffusion model are set to (288, 256, 256). Training is conducted in two stages. The Flow Teacher and Synthesis Model are first pre-trained for 100 and 150 epochs, respectively, and the diffusion model is then trained for 500 epochs with a mini-batch size of 32. The initial learning rate of the diffusion model is set to 2×10^{-4} . To improve training stability, an exponential moving average (EMA) with a decay rate of 0.9999 is applied during diffusion training. The number of diffusion time steps is set to 1000 during training, while 8 sampling steps are used during inference to improve efficiency. For motion normalization, the generated flow values are scaled by a factor of 40 to keep



the magnitude of the ITFFs within a stable numerical range for diffusion learning. For fair comparison, all baseline methods are evaluated on the same dataset under the same experimental setting.

220 **3 Experimental Results**

3.1 Overall Performance

To evaluate the proposed method, we use ERA5 reanalysis data from the region spanning 18°–34°N and 106°–130°E, where interactions among monsoonal circulation, coastal effects, and complex atmospheric dynamics produce pronounced spatiotemporal variability. Data from 2014 to 2023 are used as the test set, while data from 2024 are used for validation. We compare ITDiff with representative baselines of different types, including Linear, CNN, and U-Net. Specifically, Linear is a traditional regression-based method, whereas CNN and U-Net are representative deterministic deep learning baselines. The quantitative comparison on the 2024 test set is reported in Table 1 reports the quantitative comparison on the 2024 test set, where the full ITDiff framework, i.e., Diffusion (8 steps) + Synthesis, achieves the best performance on all evaluation metrics, including MSE, RMSE, MAE, and correlation. Compared with these baselines, ITDiff substantially reduces reconstruction errors while maintaining the highest correlation with the reference temperature fields. In addition, the comparison between “Synthesis Only” and “Diffusion (8 steps) + Synthesis” shows that introducing ITFF-based transition modeling further improves intermediate-state reconstruction.

Table 1. Overall performance comparison of different methods on the 2024 test set.

Method	MSE (K^2)	RMSE (K)	MAE (K)	Correlation
Linear	0.6163	0.7667	0.4905	0.9809
CNN	0.3611	0.5911	0.3632	0.9890
U-Net	0.2618	0.5074	0.3043	0.9917
Synthesis Only	<u>0.2143</u>	<u>0.4629</u>	<u>0.3062</u>	<u>0.9945</u>
Diffusion (8 steps) + Synthesis	0.1609	0.4010	0.2295	0.9948

Figure 4 provides a diagnostic analysis of the reconstruction results from three complementary aspects, including error variation across temperature regimes, global error statistics, and performance under different target times. From the regime-wise evaluation in Figure 4a–c, it can be observed that RMSE is lower in the dominant middle-temperature ranges and larger at the warmest extreme range, while the mean bias remains close to zero in most regimes and the spatial correlation stays consistently high, especially within the major temperature ranges that dominate the dataset. These results indicate that the proposed framework maintains stable reconstruction behavior over the main part of the temperature distribution, although reconstruction becomes more difficult at temperature extremes.

240 The global statistical characteristics shown in Figure 4d and Figure 4e further support this observation. Specifically, the prediction-error distribution is sharply centered near zero, which indicates that most reconstruction errors are small in magni-

tude, while the reconstructed temperature distribution closely matches the ERA5 reference distribution, suggesting that ITDiff preserves the overall statistical structure of the target temperature field. In addition, Figure 4f shows that the RMSE varies with target time and becomes larger during periods more strongly affected by the diurnal cycle. Nevertheless, the overall error remains stable across different target times.

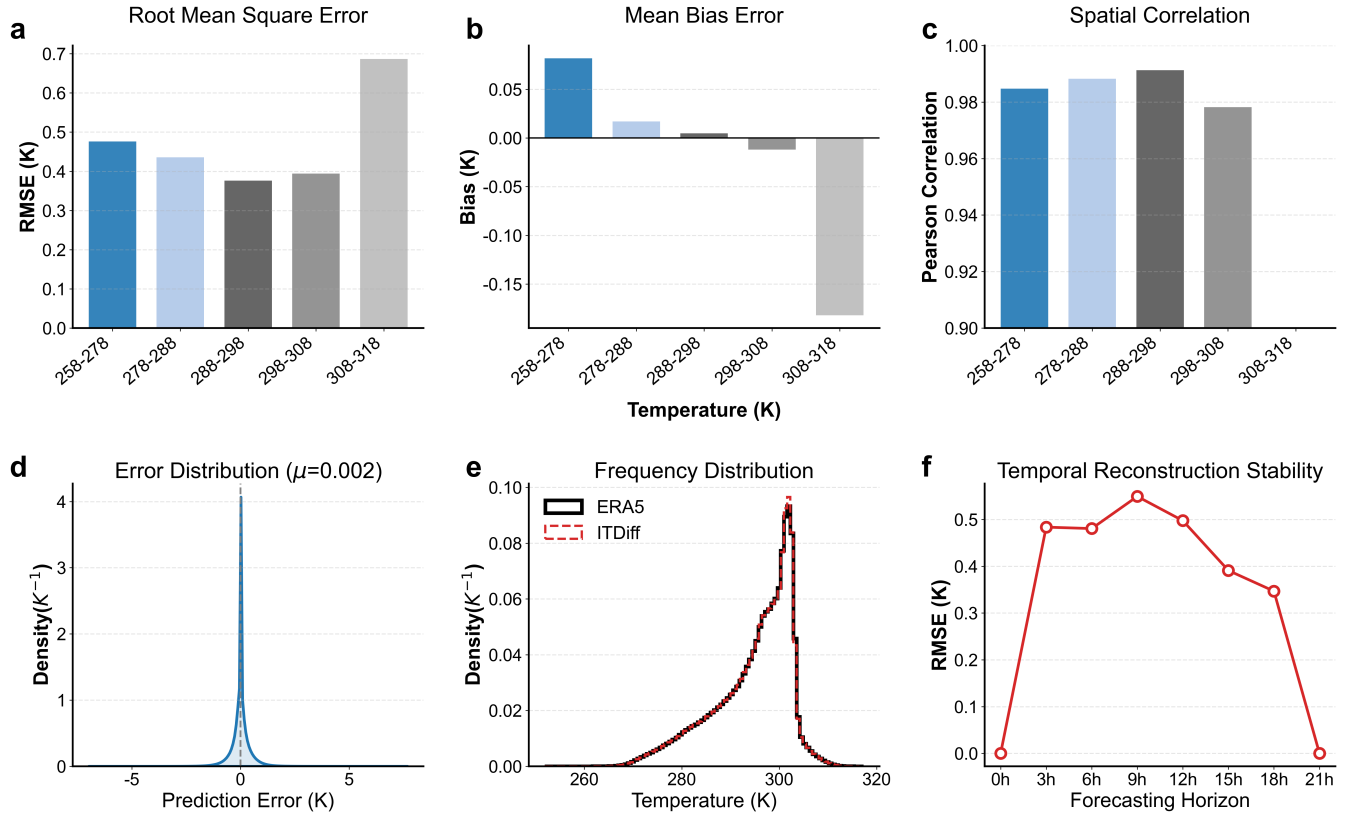


Figure 4. Comprehensive statistical evaluation of ITDiff on the 2024 test set. (a) RMSE across different temperature ranges. (b) Mean bias error across temperature ranges. (c) Spatial correlation across temperature ranges. (d) Distribution of prediction errors. (e) Comparison between the reconstructed temperature distribution and the ERA5 reference distribution. (f) RMSE at different reconstruction horizons, showing the temporal stability of the method.

