LEARNING CONTINUOUS REGIONAL TEMPERATURE FIELDS WITH LEAD-TIME AND RESOLUTION QUERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate regional near-surface temperature forecasting is fundamental to short-range weather services and downstream risk assessment. Existing deep learning-based regional forecasters commonly produce a fixed set of future frames on a prescribed grid, limiting their use when forecast products must be evaluated at query-dependent lead times or display resolutions. To overcome these fixed-output constraints, we formulate regional T2M forecasting as query-conditioned continuous spatiotemporal temperature field evaluation and propose the **C**ontinuous **S**patiotemporal **T**emperature **F**orecaster (**CSTF**), a neural field that turns forecast lead time and output resolution into explicit queries when evaluating 2-m temperature (T2M). Specifically, CSTF first encodes multi-variable ERA5 histories into latent meteorological states and then decodes T2M as a coordinate-based field. Accordingly, spatial location, forecast lead time, and output resolution are introduced as queries, enabling standard hourly forecasts, intermediate lead-time diagnostics, and resolution-controllable outputs within a unified field-evaluation framework. Furthermore, to maintain coherence across flexible field queries, we design spatial-gradient, temporal-difference, and scale-consistency objectives that regularize regional thermal structures, lead-wise evolution, and cross-resolution agreement. Experiments on the Southeast China 0–6 h ERA5-Land benchmark demonstrate that CSTF achieves the best aggregate deterministic skill, including a 17.0% reduction in Bias, with global-scope diagnostics further illustrating flexible lead-time and resolution-controllable inference.

1 INTRODUCTION

Accurate regional near-surface temperature forecasting supports weather services and decisions in electricity demand management, agricultural planning, public health warning, and urban heat-risk assessment (Mukkavilli et al., 2023; Zhou et al., 2026). Near-surface temperature evolves continuously under advection, radiation, cloud cover, land–sea contrast, and surface heterogeneity, so local gradients can be as important as domain-mean error for regional use (Arjdal et al., 2026; Chajaei & Bagheri, 2024; Liu et al., 2024c). In practice, downstream applications may require forecasts at intermediate lead times or on different spatial grids. Fixed-output models address such requests through post-hoc interpolation or resampling, which neither treats the requested lead time and resolution as part of the forecasting process nor guarantees cross-resolution consistency. This motivates a query-conditioned formulation that preserves verified hourly forecasting while supporting diagnostic non-integer lead-time queries and resolution-controllable outputs.

Data-driven meteorological models have made rapid progress in learning structured atmospheric fields from large-scale gridded data (Bi et al., 2023; Lam et al., 2023; Chen et al., 2023). Yet most regional neural forecasters still expose a discrete product interface, either emitting a fixed stack of future frames or advancing the state autoregressively at a predefined time step (Bi et al., 2023; Lam et al., 2023; Chen et al., 2023). Such designs are natural for benchmark evaluation, but they make lead time and output resolution architectural choices rather than query variables. A fixed multi-head or sequence-to-sequence model can return the frames it was built to produce, while a request such as $t + 0.5$ h at 192×192 resolution lies outside its native output space (Fig. 1a).

Autoregressive rollout can extend the forecast horizon, but it does not by itself provide dense-time field evaluation and may accumulate errors as predictions are fed back into the model (Fig. 1b)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

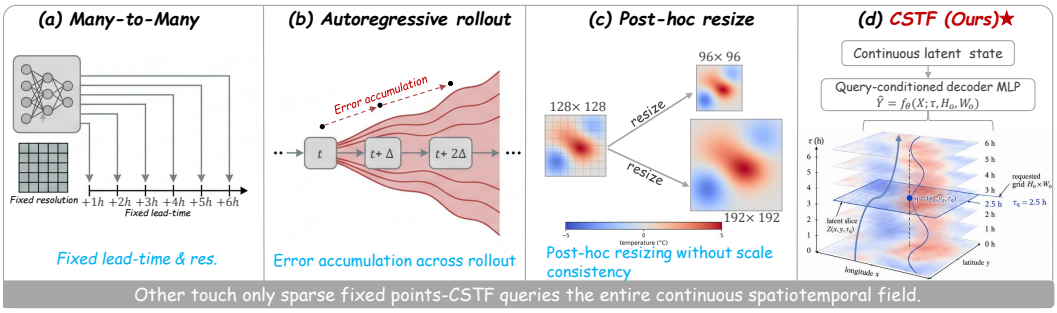

Figure 1: Comparison between conventional fixed-output and proposed query-conditioned temperature forecasting paradigms. (a) Fixed multi-head models predict predefined lead times and output resolutions. (b) Autoregressive rollout extends forecasts step by step but can accumulate errors across forecast steps. (c) Post-hoc resizing changes the display grid without enforcing consistency across resolutions. (d) Our CSTF formulates regional T2M prediction as continuous field evaluation conditioned on spatial, lead-time, and resolution queries, enabling flexible forecast requests.

(Lam et al., 2023; Tian et al., 2025). Likewise, post-hoc resizing or super-resolution can change the apparent grid after a forecast is generated, but it does not ensure that products sampled at different resolutions describe the same underlying temperature state (Fig. 1c) (Sun et al., 2024; Hess et al., 2024; Liu et al., 2024b; Gao et al., 2023). In all these cases, the available lead times and output grid are fixed by the training target or by the final prediction head. Increasing the number of supervised frames or training separate models for different resolutions can enlarge the product catalogue, but the model still predicts among predefined outputs rather than evaluating a requested field.

This limitation motivates the central question of this work: can regional temperature forecasting be formulated as continuous spatiotemporal field evaluation (Verma et al., 2024; Seong et al., 2026)? Instead of asking a network to emit a finite list of images, we seek a single field evaluator that maps the same encoded atmospheric state to T2M values at requested spatial coordinates, lead times, and output grids. Under this view, time and resolution are part of the model input, and cross-resolution consistency becomes a property that can be trained and diagnosed explicitly.

To this end, we propose the **C**ontinuous **S**patiotemporal **T**emperature **F**orecaster (**CSTF**), a query-conditioned neural field for regional 2-m temperature prediction (Fig. 1d). CSTF first encodes multivariable ERA5 histories into a shared latent meteorological state, so forecast products at different times and grids are derived from the same atmospheric context. A coordinate decoder then evaluates T2M fields from explicit spatial-coordinate, lead-time, and resolution queries, enabling hourly forecasts, intermediate lead-time queries, and resolution-controllable outputs within one field-evaluation framework. Unlike fixed prediction heads or separate scale-specific models, this design makes the requested lead time and grid part of the prediction process itself. To make these flexible queries meteorologically coherent rather than merely interpolated products, we regularize regional thermal gradients, lead-wise temperature changes, and cross-resolution agreement during training. We evaluate CSTF on a Southeast China 0–6 h T2M forecasting benchmark and further use global-scope diagnostics to examine the same query interface beyond the regional test domain. Our contributions are summarized as follows:

- We reformulate regional T2M forecasting as query-conditioned continuous spatiotemporal field evaluation, where lead time and output resolution are treated as inputs rather than fixed output dimensions.
- We propose CSTF, a shared-latent-state and coordinate-decoding architecture that directly evaluates regional temperature fields at requested spatial locations, forecast lead times, and output grids.
- We develop coherence-preserving supervision and diagnostics for query-conditioned temperature fields, validating fixed-lead skill, continuous lead-time queryability, resolution control, and cross-resolution consistency.

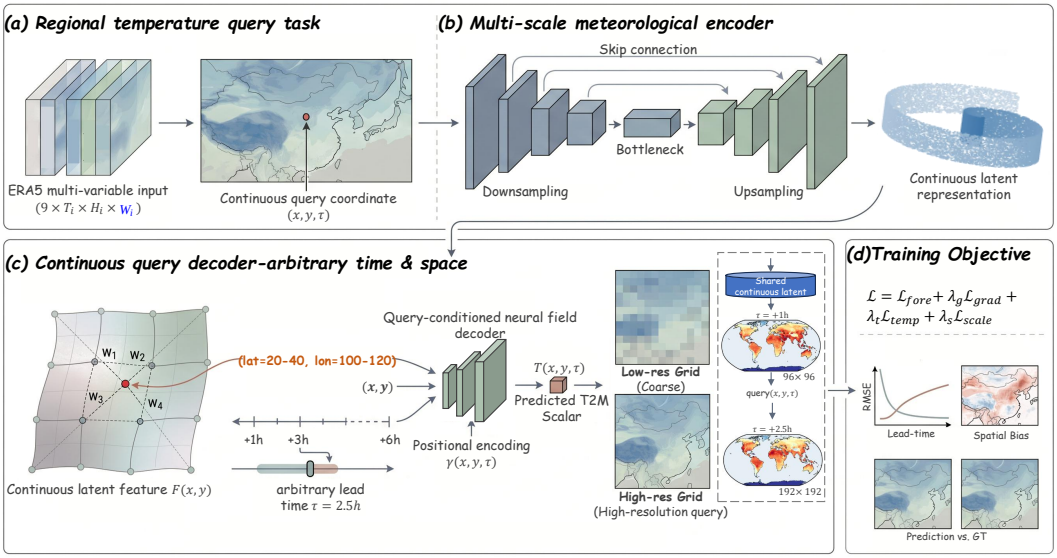

Figure 2: Overall architecture of the proposed CSTF framework. The model encodes ERA5 multi-variable input history into a latent regional weather state within a unified query-conditioned forecasting pipeline. Conditioned on this latent state, a coordinate decoder predicts regional T2M fields from lead-time, spatial-coordinate, and output-resolution queries. Training combines deterministic, gradient, temporal-difference, and scale-consistency objectives, so the same trained model supports continuous lead-time queries, resolution-controllable outputs, and scale-consistency evaluation.

