Textual Environmental Context and Spatial Graphs for LLM-Based Regional SST Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Sea surface temperature (SST) forecasting depends on local temporal persistence, regional spatial dependence, and environmental conditions that evolve with the forecast date. We study how these heterogeneous conditions can be presented to a large language model (LLM) for regional multi-step forecasting without serializing the full SST grid as text. We formulate forecasting as conditional numerical generation: historical SST and anomaly sequences, date-aligned environmental records, and static ocean knowledge form a textual context, while regional spatial state is supplied through continuous graph-derived prefixes. A static graph encodes persistent geographic–climatological relations, and a dynamic graph encodes recent SST correlations and localized tropical-cyclone influence. Two graph neural networks produce a target-node representation that is mapped by a spatial-prefix fusion and injected into the LLM input. On SST forecasting in the South China Sea, the complete configuration achieves the best MAE and R^2 among the compared methods over ten forecast steps. Alongside the numerical forecast, a rule-based module matches predicted trends and environmental-factor directions with knowledge entries to return source-linked, post-hoc contextual explanations.

1 Introduction

Sea surface temperature (SST) forecasting supports marine environmental monitoring and related applications (Liu et al., 2024a; Cui et al., 2025). SST evolution reflects local temporal persistence, spatial dependence among ocean cells, and environmental conditions such as tropical cyclones, large-scale climate states, and regional circulation (England et al., 2025). A forecasting model that uses only the historical SST sequence of one target cell must infer these heterogeneous influences indirectly. This motivates a central question: how can date-aligned environmental context and regional spatial state be presented jointly to a large language model (LLM) for multi-step SST forecasting?

Recent work has explored the use of LLMs for numerical and text-conditioned forecasting (Zhong et al., 2025; Wu et al., 2026). From News to Forecast combines selected news, auxiliary information, and historical sequences in textual prompts for numerical generation (Wang et al., 2024). In the SST domain, OKG-LLM aligns ocean-domain knowledge representations with SST observations through an LLM-based prediction pipeline (Yang et al., 2026). These studies establish that LLMs can incorporate contextual information and domain knowledge into numerical forecasting. However, they do not directly address how forecast-date environmental descriptions and explicit, date-varying regional spatial relations should be jointly supplied to an LLM for regional multi-step SST forecasting.

We address this problem by assigning textual and spatial conditions distinct roles. Related studies have paired auxiliary context with numerical histories, aligned time-series representations with language, and adapted models to heterogeneous covariates (Liu et al., 2024b; Hu et al., 2025; Han et al., 2026). Tropical-cyclone records are organized as descriptions of time, location, and intensity; the ENSO index and regional circulation provide climate and atmospheric context; and static ocean knowledge describes mixing, advection, upwelling, and air–sea exchange. These sources form a textual forecasting context together with the historical SST and anomaly sequences of the target cell. Text, however, does not itself perform neighborhood feature aggregation over the SST field: cells with similar environmental descriptions may have different local neighborhood states or storm influences. We therefore represent persistent and date-varying spatial relations separately. A static graph

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.

encodes geographic proximity and climatological features, whereas a dynamic graph encodes recent SST correlations and localized tropical-cyclone influence.

The two GNN branches aggregate regional information and produce a representation for the target cell. Rather than serializing all grid cells as text, a factorized spatial-prefix projector maps this representation into a compact sequence of continuous prefix tokens, which is concatenated with the LLM input embeddings for autoregressive SST generation. Alongside the numerical forecast, a separate rule-based module matches predicted trends and environmental-factor directions with knowledge entries to return source-linked post-hoc contextual explanations.

The main contributions are as follows:

- **A time-aligned text–graph interface for regional SST forecasting.** We formulate multi-step SST forecasting as conditional numerical generation, where environmental records and ocean knowledge are represented as text and regional spatial state is supplied through continuous graph-derived prefixes.
- **Date-specific dual-graph spatial conditioning.** We combine a static geographic–climatological graph with a dynamic graph constructed from recent SST correlations and localized tropical-cyclone influence, then inject the target-node state into the LLM without increasing the input length with the number of grid cells.
- **Lead-wise empirical evaluation of text and spatial conditions.** On South China Sea SST forecasting, we report lead-wise comparisons and ablations. Among the compared baseline methods, the full configuration achieves the best performance on average.

2 Related Work

2.1 Spatiotemporal SST forecasting

Recent SST forecasting studies have improved spatiotemporal modeling through spatial-relation learning, motion features, and physical priors. ASTGRN learns adaptive node representations and graph structures for sea-area dependencies, whereas DDSTGNN separates diffusion and intrinsic signals and learns evolving spatiotemporal dependencies with a dynamic graph (Zhao et al., 2024; Jin et al., 2025). Peng *et al.* combine optical-flow guidance with SwinLSTM features to extract motion information from SST fields (Peng et al., 2025). These methods provide different structural inductive biases for regional variation.

Physics-guided methods combine data-driven representations with ocean-process priors. DP-BICNN couples physical and data-driven processes through bidirectional information compensation, while PCFNet combines frequency-domain features with an advection–diffusion constraint (Liu et al., 2025b; 2026b). PG-DSBN combines numerical simulation priors with deterministic–stochastic modeling, and physics-constrained adapter tuning transfers meteorological foundation representations to global SST forecasting (Niu et al., 2026; Li et al., 2026a). These studies provide strong numerical spatiotemporal and physics-guided inductive biases for SST prediction. The complementary question studied here is how forecast-date environmental descriptions and a compact representation of regional spatial state can be jointly supplied to an LLM for conditional numerical generation. Our method represents the latter through static and dynamic graphs rather than imposing an explicit physical-equation constraint.