3.2 Temporal Reconstruction Accuracy

To evaluate the temporal reconstruction capability of the proposed framework, we examine its performance from two complementary aspects, namely long-term statistical stability and short-term diurnal evolution, as summarized in Figure 5.

At the monthly scale, the distribution of mean absolute error (MAE) throughout 2024 shows that ITDiff consistently achieves the lowest median error across all months while maintaining a relatively narrow interquartile range. This result indicates that the proposed method not only improves average reconstruction accuracy but also provides more stable performance under



different seasonal conditions. By comparison, the Linear, CNN, and U-Net baselines exhibit both larger median errors and stronger month-to-month variability.

At the short-term scale, we further examine four consecutive days from 2024-08-30 to 2024-09-02, during which the temperature curves contain pronounced diurnal variations and relatively sharp warming transitions, making this period suitable for evaluating reconstruction quality under rapidly changing conditions. As shown in Figure 5b–e, ITDiff follows the ground-truth temperature evolution more closely than the baseline methods across all four cases, especially during the warming stage toward the daily peak and the subsequent cooling stage. Linear interpolation consistently underestimates the peak amplitude because it cannot represent the nonlinear transition between the two boundary states, whereas CNN and U-Net reduce part of this error but still show visible deviations in amplitude or phase near the peak and during the cooling period. In contrast, ITDiff remains closer to the ground truth throughout the diurnal cycle.

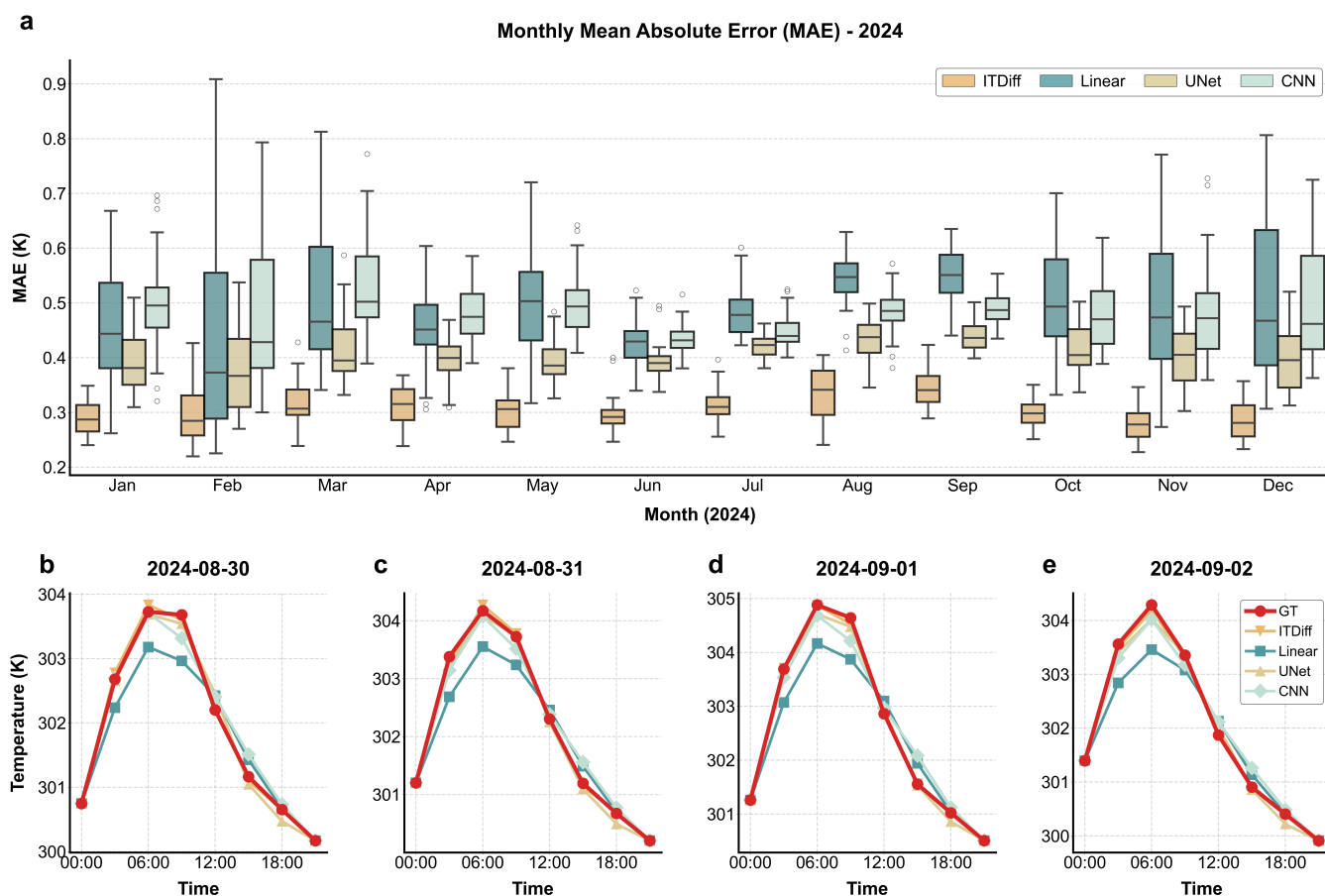


Figure 5. Evaluation of temporal reconstruction accuracy and diurnal-cycle consistency in 2024. (a) Monthly distribution of mean absolute error (MAE) for different methods. (b–e) Reconstructed diurnal temperature evolution on four consecutive days from 2024-08-30 to 2024-09-02, compared with the ground truth (GT).