2 Related Work

Data-driven weather prediction. Machine learning is now widely used for weather and Earth-system prediction. WeatherBench established a benchmark for data-driven global forecasting (Rasp et al., 2020), while FourCastNet, Pangu-Weather, GraphCast, and FengWu demonstrated skilful high-dimensional forecasts at low inference cost (Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023; Chen et al., 2023). At shorter ranges and regional scales, EarthFormer, wavelet-driven radar retrieval, and attention-based convolutional models capture useful evolution patterns from multi-channel meteorological fields (Gao et al., 2022; Shi et al., 2026). However, most systems still use a fixed-output interface: input history is mapped to predefined variables, lead times, and grids. Our work addresses this limitation for regional T2M prediction by making lead time and output resolution query variables.

Regional temperature prediction. Regional near-surface temperature prediction is a structured short-range refinement problem governed by synoptic forcing, advection, radiation, cloud cover, land–sea contrast, topography, and surface heterogeneity (Arjdal et al., 2026; Chajaei & Bagheri, 2024). ERA5 provides dynamically coherent atmospheric context, whereas ERA5-Land resolves regional thermal gradients relevant to local decisions (Hersbach et al., 2020; Wang et al., 2026). Convolutional encoder–decoders, transformer forecasters, and super-resolution models map coarse or multi-variable inputs to regional targets, but their lead times and grids are usually fixed during training (Gao et al., 2022; Sun et al., 2024; Hess et al., 2024; Liu et al., 2024c). Post-hoc interpolation changes the display grid without ensuring cross-resolution agreement for the same temperature state (Liu et al., 2024b; Gao et al., 2023); CSTF instead keeps supervised ERA5-Land verification while adding inference-time control over lead time and resolution.

Continuous and coordinate-based modelling. Recent weather-AI models relax fixed temporal interfaces by treating evolution as continuous or adapting rollout intervals (Verma et al., 2024; Liu et al., 2024a; Tian et al., 2025). Temporal interpolation and temporal super-resolution have also been explored for radar precipitation and atmospheric fields (Tatsubori et al., 2022; Demiray et al., 2023; Wang et al., 2025). Coordinate-based neural fields and operator-learning methods further represent signals or solution maps as functions evaluable at requested locations or discretizations (Li et al.,

2020; Guibas et al., 2021). Together, these ideas motivate query-based meteorological prediction, yet temporal flexibility can leave the grid fixed, and coordinate flexibility alone does not ensure cross-resolution coherence. CSTF couples lead-time, coordinate, and resolution conditioning for regional T2M forecasting, using scale-consistency training and diagnostics to separate controllable inference from post-processing.

3 METHOD

3.1 PROBLEM DEFINITION

Given a regional domain Ω discretized on a base grid of size $H \times W$, each forecast initialization provides a multi-variable ERA5 history

$$\mathbf{X} \in \mathbb{R}^{B \times T_{\text{in}} \times C_{\text{in}} \times H \times W}, \tag{1}$$

where B is the batch size, T_{in} denotes the number of historical frames, and $C_{\text{in}} = 9$ corresponds to

$$\{\text{t2m}, \text{d2m}, \text{u10}, \text{v10}, \text{msl}, \text{cp}, \text{lsp}, \text{tcc}, \text{ssrd}\}. \tag{2}$$

The supervised target is the future ERA5-Land T2M sequence $\mathbf{Y} \in \mathbb{R}^{B \times T_{\text{out}} \times 1 \times H \times W}$, defined at hourly lead times for training and standard evaluation.

The standard supervised formulation learns a fixed-output sequence mapping

$$\mathbf{X} \mapsto \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{T_{\text{out}}}\}, \tag{3}$$

where the available lead times and output grid are determined by the target sequence and prediction head. Such a predictor can only return members of its predefined product set, and therefore cannot natively evaluate the temperature field at an intermediate lead time or an alternative grid. To expose lead time and output resolution as query variables, we instead formulate regional near-surface temperature forecasting as query-conditioned spatiotemporal field evaluation:

$$\widehat{\mathbf{Y}}_{\tau}^{(H_o, W_o)} = \mathcal{F}_{\theta}(\mathbf{X}; \tau, H_o, W_o)\,, \tag{4}$$

where τ denotes the requested lead time and (H_o, W_o) denotes the requested output grid. The objective is to learn a single function $\mathcal{F}_{\theta}$ that remains accurate at supervised integer leads while supporting diagnostic queries at non-integer lead times and multiple output resolutions.

3.2 OVERVIEW

To implement this field-evaluation formulation, merely appending lead-time and resolution inputs to a fixed-grid forecaster is insufficient. If forecast products at different horizons or grids are generated by independent heads, rollout steps, or post-processing, their differences need not correspond to evaluations of a common thermal evolution. Moreover, simply resizing a prediction can change its display grid without informing the model of the requested sampling density, and pointwise supervision at hourly base-grid targets does not constrain non-integer lead queries or cross-resolution agreement. Our key insight is that query flexibility must be coupled with state sharing and consistency regularization. Accordingly, CSTF uses three coordinated design choices: a shared latent meteorological state, an explicit query-conditioned coordinate decoder, and coherence-preserving objectives for space, time, and scale. As illustrated in Fig. 2, the framework follows the mapping $\mathbf{X} \xrightarrow{\mathcal{E}_{\theta}} \mathbf{Z} \xrightarrow{\mathcal{D}_{\theta}(\tau, \mathbf{P}_{H_o,W_o}, \mathbf{R}_{H_o,W_o})} \widehat{\mathbf{Y}}_{\tau}^{(H_o,W_o)}$. Specifically, a multi-scale meteorological encoder transforms the multi-variable ERA5 history $\mathbf{X}$ into a latent regional weather state $\mathbf{Z}$. The requested lead time τ, output coordinates $\mathbf{P}_{H_o,W_o}$, and grid resolution $\mathbf{R}_{H_o,W_o}$ then condition a coordinate decoder, so each forecast product is obtained by evaluating the same state under a specified query. Training combines deterministic, spatial-gradient, temporal-difference, and scale-consistency objectives to anchor supervised accuracy while regularizing regional thermal contrast, lead-wise evolution, and cross-resolution agreement.

3.3 QUERY-CONDITIONED FIELD MODEL

We formalize CSTF as a neural field conditioned on both the encoded atmospheric state and the requested forecast product. This shared-state formulation prevents the query interface from degenerating into separate product-specific mappings. The encoder first summarizes the input history into

a shared latent regional state, after which the decoder evaluates this state at each requested lead time and grid cell:

$$\begin{aligned} \mathbf{Z} &= \mathcal{E}_\theta(\mathbf{X}) \,, \\ \widehat{\mathbf{Y}}_\tau^{(H_o,W_o)}(i,j) &= \mathcal{D}_\theta(\mathbf{Z}_{H_o,W_o}(i,j), \boldsymbol{\phi}(\tau), \mathbf{p}_{i,j}, \mathbf{r}_{H_o,W_o}) \,, \end{aligned} \tag{5}$$

where $\mathbf{Z}$ denotes the latent regional weather state, $\mathbf{Z}_{H_o,W_o}$ is its interpolation on the requested grid, $\boldsymbol{\phi}(\tau)$ is a sinusoidal lead-time embedding, $\mathbf{p}_{i,j} \in [-1,1]^2$ is the normalized coordinate of grid cell (i,j), and $\mathbf{r}_{H_o,W_o} = (1/H_o, 1/W_o)$ encodes the requested sampling density. This formulation makes the output grid an explicit part of the decoder input, enabling different spatial discretizations to be treated as direct evaluations of the same atmospheric state.

3.4 Multi-scale meteorological encoder

The encoder serves as the state-construction component of CSTF, rather than a product-specific prediction head. Because this state is later queried across lead times and resolutions, it must preserve regional thermal gradients while aggregating broader wind, pressure, moisture, cloud, and radiation cues from the multi-variable ERA5 history. We instantiate this state encoder as an encoder–decoder backbone with skip aggregation, leaving $\mathbf{Z}$ independent of any fixed lead time or output resolution and available for query-specific field evaluation by the coordinate decoder.