2.2 LLMs for numerical and text-conditioned forecasting

Some work adapt time-series to the representation space of language models. Time-LLM aligns time-series patches with text prototypes and uses a prompt prefix to provide task context (Jin et al., 2024). TimeCMA aligns numerical and prompt encodings, LangTime combines temporal-comprehension prompts with reinforcement-learning fine-tuning, and SE-LLM introduces temporal–semantic interaction and a Time-Adapter (Liu et al., 2025a; Niu et al., 2025; Liu et al., 2026a). These methods primarily address the interface between numerical patterns and language-model representations.

Another line studies the predictive value of textual context. From News to Forecast integrates selected news with historical sequences and fine-tunes an LLM for numerical generation (Wang et al., 2024).

VoT jointly uses exogenous event text and endogenous sequence descriptions, with fusion at both the representation and prediction-frequency levels; TaTS treats time-aligned text as an auxiliary variable for numerical forecasting (Wang et al., 2026; Li et al., 2026b). For the related task of text-conditioned sequence synthesis, VerbalTS generates time series from unstructured descriptions (Gu et al., 2025). Our setting focuses on regional SST forecasting, where date-aligned environmental records and static ocean knowledge form textual conditions, while a separate spatial branch summarizes regional SST state for the target cell.

2.3 Graph-conditioned inputs for language models

Graph convolution learns node representations by aggregating neighborhood information, providing a basis for regional spatial modeling (Kipf & Welling, 2017). When graph representations are supplied to an LLM, the central interface problem is converting non-sequential structure into token-level inputs. GraphToken learns graph-conditioned soft prompts from GNN representations, while G-Retriever combines subgraph retrieval, graph encoding, and soft prompting for textual-graph question answering (Perozzi et al., 2024; He et al., 2024). GOFA inserts GNN layers into an LLM for joint graph–language modeling, and LGPT studies query-aware learnable graph-pooling tokens to balance token cost and information compression (Kong et al., 2025; Kim et al., 2025).

These studies motivate the use of continuous graph-derived tokens as an interface between graph encoders and LLMs. Our setting differs from graph question answering and graph-language modeling: static and dynamic GNNs first aggregate regional SST information, and the resulting target-node state is projected into continuous prefix tokens for conditional numerical generation. The representation is conditioned on the forecast date and target location, while its prefix length remains independent of the number of regional grid cells.

2.4 Ocean knowledge and forecast interpretation

OKG-LLM aligns ocean-knowledge representations with SST observations through an LLM-based prediction pipeline (Yang et al., 2026). In contrast, our framework uses static ocean knowledge and forecast-date environmental records as textual conditions, while encoding regional spatial state with target-specific prefixes derived from static and dynamic graphs.

TimeCAP, TimeXL, and TS-RAG incorporate context generation, refinement, or retrieval into time-series prediction pipelines (Lee et al., 2025; Jiang et al., 2025; Ning et al., 2025). Our knowledge lookup runs separately after forecasting, matching predicted trends and environmental-factor directions to source-linked entries for rule-based post-hoc contextualization.

3 Method

3.1 Task formulation and framework overview

Let the study region contain N regular grid cells. Cell i has coordinates $\mathbf{c}_i$ and an ocean mask $m_i \in \{0, 1\}$. At forecast origin t, let $\mathbf{X}_t = [\mathbf{x}_{t-L+1}, \ldots, \mathbf{x}_t] \in \mathbb{R}^{L\times N}$ and $\mathbf{D}_t = [\mathbf{d}_{t-L+1}, \ldots, \mathbf{d}_t] \in \mathbb{R}^{L\times N}$ denote the regional SST and anomaly histories, respectively. The forecast-date environmental record $\mathbf{e}_t$ contains tropical-cyclone, ENSO, and regional-circulation context, while K denotes static ocean knowledge. All date-dependent environmental records and dynamic-graph statistics are constructed using observations available no later than forecast origin t.

We construct a textual condition $q_{t,i}$ from the target-cell histories, environmental record, and static knowledge, and construct a static graph $\mathcal{G}^s = (\mathbf{A}^s, \mathbf{F}^s)$ and a date-specific dynamic graph $\mathcal{G}^d_t = (\mathbf{A}^d_t, \mathbf{F}^d_t)$ for the regional SST field. For an ocean target cell i, the task is to predict the next H SST values, $\mathbf{y}_{t,i} = (x_{t+1,i}, \ldots, x_{t+H,i})$, as

$$\widehat{\mathbf{y}}_{t,i} = f_\theta\big(q_{t,i}, \mathcal{G}^s, \mathcal{G}^d_t, i\big)\,. \tag{1}$$

The model compresses the graph-derived target-node state into continuous prefix tokens and concatenates them with the embeddings of $q_{t,i}$ for autoregressive numerical generation. Each sample corresponds to one forecast origin and one target cell; samples from the same forecast origin share the regional graph inputs but receive target-specific textual conditions and spatial prefixes. Land cells