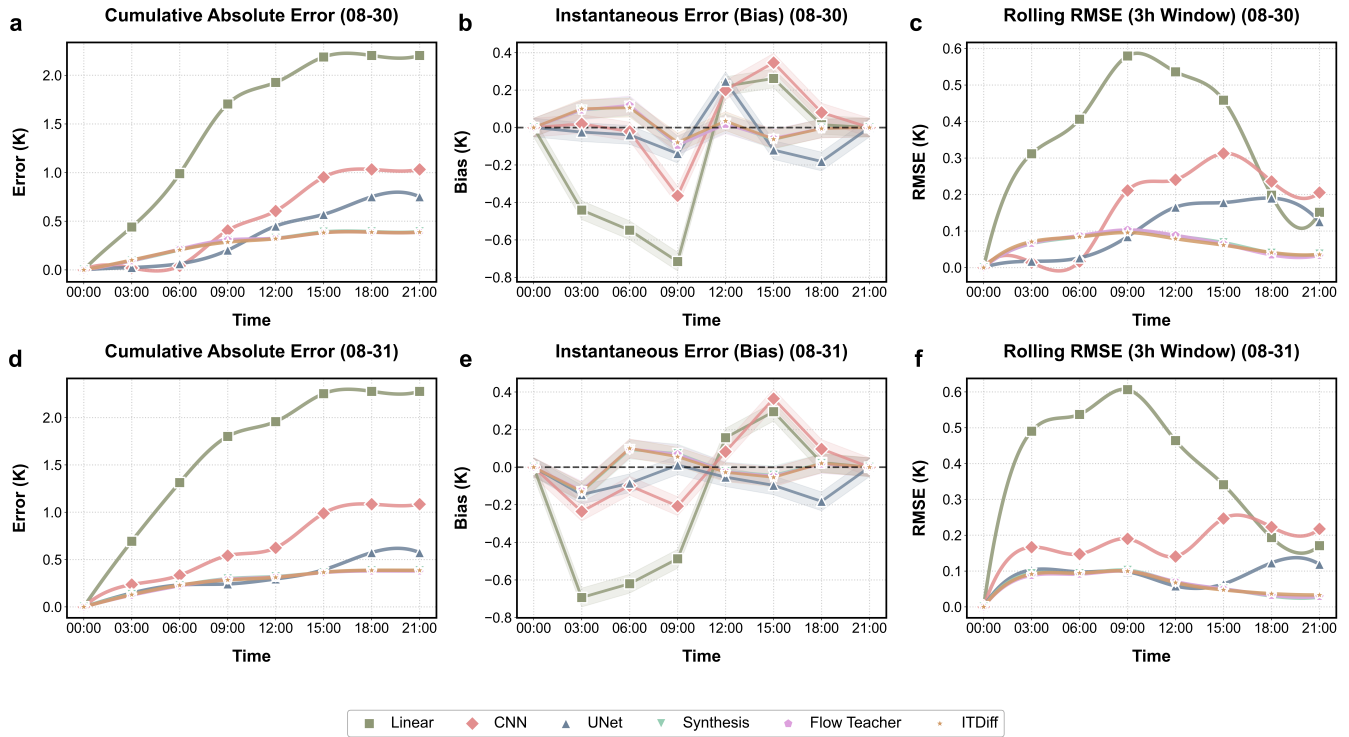


Figure 6. Short-term temporal error dynamics on 2024-08-30 and 2024-08-31. (a,d) Cumulative absolute error. (b,e) Instantaneous error (bias). (c,f) Rolling RMSE computed with a 3-hour window.

Figure 6 examines short-term error dynamics on two representative days, 2024-08-30 and 2024-08-31, both of which exhibit evident daytime warming followed by temperature adjustment. These cases therefore provide a useful test of reconstruction quality under rapidly varying conditions. Across both days, the cumulative absolute error remains substantially lower for ITDiff than for the deterministic baselines, while Linear shows the fastest error growth and CNN and U-Net also accumulate noticeably larger errors over time. By comparison, ITDiff stays at a consistently low level and remains close to the synthesis-based and teacher-guided settings.

This advantage is also reflected in the instantaneous error curves, where the largest differences appear during the rapid warming period and the following daytime transition. Linear shows a pronounced negative bias in the morning and large positive deviations in the afternoon, reflecting its limited ability to represent nonlinear temperature evolution within the reconstruction interval. CNN and U-Net reduce part of this error, but both still exhibit larger fluctuations around zero. In contrast, ITDiff remains close to zero at most reconstruction times on both days.

A similar pattern appears in the rolling RMSE curves, which provide a localized view of reconstruction difficulty over time. Linear produces the largest local error on both days, with pronounced peaks during the main daytime transition period, whereas CNN and U-Net reduce these peaks but remain clearly above ITDiff. Meanwhile, ITDiff maintains one of the lowest rolling RMSE levels throughout and stays broadly comparable to the synthesis-based and teacher-guided settings.



3.3 Structural Diagnostics of Temporal Reconstruction

This section evaluates whether the reconstructed fields preserve the structural properties of short-term temperature evolution. We focus on three diagnostics, namely temperature-gradient variation, temporal curvature, and spatial power spectra, because these quantities provide a stricter evaluation of structural fidelity than pointwise error metrics alone.

The horizontal temperature gradient magnitude is defined as

$$\nabla T = \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial y} \right) \quad (8)$$

and

$$|\nabla T| = \sqrt{\left(\frac{\partial T}{\partial x} \right)^2 + \left(\frac{\partial T}{\partial y} \right)^2}. \quad (9)$$

To assess how these thermal contrasts evolve over time, we define the temporal rate of gradient change as

$$\mathcal{R}_g = \frac{|\nabla T|_{t+\Delta t} - |\nabla T|_t}{\Delta t}. \quad (10)$$

To further characterize the nonlinearity of short-term temperature evolution, we define the temporal curvature of the temperature-evolution curve as

$$\kappa = \frac{\left| \frac{\partial^2 T}{\partial t^2} \right|}{\left(1 + \left(\frac{\partial T}{\partial t} \right)^2 \right)^{3/2}}. \quad (11)$$

The gradient-rate analysis first identifies the period of strongest structural adjustment within the diurnal cycle. As shown in Figure 7, the strongest signals are concentrated along the coastal zone, whereas inland variations are comparatively weaker and more fragmented. Structural adjustment is most active from 6h to 12h, with the strongest contrast appearing during the 9h–12h interval. After 15h, the gradient-rate signals weaken substantially and the daily mean field becomes comparatively smooth. This pattern motivates the subsequent comparison during the 9h–12h interval, when structural evolution is most pronounced.