3.5 Lead-time and resolution queries

A forecast query in CSTF specifies when the temperature field is evaluated, where it is sampled, and at what output density it is requested. For the temporal component, CSTF normalizes the requested lead τ by the maximum training horizon and maps it to a learnable temporal code:

$$\boldsymbol{\phi}(\tau) = \mathrm{MLP}\left(\mathrm{SinCos}\left(\frac{\tau}{\tau_{\max}}\right)\right). \tag{6}$$

The sinusoidal basis provides a smooth temporal representation, and the MLP adapts it to short-range temperature evolution. The same embedding function is used for supervised integer leads and non-integer diagnostic queries such as 0.5 h or 4.5 h.

For the spatial component, coordinates alone identify where to evaluate the field, but they do not indicate whether the field is being sampled sparsely or densely. We therefore encode the requested grid by both sampling locations and sampling density. The first field is the normalized coordinate grid

$$\mathbf{P}_{H_o,W_o} = \big\{\mathbf{p}_{i,j} = (x_j, y_i) : x_j, y_i \in [-1,1]\big\}, \tag{7}$$

which specifies the location of each output cell within the regional domain. The resolution field is constant over the output image:

$$\mathbf{R}_{H_o,W_o}(i,j) = (1/H_o, 1/W_o) \,. \tag{8}$$

Together, $\boldsymbol{\phi}(\tau)$, $\mathbf{P}_{H_o,W_o}$, and $\mathbf{R}_{H_o,W_o}$ define the requested forecast product passed to the decoder.

3.6 Coordinate decoder

The coordinate decoder evaluates the shared weather state under the requested product query. A conventional convolutional prediction head maps latent features to a fixed grid and is unaware of absolute sampling location or requested grid density, making resolution changes indistinguishable from post-hoc resizing. We instead interpolate $\mathbf{Z}$ to the requested grid and condition the decoder on local latent features, spatial coordinates, resolution code, and lead-time embedding:

$$\widehat{\mathbf{Y}}_\tau^{(H_o,W_o)} = \mathcal{D}_\theta[\mathbf{Z}_{H_o,W_o}, \mathbf{P}_{H_o,W_o}, \mathbf{R}_{H_o,W_o}, \boldsymbol{\phi}(\tau)] \,. \tag{9}$$

In implementation, $\mathcal{D}_\theta$ is a lightweight convolutional decoder applied on the requested grid. Because query channels are part of the decoder input, each predicted value depends on local meteorological features, spatial position, forecast lead time, and sampling scale. Thus, outputs such as 96×96, 128×128, and 192×192 fields are generated by direct field evaluation rather than by resizing a fixed-grid forecast.

3.7 TRAINING OBJECTIVE

Training uses hourly lead times $\{\tau_k\}_{k=1}^{T_{\text{out}}}$ with available ERA5-Land references. The supervised forecast loss anchors CSTF to observed ERA5-Land targets:

$$\mathcal{L}_{\text{forecast}} = \frac{1}{T_{\text{out}}} \sum_{k=1}^{T_{\text{out}}} \text{MSE}\Big(\widehat{\mathbf{Y}}_{\tau_k}^{(H,W)}, \mathbf{Y}_{\tau_k}^{(H,W)}\Big). \tag{10}$$

However, pointwise hourly supervision alone does not guarantee coherent behavior when the same field evaluator is queried across resolutions or adjacent lead times. We therefore add three regularizers for scale agreement, regional thermal structure, and lead-wise evolution. The scale-consistency loss matches a direct low-resolution query with an area-downsampled base-grid query, encouraging different resolution evaluations of the same atmospheric state to agree after aggregation:

$$\begin{aligned}\mathcal{L}_{\text{scale}} = \frac{1}{T_{\text{out}}} \sum_{k=1}^{T_{\text{out}}} \text{MSE}\Big(&\mathcal{F}_\theta(\mathbf{X}; \tau_k, H/2, W/2)\,, \\ &\text{AreaDown}\left(\mathcal{F}_\theta(\mathbf{X}; \tau_k, H, W)\right)\Big),\end{aligned} \tag{11}$$

where AreaDown denotes area downsampling from (H, W) to $(H/2, W/2)$.

The spatial-gradient loss discourages oversmoothing by matching finite-difference T2M gradients:

$$\begin{aligned}\mathcal{L}_{\text{grad}} = \frac{1}{T_{\text{out}}} \sum_{k=1}^{T_{\text{out}}} \Big(&\left\| \nabla_x \widehat{\mathbf{Y}}_{\tau_k} - \nabla_x \mathbf{Y}_{\tau_k} \right\|_1 \\ &+ \left\| \nabla_y \widehat{\mathbf{Y}}_{\tau_k} - \nabla_y \mathbf{Y}_{\tau_k} \right\|_1 \Big).\end{aligned} \tag{12}$$

The temporal-difference loss constrains short-range evolution by aligning frame-to-frame temperature changes:

$$\begin{aligned}\mathcal{L}_{\text{temp}} = \frac{1}{T_{\text{out}} - 1} \sum_{k=1}^{T_{\text{out}}-1} \Big\| &\Big(\widehat{\mathbf{Y}}_{\tau_{k+1}} - \widehat{\mathbf{Y}}_{\tau_k}\Big) \\ &- \big(\mathbf{Y}_{\tau_{k+1}} - \mathbf{Y}_{\tau_k}\big) \Big\|_1.\end{aligned} \tag{13}$$

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{forecast}} + \lambda_g \mathcal{L}_{\text{grad}} + \lambda_t \mathcal{L}_{\text{temp}} + \lambda_s \mathcal{L}_{\text{scale}}. \tag{14}$$

Although the objective contains four terms, training remains a single-stage joint optimization with fixed weights. Both $\mathcal{L}_{\text{grad}}$ and $\mathcal{L}_{\text{temp}}$ reuse the supervised forecasts without introducing additional parameters or prediction heads, while only $\mathcal{L}_{\text{scale}}$ requires an auxiliary low-resolution decoder query.

3.8 INFERENCE AND DIAGNOSTIC QUERIES

After training, inference queries the same evaluator $\mathcal{F}_\theta$ with lead-time and grid-size arguments. Integer leads are verified against ERA5-Land using deterministic forecast metrics. Non-integer leads are reported as forecast-only diagnostics because the current hourly target split has no direct half-hour references. For resolution-controllable inference, predictions at multiple output sizes are downsampled to a common base grid to quantify scale consistency. This diagnostic does not replace verification against ERA5-Land; it tests whether the query-conditioned decoder returns mutually coherent fields when the same atmospheric state is sampled at different spatial resolutions.

4 EXPERIMENTS

Table 1: Lead-wise quantitative comparison for short-range T2M forecasting over Southeast China. RMSE and MAE are reported in degrees Celsius, where lower values indicate better accuracy. The best result in each column is highlighted in green, and the second-best result is underlined.

Model	t+6 h		t+5 h		t+4 h		t+3 h		t+2 h		t+1 h	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Baseline	8.196	3.741	7.992	3.378	7.782	2.968	7.583	2.529	7.419	2.108	7.309	1.779
SMAAT-UNet (Trebing, 2021)	1.989	1.425	1.785	1.270	1.581	1.141	1.347	1.000	1.106	0.840	1.000	0.765
EarthFormer (Gao et al., 2022)	2.784	1.254	2.668	1.154	2.600	1.064	2.590	0.990	2.600	0.939	2.615	0.902
ARROW (Tian et al., 2025)	1.625	1.131	1.444	1.015	1.246	0.892	1.051	0.768	**0.870**	**0.655**	**0.739**	**0.567**
FREUD (Schusterbauer et al., 2026)	1.609	1.112	1.446	1.008	1.284	0.908	1.126	0.812	0.970	0.720	0.853	0.647
Ours	**1.414**	**1.014**	**1.284**	**0.925**	**1.162**	**0.845**	**1.036**	**0.765**	0.909	0.685	0.823	0.631

Table 2: Overall integer-lead T2M forecast skill over Southeast China. Lower is better for RMSE, MAE, and bias magnitude, while higher is better for correlation.