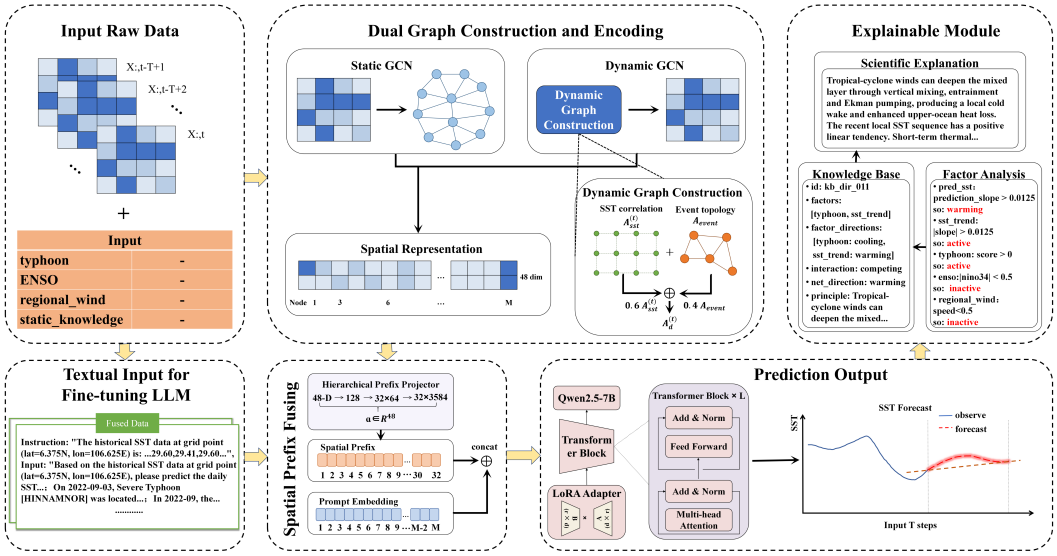

Figure 1: Overview of the framework. At forecast origin t, target-cell SST and anomaly histories, environmental records, and static ocean knowledge form a textual condition. Regional observations construct a static geographic–climatological graph and a date-specific dynamic graph. Two GNN branches encode the target-node state, which a factorized projector maps to continuous prefix tokens for autoregressive LLM forecasting. A separate rule-based module then matches forecast trends and environmental factors with knowledge entries for source-linked post-hoc contextualization.

remain in the regular indexing scheme for graph construction, but are masked and excluded from prediction targets and evaluation. Figure. 1 summarizes the framework.

3.2 Multi-source observations and textual input

Historical observations and environmental context. We use the OISST `sst` and `anom` variables (Reynolds et al., 2007) and preserve a consistent grid ordering after regional cropping. The anomaly sequence is supplied as an independent observation field. Tropical-cyclone records from IBTrACS (Knapp et al., 2010) are filtered within the target region and a 3° buffer using a 34-kt wind threshold. For each forecast date, we retain the observation with the maximum sustained wind and its location and intensity. ENSO context is derived from monthly Niño3.4 anomalies relative to a climatological reference period. Regional circulation comes from the ERA5 850-hPa wind field (Hersbach et al., 2020). We first compute cosine-latitude- weighted zonal and meridional means, then derive the resultant wind speed and meteorological direction:

$$\bar{u} = \frac{\sum_j u_j \cos\varphi_j}{\sum_j \cos\varphi_j}, \qquad \bar{v} = \frac{\sum_j v_j \cos\varphi_j}{\sum_j \cos\varphi_j}, \qquad s = \sqrt{\bar{u}^2 + \bar{v}^2}. \tag{2}$$

Unified forecasting prompt. Following the organization of text-conditioned time-series generation (Wang et al., 2024), we construct $q_{t,i} = \text{Template}(\mathbf{X}_t[:,i], \mathbf{D}_t[:,i], \mathbf{c}_i, t, L, H, E_t, K)$, where E_t is the organized environmental context. The instruction specifies the target cell and historical SST, the input contains the forecast task, date range, anomaly sequence, environmental context, and static ocean knowledge, and the response is the comma-separated future SST sequence. During training, the ground-truth response follows the prompt; during inference, the response is generated autoregressively. The environmental context describes the target date, whereas the static text provides regional background on persistence, mixing, advection, upwelling, and air–sea exchange. Both are tokenized together with the numerical history.

3.3 Static and dynamic spatial graphs

Both graphs use the same grid nodes but encode different relations. Let $\mathcal{R}(\mathbf{A})$ denote row normalization with zero-sum rows preserved. Land rows and columns are zeroed before normalization.

Static geographic graph. For great-circle distance d_{ij}, the geographic edge weight is

$$\widetilde{A}^s_{ij} = m_i m_j \mathbf{1}[i \neq j]\mathbf{1}[d_{ij} < r_s] \exp(-d_{ij}/\ell_s), \qquad \mathbf{A}^s = \mathcal{R}(\widetilde{\mathbf{A}}^s). \tag{3}$$

We use $\ell_s = 111$ km and $r_s = 1.5\ell_s$. Static node features $\mathbf{F}^s \in \mathbb{R}^{N\times 4}$ contain normalized latitude, normalized longitude, climatological SST mean, and climatological SST standard deviation.