Based on this interval, Figure 8 compares the reconstructed temperature-gradient-rate fields during the main transition period. The ground-truth field contains a sharp coastal band together with substantial fine-scale variability over land. Linear interpolation strongly smooths these structures and weakens the coastal contrast, whereas CNN and U-Net recover part of the spatial pattern but still lose local detail and blur the narrow coastal structure. The stronger reconstruction settings improve the result further, and ITDiff is visually among the closest results to the ground truth, especially in the recovery of the coastal transition band and localized inland variations. The temporal consistency of gradient reconstruction is further examined in Figure 9. The left and middle panels show the domain-mean evolution of the latitudinal and longitudinal gradient components, while the right panel shows the temporal evolution of the total gradient magnitude at the coastal point (24.5°N, 118.1°E). In both directional components, ITDiff remains close to the reference evolution throughout the day, while Linear shows the largest deviation and CNN and U-Net still exhibit visible amplitude or phase errors. The difference is especially clear in the longitudinal gradient, where the daytime transition is reproduced more faithfully by ITDiff. At the coastal point, the observed

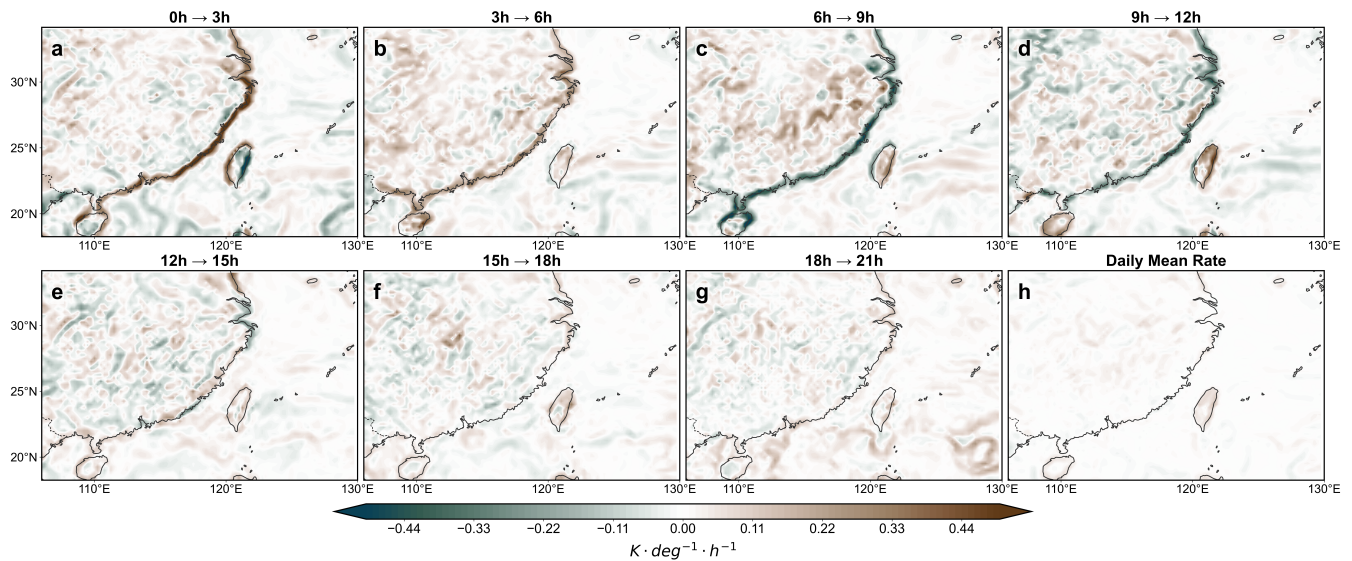


Figure 7. Temporal evolution of the temperature-gradient rate of change on 2024-08-30. Panels (a)–(g) show the gradient change rate for consecutive 3-hour intervals from 0h→3h to 18h→21h, and panel (h) shows the daily mean rate.

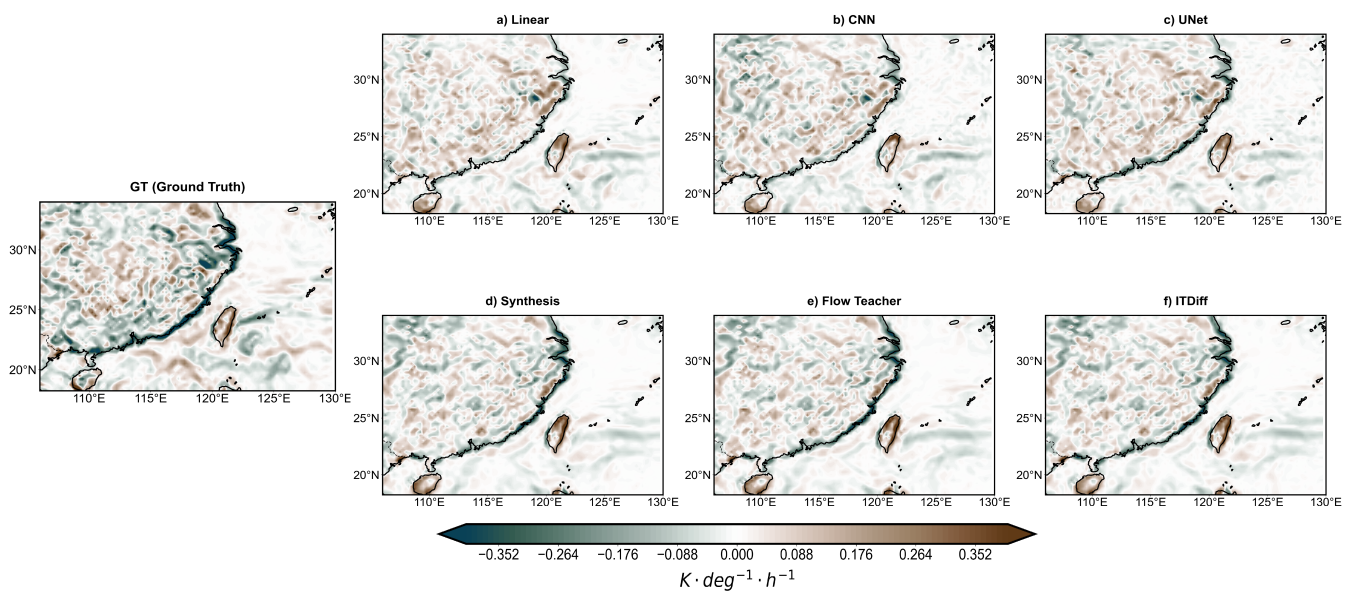


Figure 8. Comparison of temperature-gradient reconstruction quality during the 9h–12h transition period on 2024-08-30. The panels show the ground truth (GT), Linear, CNN, UNet, Synthesis, Flow Teacher, and ITDiff results, respectively.

gradient magnitude reaches a sharp peak near 06:00; Linear substantially underestimates this peak, CNN and U-Net remain visibly biased, and ITDiff reproduces both the peak intensity and the subsequent decay more closely.