Model	RMSE ↓	MAE ↓	Bias ↓	Corr. ↑
Baseline	7.720	2.751	1.247	0.576
SMAAT-UNet (Trebing, 2021)	1.510	1.074	0.329	0.987
EarthFormer (Gao et al., 2022)	2.644	1.051	0.598	0.963
ARROW (Tian et al., 2025)	1.203	0.838	0.389	0.992
FREUD (Schusterbauer et al., 2026)	1.243	0.868	0.446	0.992
Ours	**1.123**	**0.811**	**0.273**	**0.993**
Improv. vs. Second	+6.7%	+3.2%	+17.0%	+0.1%

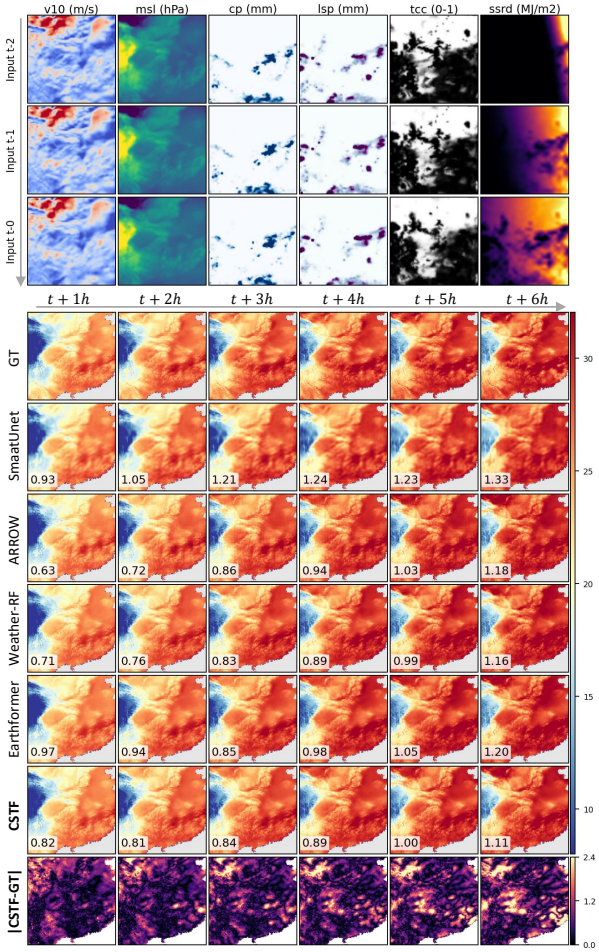

Figure 3: Regional deterministic T2M comparison over Southeast China, showing ERA5 input, CSTF forecasts, ERA5-Land references, and integer-lead absolute errors.

4.1 IMPLEMENTATION DETAILS

Experimental Setup. We evaluate CSTF on short-range 2-m temperature forecasting from ERA5 to ERA5-Land. Each sample uses a three-hour, nine-variable ERA5 history to predict six hourly future ERA5-Land T2M fields. The main benchmark is conducted over Southeast China at a base grid of 128×128, and a global-scope split is used only for qualitative query diagnostics. Detailed data splits, preprocessing, optimization, objective weights, checkpoint selection, and query-evaluation protocols are provided in Appendix.

Evaluation Metrics. Following meteorological-state evaluation protocols (Tu et al., 2025), we report MAE and RMSE in physical units for supervised integer leads. We also use bias and spatial correlation for regional temperature diagnostics, and cross-resolution scale-consistency error for resolution-controllable prediction. We define the reported scale-consistency error in normalized model space as

$$\mathcal{E}_{\text{scale}}(r) = \text{RMSE}\left(\text{Resample}^{\text{area}}_{r \to r_0}\left(\widehat{\mathbf{Y}}^{(r,r)}_{\tau}\right), \widehat{\mathbf{Y}}^{(r_0,r_0)}_{\tau}\right), \tag{15}$$

where $r_0 = 128$ is the base resolution and $\text{Resample}^{\text{area}}_{r \to r_0}$ maps each query to the base grid. This diagnostic compares predictions across query grids rather than measuring physical-unit error against ERA5-Land references.

Baseline Comparisons. We compare CSTF against the baseline and four learned models: SMAAT-UNet (Trebing, 2021), EarthFormer (Gao et al., 2022), ARROW (Tian et al., 2025), and FREUD

Table 3: Lead-wise ablation of CSTF objective terms. Southeast China RMSE and MAE are in degrees Celsius; lower is better, and Mean averages six leads.

Variant	t+6 h		t+5 h		t+4 h		t+3 h		t+2 h		t+1 h		Mean	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
w/o aux. losses	1.555	1.085	1.403	0.984	1.248	0.890	1.096	0.800	0.960	0.720	0.871	0.668	1.213	0.858
w/o $\mathcal{L}_{\text{scale}}$	1.575	1.096	1.422	0.996	1.263	0.898	1.100	0.802	0.949	0.712	0.852	0.654	1.220	0.860
w/o $\mathcal{L}_{\text{grad}}$	1.581	1.106	1.420	1.003	1.264	0.907	1.116	0.817	0.974	0.730	0.875	0.669	1.230	0.872
w/o $\mathcal{L}_{\text{temp}}$	1.476	1.043	1.337	0.948	1.203	0.861	1.059	0.773	0.921	0.689	0.831	0.636	1.160	0.825
Ours	**1.414**	**1.014**	**1.284**	**0.925**	**1.162**	**0.845**	**1.036**	**0.765**	**0.909**	**0.685**	**0.823**	**0.631**	**1.123**	**0.811**

(Schusterbauer et al., 2026). All learned baselines follow the same data, input–output, normalization, and evaluation protocol; controlled CSTF variants only remove selected objective terms.

4.2 Regional deterministic forecast

Before analyzing flexible query behavior, we first verify the deterministic forecasting skill of CSTF at the supervised hourly leads on the Southeast China benchmark.

Table 1 reports the lead-wise comparison. Although ARROW attains slightly lower errors at the first two leads, CSTF achieves the best skill from t+3 h to t+6 h, indicating more stable performance as short-range evolution becomes less constrained by the input state.

The aggregate results in Table 2 further show that CSTF obtains the lowest average RMSE, MAE, and bias, together with the highest spatial correlation among the compared methods. Relative to the strongest fixed-grid baseline, ARROW, CSTF reduces RMSE from $1.203°$C to $1.123°$C and MAE from $0.838°$C to $0.811°$C. The representative case in Fig. 3 shows that this improvement is accompanied by coherent regional thermal structures, with larger errors mainly confined to localized areas.

4.3 Global diagnostic visualization

To assess whether the query formulation preserves coherent large-scale thermal structures beyond the regional benchmark, we further apply CSTF to the global-scope split as a qualitative diagnostic. Figure 4 presents the input context, CSTF forecasts, ERA5-Land references, and absolute-error maps under a Robinson projection with polar insets. This visualization is used to inspect large-scale thermal coherence and spatial error patterns, while all quantitative comparisons are still computed on the native model grid.

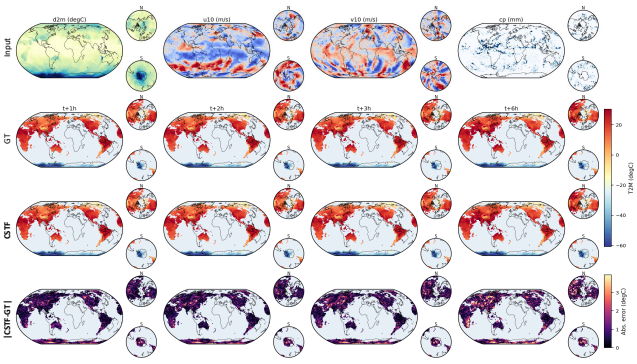

Figure 4: Global T2M diagnostic under a Robinson projection with polar insets, comparing ERA5-Land references, CSTF predictions, and absolute errors.

4.4 Continuous lead-time query

We test continuous-time queryability at verified integer and diagnostic non-integer leads, $\tau \in \{0.5, 2, 4.5, 6\}$ hours. Integer leads are compared with ERA5-Land targets; half-hour queries are forecast-only because the target split is hourly. Figures 5 and 6 show smoothly evolving regional and global temperature structures.

4.5 Resolution-controllable prediction

We assess spatial queryability on grids $(H_o, W_o) \in \{96^2, 128^2, 160^2, 192^2\}$ differing from the base resolution. Figures 7 and 8 show direct decoding at each requested grid, with consistent thermal patterns across resolutions.

Table 4: Scale-consistency ablation for resolution-controllable prediction. Normalized cross-resolution RMSE is reported after resampling to 128×128; lower is better.

Variant	$96 \to 128$	$160 \to 128$	$192 \to 128$	Mean
w/o aux. losses	1.396	1.165	1.233	1.265
w/o $\mathcal{L}_{\text{scale}}$	1.449	1.171	1.266	1.295
w/o $\mathcal{L}_{\text{grad}}$	1.111	<u>1.082</u>	0.997	<u>1.060</u>
w/o $\mathcal{L}_{\text{temp}}$	<u>1.099</u>	1.131	<u>0.988</u>	1.072
Ours	**1.098**	**1.077**	**0.982**	**1.052**

Table 5: Objective-term ablation over Southeast China. Physical forecast errors and normalized cross-resolution RMSE are reported; lower is better.