Historical SST correlation graph. Geographic distance alone cannot capture synchronous changes in the current history window. We therefore compute Pearson correlations among complete, variable ocean-node histories. Self-connections and non-positive correlations are removed, and at most $k = 10$ positive neighbors are retained per source node:

$$A^{x,t}_{ij} = \begin{cases} \dfrac{\exp(\rho^{(t)}_{ij})}{\sum_{u\in\mathcal{N}^+_i(t)} \exp(\rho^{(t)}_{iu})}, & j \in \mathcal{N}^+_i(t), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Localized tropical-cyclone graph. For a storm observation o with center $\mathbf{c}_o$ and sustained wind v_o, we define an intensity score and a finite-support influence kernel:

$$\begin{aligned} a_o &= \text{clip}(v_o/v_{99}, 0, 1), & R_o &= 111\,\text{clip}\Big(2\sqrt{v_o/64}, 1.5, 5\Big)\ \text{km}, \\ \sigma_o^2 &= -R_o^2/(2\log 0.1), & s^o_i &= m_i a_o \exp\left(-\frac{d^2_{io}}{2\sigma_o^2}\right)\mathbf{1}[d_{io} \le R_o]. \end{aligned} \tag{5}$$

Here v_{99} is the 99th percentile of valid storm winds, with 130 kt used when no valid event wind is available. The event affinity is

$$B^o_{ij} = \mathbf{1}[i \neq j]\sqrt{s^o_i s^o_j}\exp\left(-\frac{d^2_{ij}}{2\sigma_o^2}\right), \qquad \mathbf{A}^{e,o} = \text{diag}(\mathbf{s}^o)\mathcal{R}(\mathbf{B}^o). \tag{6}$$

For multiple observations, we take the element-wise maximum, and use a zero event graph when no storm is active. The dynamic adjacency is $\mathbf{A}^d_t = \mathcal{R}(0.6\mathbf{A}^{x,t} + 0.4\mathbf{A}^{e,t})$. ENSO and regional circulation are supplied as text and explanation context; they are not additional edge types in $\mathbf{A}^d_t$.

Dynamic node features. The dynamic feature matrix $\mathbf{F}^d_t \in \mathbb{R}^{N\times 7}$ contains the full-window SST mean and standard deviation, linear trend, recent seven-step mean, full-window anomaly mean, recent anomaly mean, and SST range. Features are standardized across ocean nodes for each date, while land features remain zero. Node features describe the current state of each cell, while the dynamic graph determines how these states are aggregated.

3.4 Dual-GNN spatial encoding

Geographic background and date-specific regional state evolve at different rates. We therefore use two independent two-layer GCNs. For branch $b \in \{s, d\}$ with features $\mathbf{F}^b$ and adjacency $\mathbf{A}^b$,

$$\begin{aligned} \mathbf{U}^b &= \text{Dropout}_{0.1}\big[\text{ReLU}\big(\text{BN}(\mathbf{A}^b\mathbf{F}^b\mathbf{W}^b_1 + \mathbf{b}^b_1)\big)\big], \\ \mathbf{H}^b &= \mathbf{A}^b\mathbf{U}^b\mathbf{W}^b_2 + \mathbf{b}^b_2 + \mathbf{F}^b\mathbf{W}^b_r + \mathbf{b}^b_r. \end{aligned} \tag{7}$$

Both hidden layers have width32. The static branch outputs16 dimensions and the dynamic branch outputs32 dimensions. After message passing over the entire region, we concatenate the two representations and select the target cell:

$$\mathbf{H}_t = [\mathbf{H}^s \| \mathbf{H}^d_t] \in \mathbb{R}^{N\times 48}, \qquad \mathbf{h}_{t,i} = \mathbf{H}_t[i,:] \in \mathbb{R}^{48}. \tag{8}$$

3.5 Hierarchical spatial-prefix fusion

The joint spatial vector and the LLM input have different dimensions. We map the target representation to $P = 32$ continuous prefix tokens, with intermediate token width $d_m = 64$ and LLM hidden width d_{LM}:

$$\begin{aligned}\mathbf{u}_{t,i} &= \text{ReLU}(\mathbf{W}_1\mathbf{h}_{t,i} + \mathbf{b}_1) \in \mathbb{R}^{128},\\ \mathbf{V}_{t,i} &= \text{reshape}_{P\times d_m}[\text{ReLU}(\mathbf{W}_2\mathbf{u}_{t,i} + \mathbf{b}_2)],\\ \mathbf{Z}_{t,i} &= \alpha(\mathbf{V}_{t,i}\mathbf{W}_3 + \mathbf{b}_3) \in \mathbb{R}^{P\times d_{\text{LM}}}.\end{aligned} \tag{9}$$

The third linear layer shares parameters across prefix positions. For Qwen2.5-7B, $d_{\text{LM}} = 3584$, yielding $48 \to 128 \to 32 \times 64 \to 32 \times 3584$. The projector contains 503,425 trainable parameters, compared with 5,619,713 for a direct $48 \to 32 \times 3584$ mapping, a reduction of approximately 91% for this comparison. The learnable scalar α is initialized as 1 and adjusts amplitude during training.

Let $\mathbf{E}_{t,i}$ denote the token embeddings of the prompt. The LLM input is $\mathbf{S}_{t,i} = [\mathbf{Z}_{t,i}; \mathbf{E}_{t,i}]$, where the semicolon denotes concatenation along the sequence dimension. Each sample receives only the target cell's prefix, so the prefix length does not scale with the number of regional nodes.

3.6 Autoregressive training and forecasting

The ground-truth numerical sequence is appended to the prompt as the esponse. We train the model with the standard autoregressive language-model objective, using supervised fine-tuning to predict the next response token from the prompt and preceding tokens:

$$\mathcal{L} = -\frac{1}{\sum_{(t,i),j} a_{t,i,j}} \sum_{(t,i),j} a_{t,i,j} \log p_\theta(w_{t,i,j} \mid \mathbf{Z}_{t,i}, w_{t,i,<j}). \tag{10}$$

The parameters of the dual GNNs, prefix projector, prefix scale, and LoRA adapters are jointly optimized. At inference, the numerical response is generated autoregressively under the same conditioning sequence. Qwen2.5-7B is loaded in 8-bit precision with BF16 computation, and rank-8 LoRA adapters with scale 16 are attached to the attention and feed-forward projections (Hu et al., 2022).