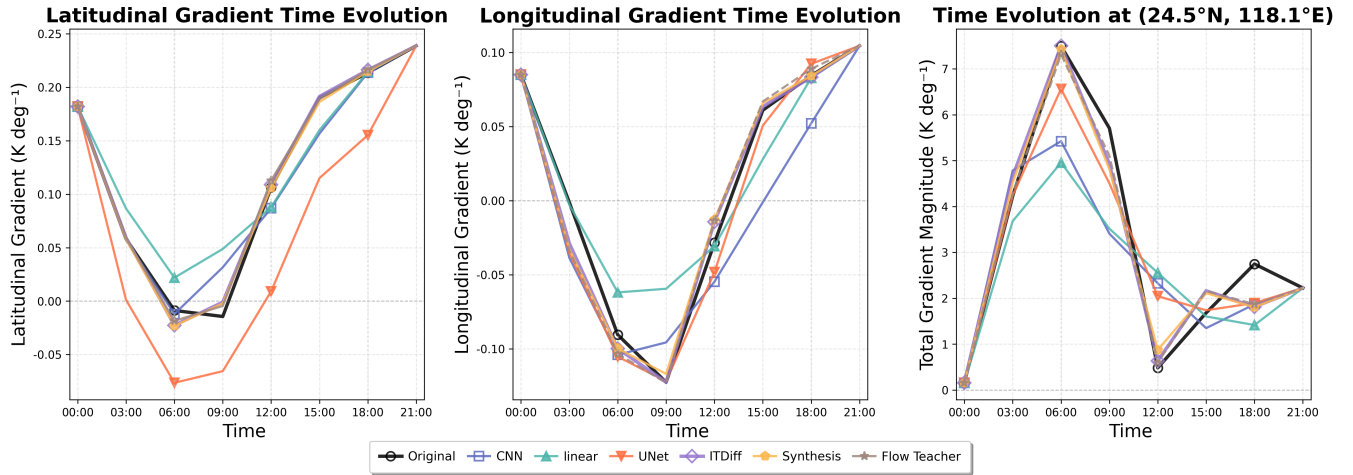


Figure 9. Temporal evolution of temperature gradients on 2024-08-30. The left and middle panels show the latitudinal and longitudinal gradient evolution, respectively. The right panel shows the temporal evolution of the total gradient magnitude at the coastal point (24.5°N, 118.1°E).

Temporal curvature provides a complementary view of structural fidelity beyond first-order gradient behavior. As shown in Figure 10, the ground-truth field contains several spatially organized high-curvature regions, which mark where short-term temperature evolution is most nonlinear. Linear interpolation reproduces the broad distribution of these regions but noticeably smooths their boundaries and enlarges their spatial extent. CNN and U-Net recover the main high-curvature structures more clearly, yet visible deviations remain in local extent and boundary sharpness. Among the compared methods, ITDiff provides the closest match to the reference field, particularly in the location, extent, and boundary definition of the dominant high-curvature regions. A broader comparison across spatial scales is provided by Figure 11, which reports the ratio of reconstructed to reference power spectral density in four seasons. Across all seasons, ITDiff remains closest to the reference ratio of 1.0 over most wavenumbers, showing better preservation of the spatial spectral structure. By comparison, U-Net exhibits the largest spectral deficit at medium-to-high wavenumbers in most seasons, while CNN also remains systematically below the reference level. Linear interpolation shows stronger seasonal dependence: although it loses spectral energy in spring and summer, it stays closer to the reference ratio in autumn and winter. Overall, ITDiff preserves multi-scale spatial variability more consistently than the baseline methods, with the clearest advantage appearing in summer, when temperature variability is strongest.

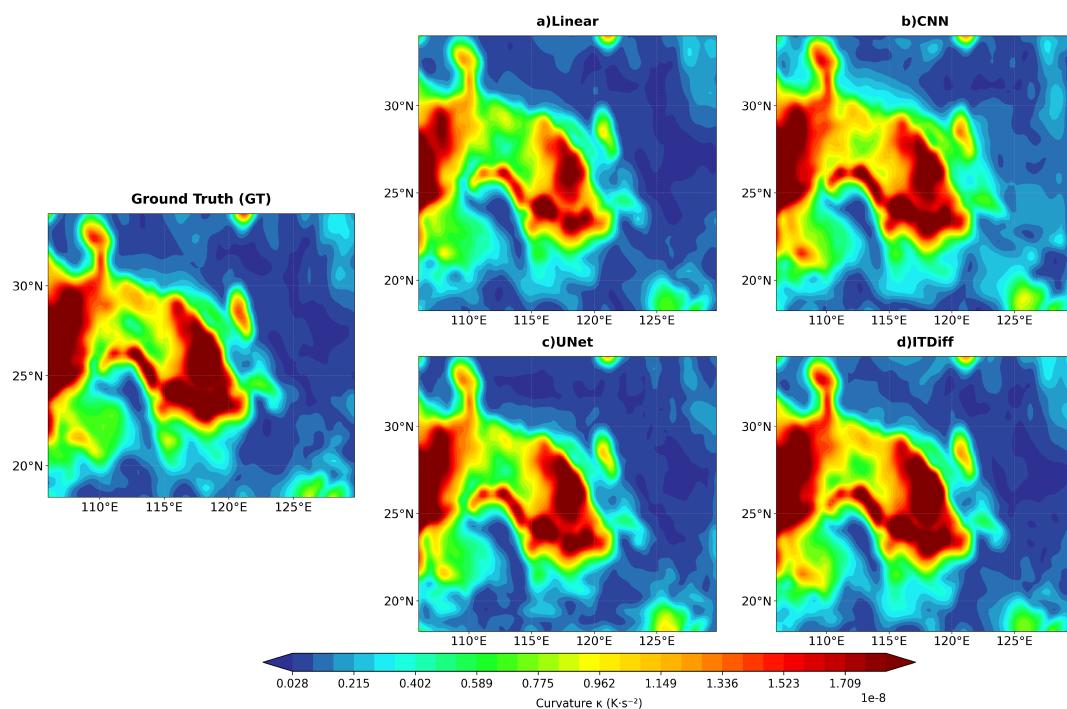


Figure 10. Comparison of the spatial distribution of temporal curvature κ for different reconstruction methods. The panels show the ground truth (GT), Linear, CNN, UNet, and ITDiff results, respectively.