Variant	RMSE ↓	MAE ↓	Bias ↓	Grad. err. ↓	Scale err. ↓
w/o aux. losses	1.213	0.858	0.339	0.468	1.265
w/o $\mathcal{L}_{\text{scale}}$	1.220	0.860	0.350	0.382	1.295
w/o $\mathcal{L}_{\text{grad}}$	1.230	0.872	0.356	0.455	<u>1.060</u>
w/o $\mathcal{L}_{\text{temp}}$	<u>1.160</u>	<u>0.825</u>	<u>0.281</u>	<u>0.376</u>	1.072
Ours	**1.123**	**0.811**	**0.273**	**0.361**	**1.052**

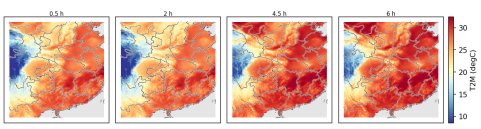

Figure 5: Continuous lead-time queries at 0.5, 2, 4.5, and 6 h; half-hour panels are forecast-only, while integer-hour panels have supervised targets.

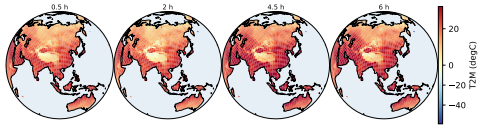

Figure 6: Global continuous lead-time query on a Robinson projection with polar insets. The checkpoint is queried at 0.5, 2, 4.5, and 6 hours.

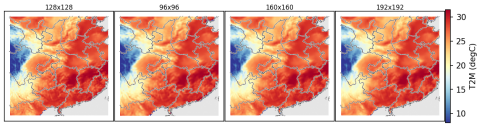

Figure 7: Resolution-conditioned decoding at 96^2, 128^2, 160^2, and 192^2 grids from a fixed input and lead time, without post-hoc resizing.

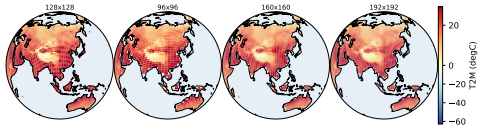

Figure 8: Global resolution-controllable prediction. Orthographic visualizations are generated for different output grid sizes from the same input state and lead-time query.

4.6 Scale-consistency analysis

Resolution-controllable visualization alone does not ensure compatible temperature states across grids. We therefore aggregate non-base queries to the reference grid and compare them with the base-grid prediction. The resulting diagnostic, reported in Table 4, isolates the resolution-query training signal under controlled ablations.

4.7 Ablation Studies

We evaluate four controlled variants: w/o aux. losses retains only $\mathcal{L}_{\text{forecast}}$; w/o $\mathcal{L}_{\text{scale}}$ removes only the scale-consistency term; w/o $\mathcal{L}_{\text{grad}}$ removes only the spatial-gradient term; and w/o $\mathcal{L}_{\text{temp}}$ removes only the temporal-difference term. As shown in Table 3, the full objective improves RMSE and MAE across all supervised leads, with larger gains at longer horizons, indicating more stable lead-wise temperature evolution. Table 4 shows that removing $\mathcal{L}_{\text{scale}}$ increases cross-resolution error after aggregation, supporting its role in coherent resolution-query prediction. Table 5 further indicates that the full objective provides balanced gains in aggregate accuracy, thermal-structure fidelity, and scale coherence.

5 Conclusion

We introduced CSTF, a query-conditioned forecaster that encodes multi-variable ERA5 histories into a shared latent state and decodes T2M fields at requested spatial coordinates, lead times, and resolutions. On the Southeast China benchmark, CSTF improves aggregate deterministic skill at supervised hourly leads, while additional diagnostics demonstrate non-integer lead-time queryability and resolution-controllable inference. The forecast and three auxiliary objectives jointly promote predictive accuracy, thermal-structure fidelity, and cross-resolution coherence through single-stage optimization. Because the targets are hourly, non-integer outputs should be interpreted as queryability diagnostics rather than quantitatively validated continuous-time forecasts.

AI USE STATEMENT

Generative AI tools were used to assist with methodological discussion, code development and debugging, literature search. They were also used to improve the clarity, grammar, and organization of the manuscript. All AI-assisted methodological suggestions were reviewed and validated by the authors, all generated or modified code was tested by the authors, and cited literature was checked against the original sources. Generative AI tools were not used to generate experimental data or replace the authors' scientific judgment. The authors take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

REFERENCES

Khadija Arjdal, Fatima Driouech, Saloua Balhane, and Rodrigo Manzanas. Compound heatwaves in north africa: Mechanisms and model capabilities. *Earth Systems and Environment*, pp. 1–25, 2026.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Fatemeh Chajaei and Hossein Bagheri. Machine learning framework for high-resolution air temperature downscaling using lidar-derived urban morphological features. *Urban Climate*, 57:102102, 2024.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023.

Bekir Z Demiray, Muhammed Sit, and Ibrahim Demir. Efficienttempnet: Temporal super-resolution of radar rainfall. *International Conference on Learning Representations (ICLR) 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023.

Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10021–10030, 2023.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Philipp Hess, Michael Aich, Baoxiang Pan, and Niklas Boers. Fast, scale-adaptive, and uncertainty-aware downscaling of earth system model fields with generative foundation models. *arXiv preprint arXiv:2403.02774*, 2024.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Peiyuan Liu, Tian Zhou, Liang Sun, and Rong Jin. Mitigating time discretization challenges with weatherode: A sandwich physics-driven neural ode for weather forecasting. *arXiv preprint arXiv:2410.06560*, 2024a.

Zili Liu, Hao Chen, Lei Bai, Wenyuan Li, Keyan Chen, Zhengyi Wang, Wanli Ouyang, Zhengxia Zou, and Zhenwei Shi. Deriving accurate surface meteorological states at arbitrary locations via observation-guided continous neural field modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 2024b.

Zili Liu, Hao Chen, Lei Bai, Wenyuan Li, Wanli Ouyang, Zhengxia Zou, and Zhenwei Shi. Mambads: Near-surface meteorological field downscaling with topography constrained selective state space modeling. *arXiv preprint arXiv:2408.10854*, 2024c.

S Karthik Mukkavilli, Daniel Salles Civitarese, Johannes Schmude, Johannes Jakubik, Anne Jones, Nam Nguyen, Christopher Phillips, Sujit Roy, Shraddha Singh, Campbell Watson, et al. Ai foundation models for weather and climate: Applications, design, and implementation. *arXiv preprint arXiv:2309.10808*, 2023.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

Johannes Schusterbauer, Jannik Wiese, Nick Stracke, Timy Phan, and Björn Ommer. Probabilistic precipitation nowcasting with rectified flow transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25742–25756, 2026.

Yeeun Seong, Doyi Kim, Minseok Seo, and Changick Kim. Rainode: Continuous-time precipitation forecasting with latent neural odes. *arXiv preprint arXiv:2606.29855*, 2026.

Chunlei Shi, Han Xu, Yinghao Li, Yi-Lin Wei, Yongchao Feng, Yecheng Zhang, and Dan Niu. Wavec2r: Wavelet-driven coarse-to-refined hierarchical learning for radar retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8951–8959, 2026.

Yongjian Sun, Kefeng Deng, Kaijun Ren, Jia Liu, Chongjiu Deng, and Yongjun Jin. Deep learning in statistical downscaling for deriving high spatial resolution gridded meteorological data: A systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:14–38, 2024.

Michiaki Tatsubori, Takao Moriyama, Tatsuya Ishikawa, Paolo Fraccaro, Anne Jones, Blair Edwards, Julian Kuehnert, and Sekou L Remy. Deep temporal interpolation of radar-based precipitation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1685–1689. IEEE, 2022.

Jindong Tian, Yifei Ding, Ronghui Xu, Hao Miao, Chenjuan Guo, and Bin Yang. Arrow: An adaptive rollout and routing method for global weather forecasting. *arXiv preprint arXiv:2510.09734*, 2025.

Kevin Trebing. Smaat-unet: Precipitation nowcasting using a small attention-unet architecture. *Pattern Recognition Letters*, 145:178–186, 2021.

Siwei Tu, Ben Fei, Weidong Yang, Fenghua Ling, Hao Chen, Zili Liu, Kun Chen, Hang Fan, Wanli Ouyang, and Lei Bai. Satellite observations guided diffusion model for accurate meteorological states at arbitrary resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28071–28080, 2025.

Yogesh Verma, Markus Heinonen, and Vikas Garg. Climode: Climate and weather forecasting with physics-informed neural odes. *arXiv preprint arXiv:2404.10024*, 2024.

Liwen Wang, Qian Li, Qi Lv, Xuan Peng, and Wei You. Temdeep: a self-supervised framework for temporal downscaling of atmospheric fields at arbitrary time resolutions. *Geoscientific Model Development*, 18(8):2427–2442, 2025.

Wenhong Wang, Shiao Feng, Yonggen Zhang, Zhongwang Wei, Jianzhi Dong, Lutz Weihermüller, Cong-Qiang Liu, and Harry Vereecken. Fusing era5-land and smap l4 for an improved global soil moisture product (1950–2025). *Earth System Science Data*, 18(2):1061, 2026.