3.7 Post-hoc factor-direction explanation

The explanation module reads the forecast, historical SST, and environmental records after generation. For a temperature sequence $\mathbf{v}$, we compute a linear slope and classify its direction with threshold τ:

$$\beta(\mathbf{v}) = \frac{\sum_j (j - \bar{j})(v_j - \bar{v})}{\sum_j (j - \bar{j})^2}, \qquad \delta(\mathbf{v}) = \begin{cases} +1, & \beta(\mathbf{v}) > \tau, \\ -1, & \beta(\mathbf{v}) < -\tau, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The current rule activates four factors: tropical cyclone influence, ENSO state, regional wind, and historical SST trend. It then matches the active factor directions and forecast direction against 52 explicit knowledge entries. Neutral forecasts receive a neutral explanation, while non-neutral forecasts without a supporting active factor receive a conflict explanation. Otherwise, the module returns the matched knowledge identifier, mechanism description, and references.

4 Experiments

4.1 Dataset and evaluation setting

Dataset: We study SST forecasting in the South China Sea using 0.25° OISST data. The region is represented by a 64×64 regular grid, with scores computed only on ocean cells. The data span 1 September 1981 to 16 February 2025 and are partitioned chronologically into training, validation, and testing periods. Each sample uses ten historical observations to predict the next ten steps, and no window crosses a partition boundary.

Evaluation setting: Training and validation samples are drawn from valid windows in their respective periods with coverage across spatial regions, coastal types, environmental events, temperature

Table 1: SST forecasting results over ten prediction steps. **Bold** entries mark the best results.

Model	Metric	Days										Mean
		1	2	3	4	5	6	7	8	9	10	
Ours	RMSE	**.212**	.322	.381	.424	.457	.485	**.512**	**.535**	**.553**	**.566**	**.445**
	MAE	**.131**	**.216**	**.269**	**.306**	**.335**	**.358**	**.378**	**.396**	**.413**	**.427**	**.323**
	R^2	**.991**	.979	.970	.963	**.956**	**.950**	**.943**	**.937**	**.931**	**.927**	**.955**
PCFNet	RMSE	.273	**.294**	**.352**	**.401**	**.444**	**.484**	.514	.545	.574	.597	.448
	MAE	.207	.223	.269	.307	.340	.372	.395	.419	.444	.464	.344
	R^2	.983	**.980**	**.972**	**.964**	.956	.948	.941	.935	.928	.922	.953
CoTCN	RMSE	.281	.336	.382	.430	.477	.515	.540	.568	.588	.602	.472
	MAE	.218	.261	.298	.337	.375	.405	.425	.446	.461	.469	.370
	R^2	.982	.974	.967	.958	.949	.941	.935	.929	.924	.921	.948
PANN	RMSE	.297	.305	.384	.411	.484	.529	.548	.578	.590	.619	.475
	MAE	.239	.264	.295	.349	.382	.430	.453	.472	.487	.493	.386
	R^2	.981	.971	.967	.956	.952	.943	.936	.920	.920	.915	.946
EarthFarseer	RMSE	.312	.374	.424	.478	.530	.572	.600	.631	.653	.669	.524
	MAE	.243	.290	.331	.374	.416	.450	.472	.496	.512	.521	.411
	R^2	.978	.968	.959	.949	.937	.927	.920	.912	.906	.903	.936
Swin Transformer	RMSE	.404	.434	.487	.511	.543	.565	.586	.604	.633	.639	.541
	MAE	.331	.355	.400	.415	.439	.457	.472	.486	.511	.515	.438
	R^2	.963	.957	.946	.941	.934	.929	.924	.920	.912	.911	.934
PredRNN	RMSE	.449	.483	.541	.567	.603	.628	.651	.671	.703	.710	.601
	MAE	.368	.394	.444	.461	.488	.507	.524	.540	.568	.572	.487
	R^2	.954	.947	.934	.927	.918	.912	.906	.901	.891	.890	.918
ConvLSTM	RMSE	.464	.481	.512	.545	.574	.592	.606	.620	.631	.637	.566
	MAE	.396	.408	.431	.458	.481	.494	.503	.513	.521	.524	.473
	R^2	.951	.947	.941	.933	.926	.922	.919	.915	.913	.911	.928
CNN	RMSE	.515	.544	.566	.602	.641	.687	.712	.742	.750	.763	.652
	MAE	.427	.447	.461	.489	.520	.556	.574	.597	.604	.615	.529
	R^2	.939	.932	.927	.918	.908	.895	.888	.878	.877	.873	.904

states, and trajectory trends. Future trend labels are used only to construct training coverage and are never provided as prediction inputs. The test set covers all valid test dates and ocean cells, following the complete evaluation protocol used for the PCFNet comparison. Qwen2.5-7B-Instruct is loaded in 8-bit precision with BF16 computation. LoRA uses rank 8 and scale 16; the learning rate is 10^{-4} with 100 warm-up steps, one training epoch, and random seed as 42.