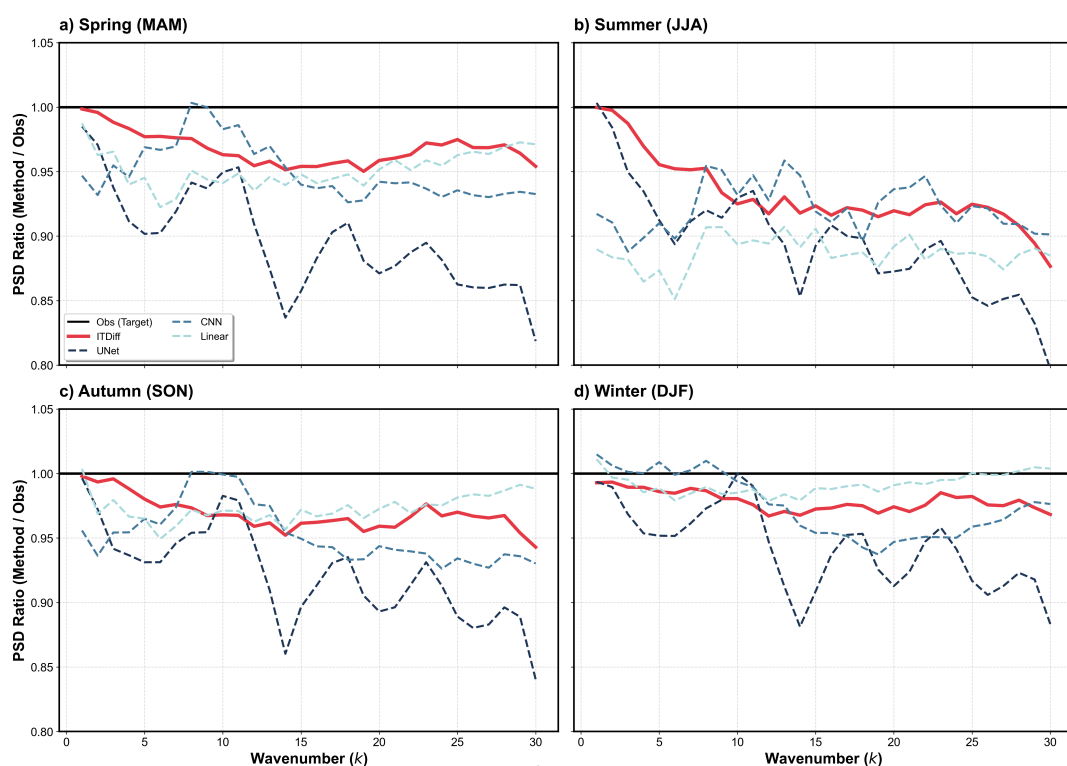


Figure 11. Seasonal comparison of the spatial power spectral density ratio between reconstructed fields and the ERA5 reference on the 2024 test set. Panels show the results for (a) spring (MAM), (b) summer (JJA), (c) autumn (SON), and (d) winter (DJF).

3.4 Case Study: January 2024 Cold Wave

The statistical evaluations above describe the overall behavior of the proposed framework over the full test period, whereas the following case study focuses on its performance during a strong cold-wave event in January 2024, when the temperature field exhibits sharp spatial contrasts and rapid transitions. Under such conditions, reconstruction quality depends not only on the recovery of the large-scale cold anomaly but also on the preservation of the associated frontal structure.

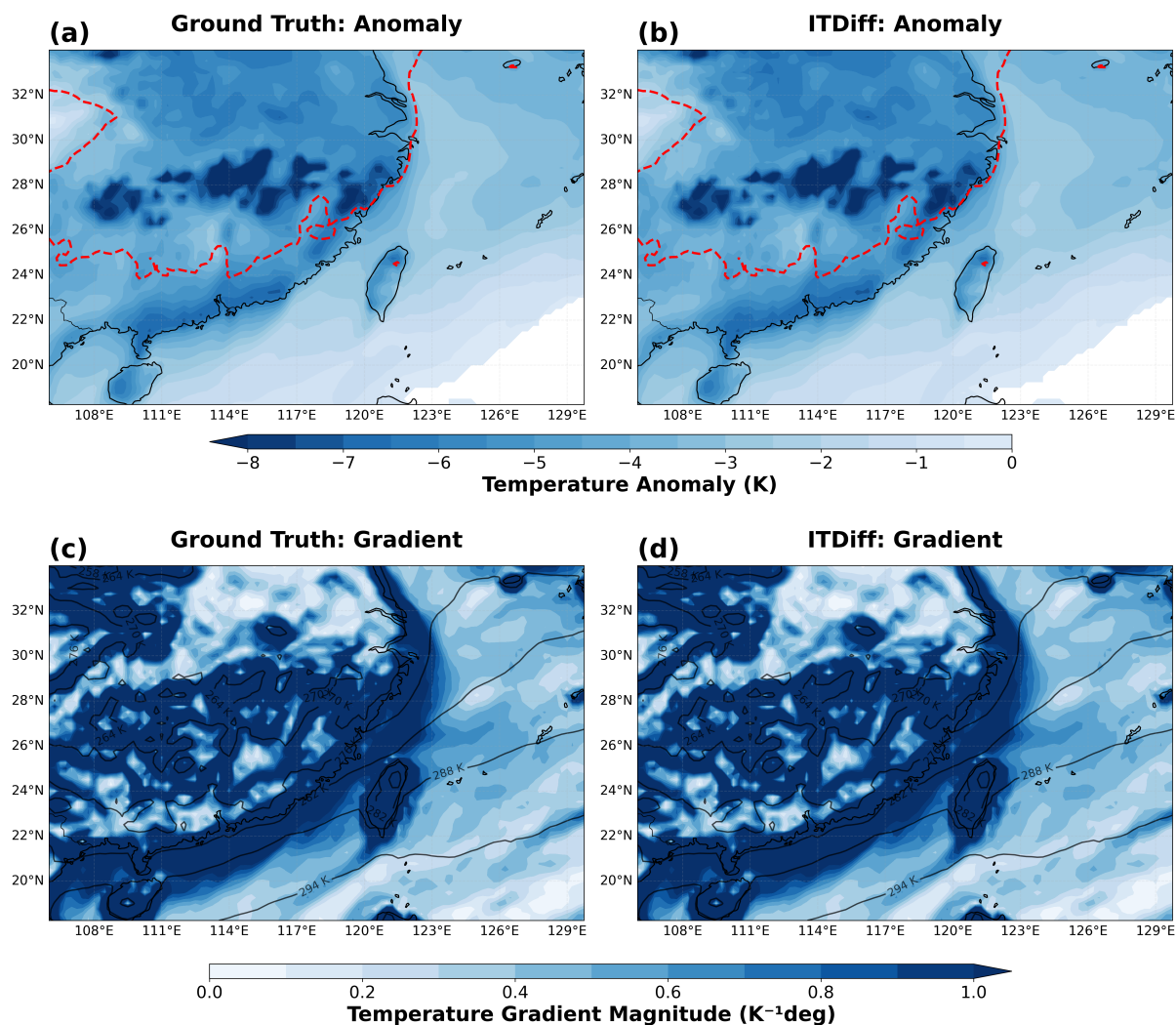


Figure 12. Spatial comparison during the January 2024 cold-wave event. Panels (a) and (b) show the ground-truth and ITDiff-reconstructed temperature anomaly fields, respectively. Panels (c) and (d) show the corresponding temperature-gradient magnitude fields. The red dashed contour in the upper panels denotes the 273.15 K isotherm.