Jun Zhou, Feng Chen, Yu Zhu, Fuming Xie, Caixia Qin, Hanqiang Tian, Yiyuan Shen, Xin Yang, Yunpeng Duan, Muhammad Mannan Afzal, et al. Reconstruction of temperature, precipitation, and identification of extreme climate events in high mountain asia over 500 years using multi-method enkf. *Scientific Reports*, 2026.

This supplementary material provides additional implementation details in Sec. A, extended temperature forecast case studies in Sec. B, and query, training, benchmark, and ablation diagnostics in Sec. C.

A Additional Implementation Details

Data split and preprocessing. The Southeast China (SE China) domain spans $100.0°$E–$120.0°$E and $20.0°$N–$40.0°$N. The Southeast China and global-scope splits both contain 4,615 training samples, 511 validation samples, and 714 test samples. The data are chronologically split over 2024–2025, with validation drawn from the training period using a 0.1 fraction and a held-out test subset used only for final evaluation. Unless otherwise stated, fields are resized to 128×128 for training and supervised integer-lead evaluation.

Training protocol. All reported CSTF models are trained for T2M prediction with a three-frame input history and a six-frame hourly forecast target. The main CSTF model is trained for 50 epochs with batch size 8, AdamW optimization, an initial learning rate of 10^{-3}, weight decay of 10^{-4}, cosine learning-rate annealing, gradient clipping at 1.0, and random seed 42. The final model is selected by validation loss and then evaluated on the held-out test split. For the full objective, the auxiliary loss weights are $\lambda_s = 0.05$, $\lambda_g = 0.05$, and $\lambda_t = 0.01$. Test metrics are computed with batch size 32 using the same normalization and evaluation script across CSTF and all fixed-grid baselines.

Query diagnostics. The standard benchmark verifies supervised integer-lead T2M forecasts against ERA5-Land. Continuous lead-time diagnostics use a single learned parameterization at both integer and non-integer lead times; non-integer leads are shown as forecast-only examples because the current hourly target split has no direct half-hour references. Resolution diagnostics evaluate the same field parameterization at 96×96, 128×128, 160×160, and 192×192 output grids and assess cross-resolution consistency after resampling to the base grid. The scale-consistency values in the main tables are reported as normalized RMSE diagnostics rather than physical-unit forecast errors.

Ablation setting. All controlled CSTF variants keep the architecture, data split, input variables, input length, output horizon, training resolution, and model-selection rule fixed. The ablations therefore isolate objective terms rather than changes in model capacity or input information.

Training objective rationale. The forecast MSE provides the main deterministic supervision on the hourly ERA5-Land targets and is computed over all supervised leads, channels, and grid cells in each training batch. The scale-consistency term is designed for the queryable-resolution interface: for the same atmospheric input and lead time, a direct low-resolution query should agree with an area-downsampled high-resolution query after both are represented on the same grid. In the implementation, the downsampled high-resolution branch is detached in this term, so the auxiliary gradient primarily aligns the direct low-resolution query with the high-resolution prediction used as the reference. The spatial-gradient term compares finite differences along the two image axes and discourages overly smooth temperature maps by penalizing mismatches in local thermal contrasts. The temporal-difference term compares adjacent-lead changes rather than only absolute values, encouraging the predicted sequence to follow the observed short-range warming or cooling tendency across the six supervised lead times. Together, these auxiliary objectives are intended to support the two query diagnostics used in the experiments: coherent multi-resolution products and smooth intermediate lead-time behavior.

B Additional Temperature Forecast Case Studies

This section provides additional deterministic T2M forecast visualizations before the query-specific diagnostics in Sec. C. The purpose is to separate two forms of evidence: first, whether the learned forecaster produces meteorologically coherent temperature fields at the supervised hourly leads, and second, whether the same model can later be interrogated through lead-time and resolution queries. Figures A and B therefore focus on standard deterministic forecast behavior rather than on flexible queryability.

Figure A extends the global-scope qualitative diagnostic beyond the main-text initialization. Because the global split is used only for diagnostic visualization, the figure should be interpreted as a

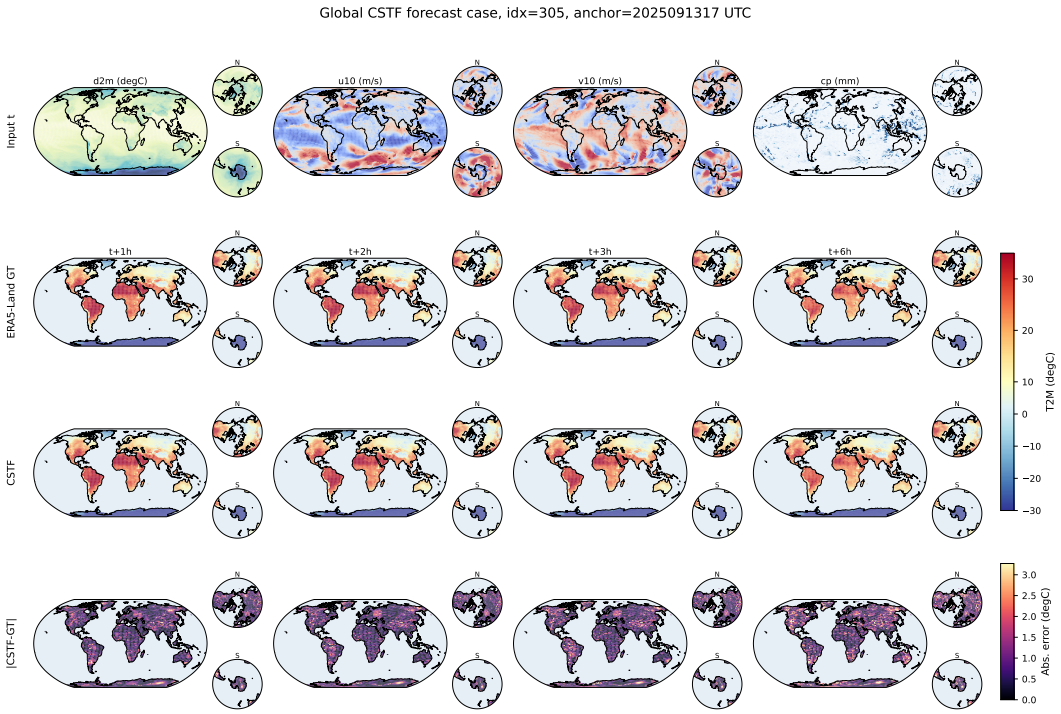

Figure A: Additional deterministic forecast diagnostic for the global domain (180.0°W–180.0°E, 90.0°S–90.0°N). The case shows input context, ERA5-Land references, CSTF predictions, and absolute-error maps under the main-text global visualization protocol. It provides an additional global forecast case and assesses whether the qualitative diagnostic remains consistent across forecast initializations.

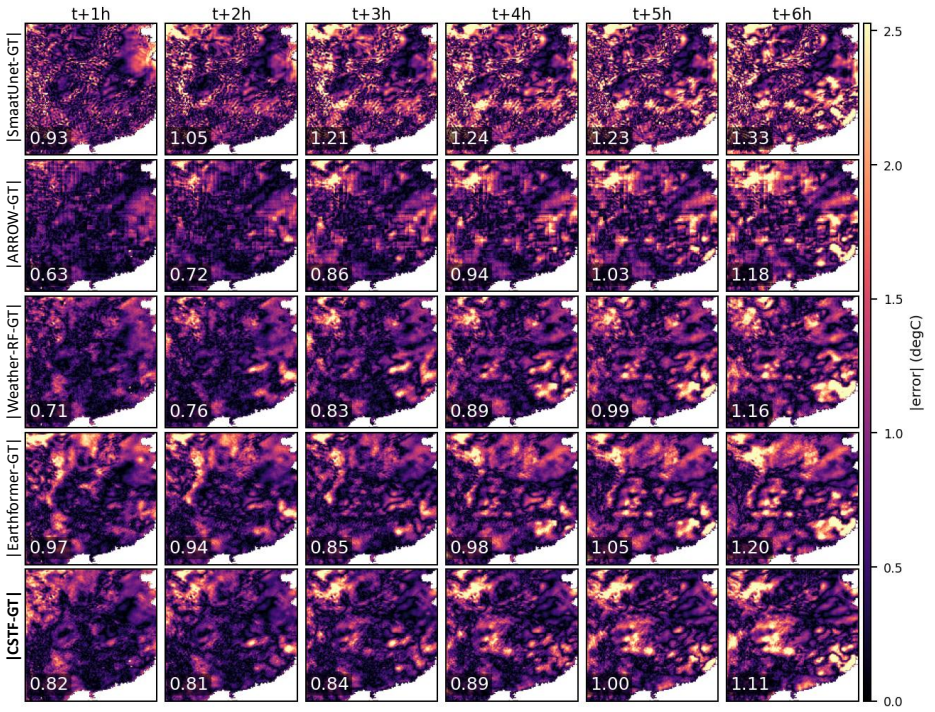

Figure B: Lead-wise absolute-error maps for the Southeast China case study (100.0°E–120.0°E, 20.0°N–40.0°N). Errors are shown over the six supervised integer lead times, providing a spatial diagnostic of where forecast discrepancies grow with lead time. The overlaid number on each panel reports the per-frame spatial RMSE (°C) against the ERA5-Land T2M reference over valid land pixels.

spatial plausibility check: it examines whether large-scale thermal patterns, land–sea contrasts, and error distributions remain coherent when the regional field-evaluation formulation is applied under a global rendering protocol. The inclusion of input context, ERA5-Land references, CSTF predictions, and absolute-error maps allows the reader to distinguish forecast structure from residual error patterns.