We report RMSE, MAE, and the coefficient of determination R^2 at each lead, and use the arithmetic mean across the ten leads for the summary value:

$$\text{RMSE}_h = \sqrt{\frac{1}{|\mathcal{S}_h|}\sum_{n\in\mathcal{S}_h}(\widehat{y}_{n,h}-y_{n,h})^2}, \quad \text{MAE}_h = \frac{1}{|\mathcal{S}_h|}\sum_{n\in\mathcal{S}_h}|\widehat{y}_{n,h}-y_{n,h}|, \quad R^2_h = 1-\frac{\sum_{n\in\mathcal{S}_h}(\widehat{y}_{n,h}-y_{n,h})^2}{\sum_{n\in\mathcal{S}_h}(y_{n,h}-\bar{y}_h)^2}$$

Table 2: SST forecasting results across four LLM backbones. **Bold** values mark the best result for each metric and lead; ties at the displayed precision are also bold.

Model	Metric	Days										Mean
		1	2	3	4	5	6	7	8	9	10	
Qwen2.5-7B-Instruct	RMSE	.212	.322	.381	.424	.457	.485	**.512**	**.535**	**.553**	**.566**	.445
	MAE	.131	.216	.269	.306	.335	**.358**	**.378**	**.396**	**.413**	**.427**	**.323**
	R^2	.991	.979	.970	**.963**	**.956**	**.950**	**.943**	**.937**	**.931**	**.927**	**.955**
Llama-3.1-8B-Instruct	RMSE	**.170**	**.271**	**.341**	**.397**	**.444**	**.483**	.517	.546	.572	.595	**.434**
	MAE	**.115**	**.194**	**.251**	**.297**	**.334**	.366	.393	.417	.439	.458	.326
	R^2	**.993**	**.982**	**.971**	.961	.952	.943	.934	.927	.920	.913	.950
Llama-2-7b-chat-hf	RMSE	.171	.272	.342	.399	.446	.487	.521	.551	.577	.600	.436
	MAE	.116	.195	.252	**.297**	.335	.367	.395	.419	.440	.459	.327
	R^2	**.993**	**.982**	**.971**	.961	.951	.942	.933	.926	.918	.912	.949
Mistral-7B-Instruct-v0.3	RMSE	.176	.277	.347	.403	.450	.491	.526	.557	.584	.608	.442
	MAE	.120	.200	.256	.301	.339	.372	.401	.425	.448	.468	.333
	R^2	.992	.981	.970	.960	.950	.941	.932	.924	.916	.909	.948

4.2 Main results

Table 1 reports lead-wise results against PCFNet and the other specialized models listed in the corresponding SST benchmark table. Compared with the best baseline PCFNet, the complete configuration reduces mean RMSE from 0.448 to 0.445 and mean MAE from 0.344 to 0.323, while increasing mean R^2 from 0.953 to 0.955. The RMSE comparison is lead-dependent: our model performs better at the first prediction step and at steps 7–10, whereas PCFNet attains lower RMSE at steps 2–6. Overall, the results indicate improved mean RMSE, MAE and R^2.

4.3 Comparison across LLM backbones

We compare the Qwen2.5-7B-Instruct backbone used in the full model with Llama-2-7b-chat-hf, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3. Table 2 reports results at each forecast lead and the arithmetic mean over ten steps. Llama-3.1-8B-Instruct achieves the lowest mean RMSE (0.434), 0.011 below Qwen (0.445), whereas Qwen obtains the lowest mean MAE (0.323) and the highest mean R^2 (0.955). The RMSE comparison varies with forecast lead: Llama-2 and Mistral have lower RMSE than Qwen at steps 1–5, and Llama 3.1 does so at steps 1–6; Qwen has lower RMSE than all three alternatives at steps 7–10. Thus, the backbone comparison shows a metric- and horizon-dependent trade-off rather than a uniform improvement.

4.4 Ablation analysis

We evaluate four variants. `no_static` removes static ocean-knowledge text, and `no_dynamic` removes dynamic environmental text; both retain the two GNN branches. `no_knowledge` removes both text sources but retains the spatial branch. `no_knowledge_no_gnn` removes the GNN branch entirely and therefore also has no spatial prefix input.

The ablations show that the two text sources contribute complementary information. Removing dynamic context gives RMSE 0.449, MAE 0.340, and $R^2 = 0.946$, whereas removing static knowledge gives RMSE 0.437, MAE 0.330, and $R^2 = 0.949$. The complete model has the best MAE and R^2 among these text configurations, indicating that combining static background with date-specific context improves overall consistency. Removing both text sources further reduces performance to RMSE 0.456, MAE 0.347, and $R^2 = 0.944$.

Table 3: Ablation results on the variants of our proposed framework.

Variants	Static-knowledge	Dynamic-context text	dual-GNN	RMSE	MAE	R^2
Full	yes	yes	yes	.445	**.323**	**.955**
No static text	no	yes	yes	**.437**	.330	.949
No dynamic text	yes	no	yes	.449	.340	.946
No knowledge text	no	no	yes	.456	.347	.944
No GNN	no	no	no	.464	.354	.943

Under the no-knowledge condition, removing the GNN branch increases RMSE from 0.456 to 0.464 and MAE from 0.347 to 0.354, while reducing R^2 from 0.944 to 0.943. This result supports the contribution of regional spatial information encoded by the dual GNNs. The current ablation evaluates the spatial branch as a whole and does not separately vary the two GCNs or the prefix projector.