As shown in Figure 12, the reference field during the cold-wave event is characterized by a broad cold anomaly over southeastern China together with a southward intrusion of cold air. The ITDiff reconstruction reproduces the main large-scale anomaly pattern, while the corresponding gradient field retains the major transition zone associated with the frontal boundary.

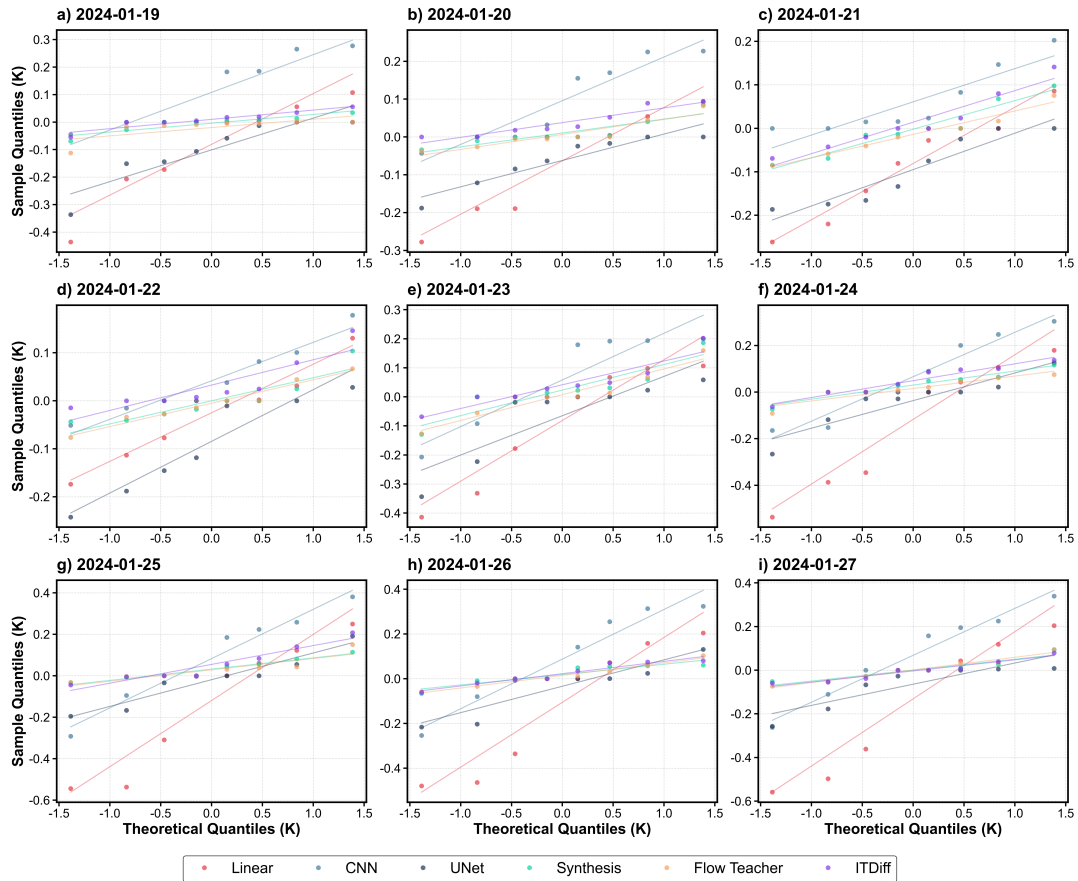


Figure 13. Q–Q plots of reconstruction residuals during 19–27 January 2024. Panels (a)–(i) correspond to individual days from 2024-01-19 to 2024-01-27. Points and fitted lines are shown for the compared methods, including Linear, CNN, UNet, Synthesis, Flow Teacher, and ITDiff. Curves closer to the theoretical diagonal indicate better agreement between the residual distribution and the reference normal distribution.

325 The residual behavior during this event is further examined over 19–27 January 2024 using the Q–Q plots in Figure 13. Across these nine days, Linear shows the largest departures at both lower and upper quantiles, reflecting stronger tail errors during the event. CNN and U-Net reduce part of this deviation, but their residual distributions remain more dispersed than those of the stronger reconstruction settings. By comparison, the curves of Synthesis, Flow Teacher, and ITDiff stay more concentrated and stable across days, with ITDiff remaining among the closest to the reference behavior over most quantiles.

330 Taken together, the cold-wave results show that the proposed framework maintains both better structural reconstruction and more stable residual behavior under strong thermal contrasts.



4 Conclusions

This study presents ITDiff, a two-stage motion-aware framework for high-frequency temperature reconstruction based on Implicit Temperature Flow Fields. By decoupling transition generation from temperature-state synthesis and introducing diffusion modeling in the latent flow space, the proposed framework improves intermediate-state reconstruction accuracy while better preserving structural characteristics of temperature evolution, including spatial gradients, temporal curvature, and spatial spectral structure. Experiments on ERA5 data over East Asia, with 2024 used as the test year, show that ITDiff consistently outperforms representative baselines under both routine diurnal conditions and cold-wave events.

Future work will extend the current framework in several directions. First, we will generalize ITDiff to jointly reconstruct multiple atmospheric variables, such as humidity and wind fields, so that cross-variable dependencies can be better represented. Second, we will incorporate stronger physics-informed constraints to improve physical consistency during reconstruction. Third, we will evaluate cross-region generalization to assess the applicability of the framework under different climatic conditions. These extensions will further improve the usefulness of ITDiff for high-frequency meteorological reconstruction.

Code availability. The exact model version used for all simulations and analyses in this study is archived in Zenodo under a persistent DOI:<https://doi.org/10.5281/zenodo.2331888>.

Data availability. All data necessary to reproduce the results of this work can be downloaded from the Copernicus Climate Change Service (Copernicus Climate Change Service, 2024a, b) at <https://doi.org/10.24381/cds.adbb2d47>.

Author contributions. GLiao was primarily responsible for the design of the model and conducting the experiments. JWang prepared the experimental datasets and organized the entire research project.QLi, LWW, TH and SLi contributed to the optimization of the experimental code

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.



355 *Acknowledgements.* This research was funded by the National Natural Science Foundation of China (Grant No. 42075139, U2242201, 41305138, 42105146), the China Postdoctoral Science Foundation (Grant No. 2017M621700), Hunan Province Natural Science Foundation (Grant No. 2021JC0009).