Figure B complements the global case by focusing on the supervised Southeast China benchmark used for quantitative evaluation. Instead of summarizing skill with a single aggregate score, the figure reports absolute-error maps at each integer lead time, making it possible to inspect where regional discrepancies emerge and how their spatial distribution changes across the 0–6 h forecast horizon. The number overlaid on each panel is the corresponding spatial RMSE in degrees Celsius, computed between the model prediction and the ERA5-Land T2M reference over valid land pixels. This view is particularly relevant for T2M forecasting because localized thermal contrasts can be smoothed or displaced even when average error remains low.

C Additional Query and Quantitative Diagnostics

This section expands the evidence for the query-conditioned field interface beyond the main-text examples. Figures C–G are organized to separate the two query modes and the two diagnostic domains. Specifically, Figs. C and D examine continuous lead-time and resolution-conditioned behavior over Southeast China (100.0°E–120.0°E, 20.0°N–40.0°N), whereas Figs. E and F provide the corresponding global-domain diagnostics (180.0°W–180.0°E, 90.0°S–90.0°N). Figure G then summarizes training, benchmark, scale-consistency, and ablation evidence to connect these qualitative diagnostics with the quantitative claims in the main text.

Figure C extends the regional continuous-time analysis from a single main-text example to three independent Southeast China forecast initializations. The purpose of this figure is to test whether

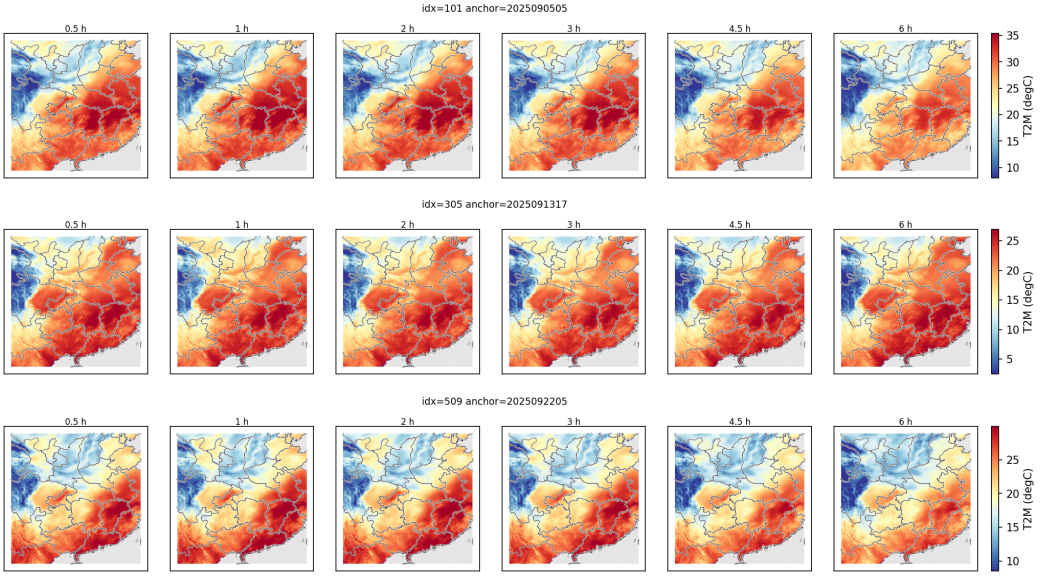

Figure C: Continuous lead-time query examples over Southeast China (100.0°E–120.0°E, 20.0°N–40.0°N). The three rows are independent test initializations at 2025-09-05 05:00 UTC, 2025-09-13 17:00 UTC, and 2025-09-22 05:00 UTC. Within each row, the same atmospheric input is queried at multiple lead times, including supervised hourly leads and non-integer lead times. The figure is therefore used to inspect whether the learned field evaluator produces temporally smooth and spatially coherent T2M fields when the requested forecast time is varied continuously. Non-integer lead panels are forecast-only diagnostics because the current ERA5-Land target sequence is available at hourly intervals.

varying the lead-time query produces physically plausible, spatially coherent temperature evolution rather than isolated artifacts at non-integer query times.

Figure D evaluates the regional resolution-query interface under the same Southeast China setting. By holding the atmospheric input and forecast case fixed while changing only the requested output grid, the figure checks whether CSTF behaves as a direct field evaluator instead of a fixed-grid predictor followed by post-hoc resizing.

Figure E transfers the continuous lead-time diagnostic to the global domain. This figure is not intended as a separate global benchmark; rather, it examines whether the same query-conditioned model produces coherent large-scale thermal structures when lead time is varied under a global diagnostic view.

Figure F provides the complementary global-domain resolution-query diagnostic. It is used to inspect whether large-scale temperature patterns remain stable across requested output grids, which supports the interpretation of resolution control as query-conditioned field evaluation rather than a visualization-only operation.

Figure G collects the quantitative diagnostics that support the preceding visual evidence. The panels cover optimization behavior, deterministic forecast skill, lead-wise degradation, cross-resolution consistency, and objective-term ablations, providing a compact check that the queryable behavior is accompanied by stable training and controlled test performance.

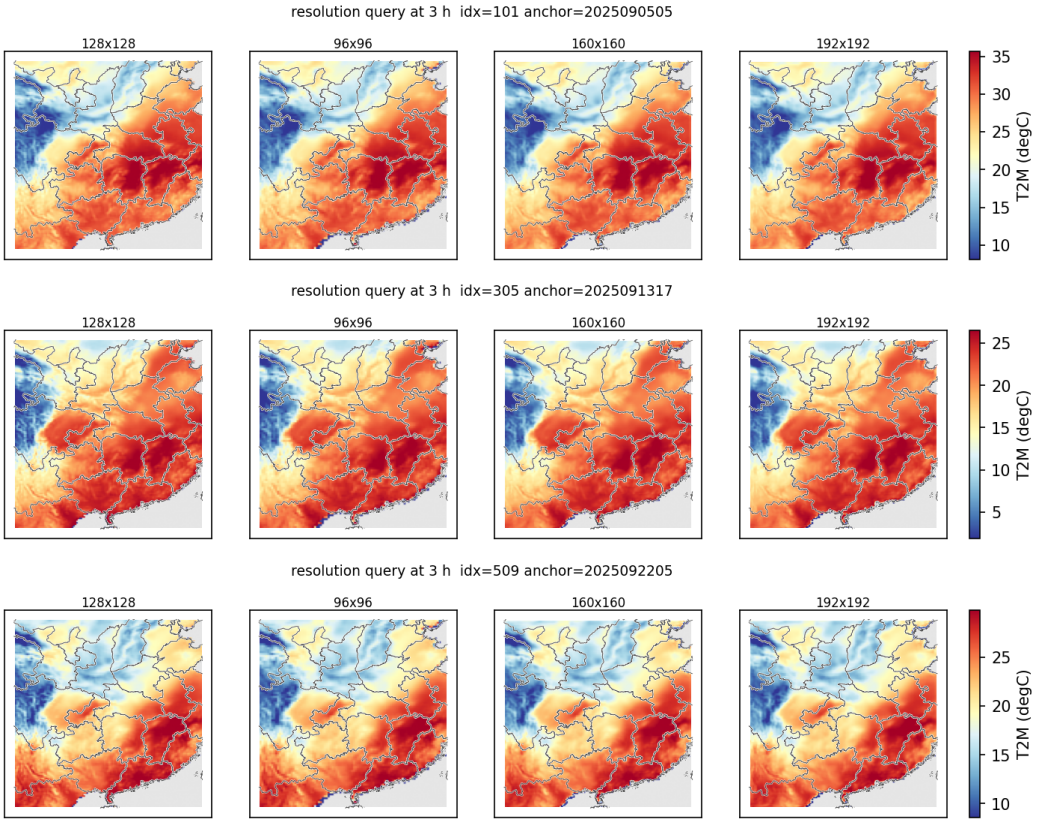

Figure D: Resolution-conditioned query examples over Southeast China (100.0°E–120.0°E, 20.0°N–40.0°N). The three rows are independent test initializations at 2025-09-05 05:00 UTC, 2025-09-13 17:00 UTC, and 2025-09-22 05:00 UTC. Within each row, CSTF directly evaluates the same forecast state at different requested output grids rather than resizing a single fixed-grid prediction. The visual comparison checks whether regional thermal structures remain consistent as the output resolution changes, complementing the cross-resolution RMSE values reported in the main text.