The principle of kb_dir_011	The principle of kb_dir_046
Tropical-cyclone winds can deepen the mixed layer through vertical mixing, entrainment and Ekman pumping, producing a local cold wake and enhanced upper-ocean heat loss. The recent local SST sequence has a positive linear tendency. Short-term thermal persistence makes continued warming a consistent interpretation over the forecast horizon, but the tendency alone is not a causal proof. The warming signals (recent local SST tendency) and cooling signals (tropical-cyclone forcing) compete; the observed net prediction direction is warming, without implying that an internal model weight or physical causal dominance has been measured.	**Tropical-cyclone winds can deepen the mixed layer through vertical mixing, entrainment and Ekman pumping, producing a local cold wake and enhanced upper-ocean heat loss. In the South China Sea, seasonal regional winds can enhance latent-heat loss and upper-ocean mixing; where the circulation supports offshore Ekman transport, upwelling and cold-water entrainment further favor SST cooling. A warm-phase ENSO background can alter regional winds, clouds, surface heat fluxes and ocean advection through teleconnections, providing a warming tendency whose sign depends on location and season. The recent local SST sequence has a positive linear tendency. Short-term thermal persistence makes continued warming a consistent interpretation over the forecast horizon, but the tendency alone is not a causal proof. The warming signals (ENSO teleconnection, recent local SST tendency) and cooling signals (tropical-cyclone forcing, regional wind forcing) compete; the observed net prediction direction is warming, without implying that an internal model weight or physical causal dominance has been measured.**
NEUTRAL_EXPLANATION	
The combined signals from the four core factors do not indicate a clear unidirectional warming or cooling effect. The predicted values broadly maintain the historical SST state with only small fluctuations.	
CONFLICT_EXPLANATION	
The combined signals from the four core factors are inconsistent with the predicted trend. The current explanation framework cannot provide a clear explanation for this prediction.	

Figure 2: Illustrative explanation outputs. The left column shows the principle for `kb_dir_011` and the neutral and conflict fallback messages; the right column shows the principle for `kb_dir_046`.

4.5 Qualitative explanation examples

The explanation module establishes a deterministic mapping from activated factor directions to knowledge entries. In `kb_dir_011`, tropical-cyclone influence is cooling while the recent SST trend is warming, producing a competing two-factor pattern with a warming forecast direction. In `kb_dir_046`, all four factors are active: tropical cyclone and regional wind are cooling, whereas ENSO and the recent SST trend are warming. The retrieved identifier and references make the rule-level explanation traceable, but do not constitute an internal attribution or a causal estimate. The module returns a neutral message for forecasts within the neutral threshold, and a conflict message when no active factor supports a directional forecast. Figure 2 collects the two knowledge principles and both fallback messages discussed above.

5 Conclusion

We presented an LLM-based framework for regional SST forecasting that combines textual environmental context with static and date-specific dynamic graph representations of regional spatial state. A factorized spatial-prefix projector maps the target-node representation to continuous tokens for autoregressive generation, while a separate rule-based module provides source-linked post-hoc contextualization. On South China Sea SST forecasting, the complete configuration achieves the lowest mean RMSE, MAE and highest mean R^2 among the compared methods. Ablations show metric-dependent effects of individual text components. Future work will evaluate additional backbones and generalization across regions.

AI USE STATEMENT

In preparing this submission, the authors used a generative AI tool to assist with LaTeX formatting, translation, and preparation of tables from existing experiment logs.

REFERENCES

Yingzhe Cui, Ruohan Wu, Xiang Zhang, Ziqi Zhu, Bo Liu, Jun Shi, Junshi Chen, Hailong Liu, Shenghui Zhou, Liang Su, Zhao Jing, Hong An, and Lixin Wu. Forecasting the eddying ocean with a deep neural network. *Nature Communications*, 16:2268, 2025. doi: 10.1038/s41467-025-57389-2.

Matthew H. England, Zhi Li, Maurice F. Huguenin, Andrew E. Kiss, Alex Sen Gupta, Ryan M. Holmes, and Stefan Rahmstorf. Drivers of the extreme North Atlantic marine heatwave during 2023. *Nature*, 642:636–643, 2025. doi: 10.1038/s41586-025-08903-5.

Shuqi Gu, Chuyue Li, Baoyu Jing, and Kan Ren. VerbalTS: Generating time series from texts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 20448–20476, 2025. URL https://proceedings.mlr.press/v267/gu25a.html.

Lu Han, Yu Liu, Lan Li, Qiwen Deng, Jian Jiang, Yinbo Sun, Zhe Yu, Binfeng Wang, Xingyu Lu, Lintao Ma, Han-Jia Ye, and De-Chuan Zhan. UniCA: Unified covariate adaptation for time series foundation model. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0b5eb45a22ff33956c043dd271f244ea-Abstract-Conference.html.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/efaf1c9726648c8ba363a5c927440529-Abstract.html.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Mu noz 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. doi: 10.1002/qj.3803.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yuxiao Hu, Qian Li, Dongxiao Zhang, Jinyue Yan, and Yuntian Chen. Context-alignment: Activating and enhancing LLM capabilities in time series. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e1de63ec74f40d3234c4e053f3528e18-Paper-Conference.pdf.

Yushan Jiang, Wenchao Yu, Geon Lee, Dongjin Song, Kijung Shin, Wei Cheng, Yanchi Liu, and Haifeng Chen. TimeXL: Explainable multi-modal time series prediction with LLM-in-the-loop. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c5433ab4056ca58db67be4578c384cba-Abstract-Conference.html.

Kundong Jin, Xiaoyu He, Jinkun Yang, Rong Jin, and Feng Fu. Decoupled dynamic spatial–temporal graph neural network for sea surface temperature prediction. *IEEE Geoscience and Remote Sensing Letters*, 22, 2025. doi: 10.1109/LGRS.2025.3558739.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Unb5CVPtae.