References

- Cachay, S. R., Aittala, M., Kreis, K., Brenowitz, N., Vahdat, A., Mardani, M., and Yu, R.: Elucidated Rolling Diffusion Models for Probabilistic Forecasting of Complex Dynamics, arXiv preprint arXiv:2506.20024, 2025.
- Cannon, A. J.: Multivariate Quantile Mapping Bias Correction: An N-Dimensional Probability Density Function Transform for Climate Model Simulations of Multiple Variables, *Climate Dynamics*, 50, 31–49, <https://doi.org/10.1007/s00382-017-3580-6>, 2018.
- Giorgi, F.: Thirty Years of Regional Climate Modeling: Where Are We and Where Are We Going Next?, *Journal of Geophysical Research: Atmospheres*, 124, 5696–5723, <https://doi.org/10.1029/2018JD030094>, 2019.
- Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., and Palmer, T. N.: A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts, *Journal of Advances in Modeling Earth Systems*, 14, e2022MS003120, <https://doi.org/10.1029/2022MS003120>, 2022.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Holm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- Larsson, E., Oskarsson, J., Landelius, T., and Lindsten, F.: Diffusion-LAM: Probabilistic Limited Area Weather Forecasting with Diffusion, arXiv preprint arXiv:2502.07532, 2025.
- Li, J., Santek, D., Li, Z., Lim, A., Di, D., Min, M., Velden, C., and Menzel, W. P.: Tracking Atmospheric Motions for Obtaining Wind Estimates Using Satellite Observations: from 2D to 3D, *Bulletin of the American Meteorological Society*, 106, E344–E363, <https://doi.org/10.1175/BAMS-D-24-0027.1>, 2025.
- Ling, F., Lu, Z., Luo, J.-J., Bai, L., Behera, S. K., Jin, D., Pan, B., Jiang, H., and Yamagata, T.: Diffusion Model-Based Probabilistic Downscaling for 180-Year East Asian Climate Reconstruction, *npj Climate and Atmospheric Science*, 7, 111, <https://doi.org/10.1038/s41612-024-00679-1>, 2024.
- Loganathan, P., Zea, E., Vinuesa, R., et al.: Regional climate projections using a deep-learning-based model-ranking and downscaling framework: application to European climate zones, *Environmental Science and Pollution Research*, 32, 20133–20153, <https://doi.org/10.1007/s11356-025-36872-9>, 2025.
- Lopez-Gomez, I., Wan, Z. Y., Zepeda-Núñez, L., Schneider, T., Anderson, J., and Sha, F.: Dynamical-Generative Downscaling of Climate Model Ensembles, *Proceedings of the National Academy of Sciences*, 122, e2420288122, <https://doi.org/10.1073/pnas.2420288122>, 2025.
- Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Nabian, M. A., Ge, T., Subramaniam, A., Kashinath, K., Kautz, J., and Pritchard, M.: Residual Corrective Diffusion Modeling for km-Scale Atmospheric Downscaling, *Communications Earth & Environment*, 6, 124, <https://doi.org/10.1038/s43247-025-02042-5>, 2025.
- Qin, J., Pan, W., He, M., Lu, N., Yao, L., Jiang, H., and Zhou, C.: A Long-Term 1 km Monthly Near-Surface Air Temperature Dataset over the Tibetan Glaciers by Fusion of Station and Satellite Observations, *Earth System Science Data*, 15, 331–346, <https://doi.org/10.5194/essd-15-331-2023>, 2023.



- Semenov, M. A. and Barrow, E. M.: Use of a stochastic weather generator in the development of climate change scenarios, *Climatic Change*, 35, 397–414, <https://doi.org/10.1023/A:1005342632279>, 1997.
- Teed, Z. and Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 402–419, https://doi.org/10.1007/978-3-030-58536-5_24, 2020.
- Uusinoka, M., Haapala, J., and Polojärvi, A.: Deep Learning-Based Optical Flow in Fine-Scale Deformation Mapping of Sea Ice Dynamics, *Geophysical Research Letters*, 52, e2024GL112 000, <https://doi.org/10.1029/2024GL112000>, 2025.
- 400 Wang, B., Feng, P., Liu, D. L., et al.: Sources of uncertainty for wheat yield projections under future climate are site-specific, *Nature Food*, 1, 720–728, <https://doi.org/10.1038/s43016-020-00181-w>, 2020.
- Wang, L., Li, Q., Peng, X., and Lv, Q.: A Temporal Downscaling Model for Gridded Geophysical Data with Enhanced Residual U-Net, *Remote Sensing*, 16, 442, <https://doi.org/10.3390/rs16030442>, 2024.
- Wang, L., Li, Q., Lv, Q., Peng, X., and You, W.: TemDeep: a self-supervised framework for temporal downscaling of atmospheric fields at
405 arbitrary time resolutions, *Geoscientific Model Development*, 18, 2427–2442, <https://doi.org/10.5194/gmd-18-2427-2025>, 2025a.
- Wang, L., Li, Q., Zhou, Z., and Peng, X.: A diffusion framework for temperature fields reconstruction with embedded vertical dynamics, *Journal of Geophysical Research: Atmospheres*, 130, e2024JD042 742, <https://doi.org/10.1029/2024JD042742>, 2025b.
- Wang, Y., Karimi, H. A., and Jia, X.: Reconstruction of Continuous High-Resolution Sea Surface Temperature Data Using Time-Aware Implicit Neural Representation, *Remote Sensing*, 15, 5646, <https://doi.org/10.3390/rs15245646>, 2023.
- 410 Weyn, J. A., Durran, D. R., and Caruana, R.: Can Machines Learn to Predict Weather? Using Deep Learning to Predict Gridded 500-hPa Geopotential Height From Historical Weather Data, *Journal of Advances in Modeling Earth Systems*, 11, 2680–2693, <https://doi.org/10.1029/2019MS001705>, 2019.
- Wilks, D. S.: Adapting stochastic weather generation algorithms for climate change studies, *Climatic Change*, 22, 67–84, <https://doi.org/10.1007/BF00143344>, 1992.
- 415 Yang, Y., You, Q., Jin, Z., Zuo, Z., Zhang, Y., and Kang, S.: The Reconstruction for the Monthly Surface Air Temperature over the Tibetan Plateau during 1901–2020 by Deep Learning, *Atmospheric Research*, 285, 106 635, <https://doi.org/10.1016/j.atmosres.2023.106635>, 2023.
- Zhao, Y., Li, J., Chen, C., and Guan, Q.: A Diffusion–Attention–Enhanced Temporal (DATE-TM) Model: A Multi-Feature-Driven Model for Very-Short-Term Household Load Forecasting, *Energies*, 18, 486, <https://doi.org/10.3390/en18030486>, 2025.