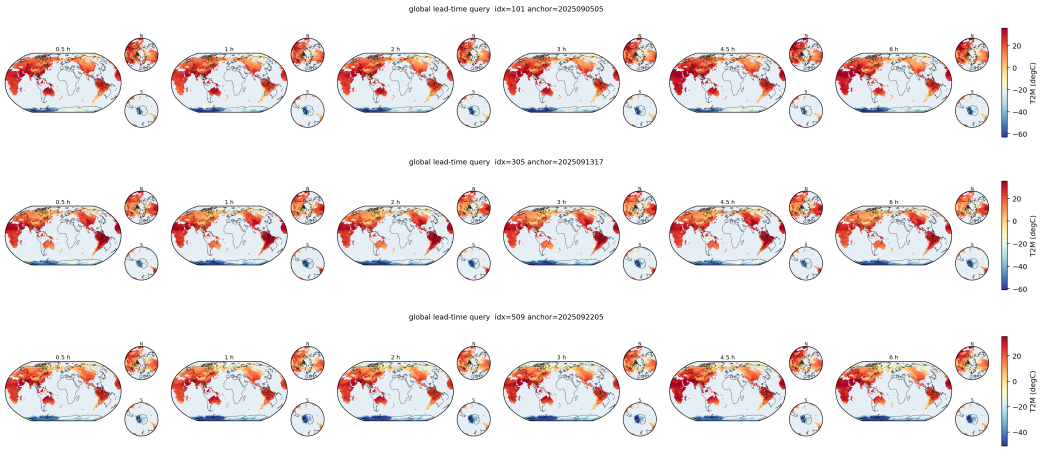

Figure E: Continuous lead-time query diagnostics over the global domain (180.0°W–180.0°E, 90.0°S–90.0°N). The three rows are independent test initializations at 2025-09-05 05:00 UTC, 2025-09-13 17:00 UTC, and 2025-09-22 05:00 UTC, rendered with the same globe-style projection as the main-text global diagnostic. Within each row, the queried lead time is varied while the input state is fixed, allowing visual inspection of large-scale thermal evolution under the continuous-time interface. Non-integer leads are forecast-only diagnostics because the current target split provides hourly references.

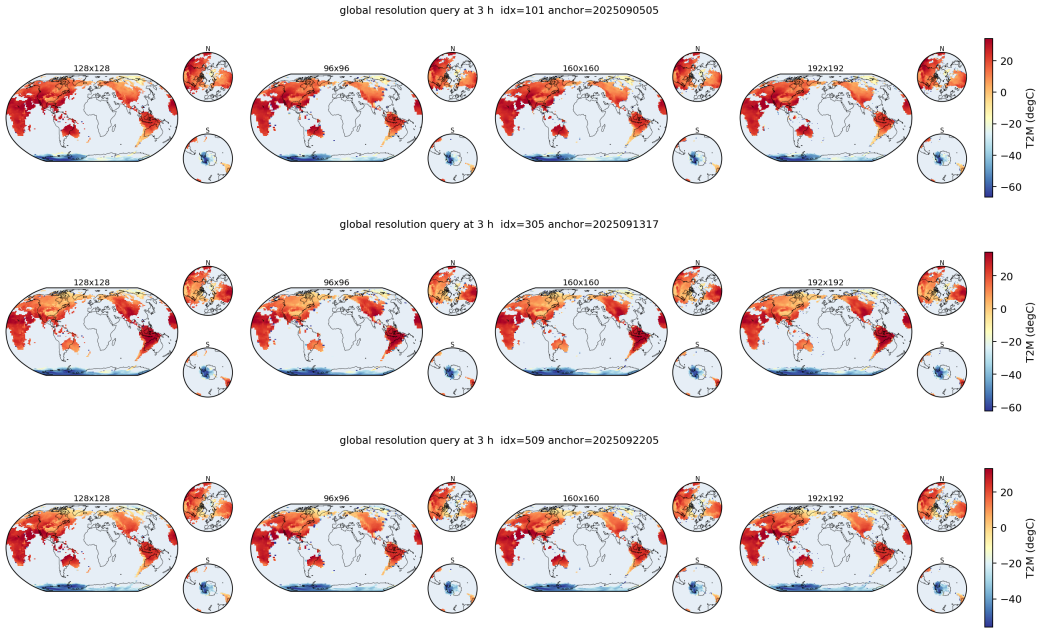

Figure F: Resolution-conditioned query diagnostics over the global domain (180.0°W–180.0°E, 90.0°S–90.0°N). The three rows are independent test initializations at 2025-09-05 05:00 UTC, 2025-09-13 17:00 UTC, and 2025-09-22 05:00 UTC. Within each row, the model is queried at multiple output grid sizes for the same forecast case, so changes across columns reflect direct resolution-conditioned field evaluation. These visualizations are used as qualitative checks of large-scale coherence under resolution queries, whereas quantitative cross-resolution consistency is reported separately in the main tables.

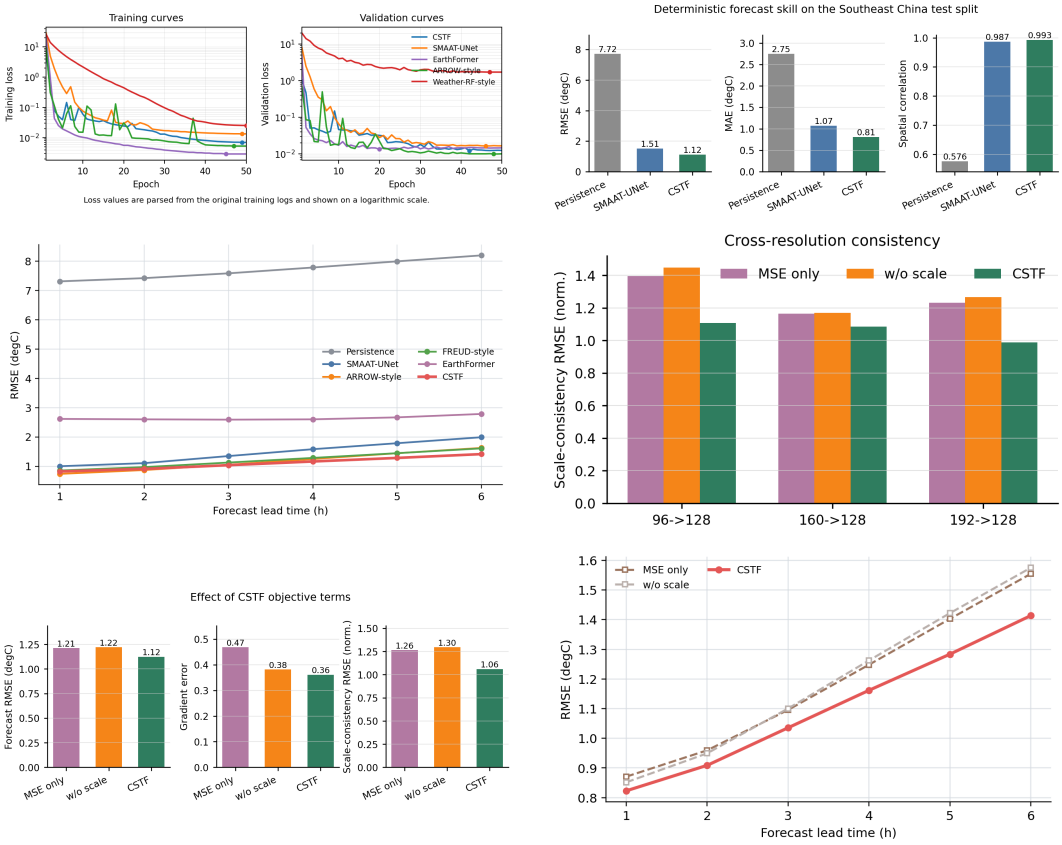

Figure G: Additional training, benchmark, and ablation diagnostics supporting the main quantitative results. From left to right and top to bottom, the panels show optimization curves, aggregate deterministic skill, lead-wise RMSE, cross-resolution scale consistency, objective-term ablation metrics, and lead-wise ablation behavior. The first panel checks whether model training is stable, the next two panels summarize forecast accuracy on the supervised hourly benchmark, the scale-consistency panel evaluates the resolution-query interface, and the ablation panels isolate the effects of the auxiliary objective terms. These plots are diagnostic summaries; the standardized test metrics and primary comparisons are reported in the main-text tables.