Wooyoung Kim, Byungyoon Park, and Wooju Kim. Query-aware learnable graph pooling tokens as prompt for large language models. *arXiv preprint arXiv:2501.17549*, 2025. URL https://arxiv.org/abs/2501.17549.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Kenneth R. Knapp, Michael C. Kruk, David H. Levinson, Howard J. Diamond, and Charles J. Neumann. The international best track archive for climate stewardship (IBTrACS): Unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010. doi: 10.1175/2009BAMS2755.1.

Lecheng Kong, Jiarui Feng, Hao Liu, Chengsong Huang, Jiaxin Huang, Yixin Chen, and Muhan Zhang. GOFA: A generative one-for-all model for joint graph language modeling. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/652c104b5b0652a03684efeaf805463b-Abstract-Conference.html.

Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. TimeCAP: Learning to contextualize, augment, and predict time series events with large language model agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18082–18090, 2025. doi: 10.1609/aaai.v39i17.33989.

Tianbao Li, Yuting Su, Dan Song, Wenhui Li, Lanjun Wang, Zhiqiang Wei, and An-An Liu. Physics-constrained adapter tuning of meteorological foundation models for global SST forecasting. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026a. doi: 10.1109/TGRS.2026.3676064.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, and Jingrui He. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=a1zBg9cBvt.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards LLM-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18780–18788, 2025a. doi: 10.1609/aaai.v39i18.34067.

Fukai Liu, Fengfei Song, and Yiyong Luo. Human-induced intensified seasonal cycle of sea surface temperature. *Nature Communications*, 15:3948, 2024a. doi: 10.1038/s41467-024-48381-3.

Hao Liu, Xiaoxing Zhang, Chun Yang, and Xiaobin Zhu. Semantic-enhanced time-series forecasting via large language models. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2508.07697.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-MMD: Multi-domain multimodal dataset for time series analysis. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8e7768122f3eeec6d77cd2b424b72413-Abstract-Datasets_and_Benchmarks_Track.html.

Xiong Liu, Ning Song, Jie Nie, Min Ye, Jun Ma, Yuchen Yuan, and Zhiqiang Wei. DP-BICNN: A bidirectional information compensation neural network coupled with data-driven and physical information for sea surface temperature prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 2025b. doi: 10.1109/TGRS.2025.3560717.

Yingbing Liu, Cong Xiao, Guangwen Peng, Wenying Du, and Changjiang Xiao. PCFNet: A sea surface temperature prediction network fusing physical constraints with frequency domain feature extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–15, 2026b. doi: 10.1109/TGRS.2026.3694412.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Yiming Zhang, Kashif Rasul, Anderson Schneider, Lintao Ma, Yuriy Nevmyvaka, and Dongjin Song. TS-RAG: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/eed25c037bc08afcbefab6f7a6b700e0-Abstract-Conference.html.

Di Niu, Caichao Lv, Ning Song, Yufei Guo, Yanfeng Pan, Xinyue Liang, and Jie Nie. PG-DSBN: A physics-guided deterministic–stochastic balanced network for sea surface temperature forecasting. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026. doi: 10.1109/TGRS.2026.3707295.

Wenzhe Niu, Zongxia Xie, Yanru Sun, Wei He, Man Xu, and Chao Hao. LangTime: A language-guided unified model for time series forecasting with proximal policy optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 46712–46734, 2025. URL https://proceedings.mlr.press/v267/niu25e.html.

Guangwen Peng, Yingbing Liu, Juncheng Luo, Cong Xiao, Wenying Du, and Changjiang Xiao. A dual-path recurrent framework integrating optical flow guidance and spatiotemporal-aware learning for sea surface temperature prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 2025. doi: 10.1109/TGRS.2025.3640632.

Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. Let your graph do the talking: Encoding structured data for LLMs. *arXiv preprint arXiv:2402.05862*, 2024. URL https://arxiv.org/abs/2402.05862.

Richard W. Reynolds, Thomas M. Smith, Chunying Liu, Dudley B. Chelton, Kenneth S. Casey, and Michael G. Schlax. Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20(22):5473–5496, 2007. doi: 10.1175/2007JCLI1824.1.

Siyuan Wang, Peng Chen, Yihang Wang, Wanghui Qiu, Chenjuan Guo, Bin Yang, and Yang Shu. Unlocking the value of text: Event-driven reasoning and multi-level alignment for time series forecasting. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0TAFiyHgEl.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in LLM-based time series forecasting with reflection. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6aef8bffb372096ee73d98da30119f89-Abstract-Conference.html.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a8e18adf6489fd3a27417d128660eaa9-Abstract-Conference.html.

Hanchen Yang, Jiaqi Wang, Jiannong Cao, Wengen Li, Jialun Zheng, Yangning Li, Chunyu Miao, Jihong Guan, Shuigeng Zhou, and Philip S. Yu. OKG-LLM: Aligning ocean knowledge graph with observation data via LLMs for global sea surface temperature prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2026. doi: 10.1109/TKDE.2026.3674110. Accepted author manuscript.

Xingyu Zhao, Zhongyu Wang, Zhihao Zhang, Fangbo Lu, Yanwei Yu, and Junyu Dong. Adaptive spatiotemporal graph recurrent network for sea surface temperature forecasting. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024. doi: 10.1109/TGRS.2024.3502841.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-VLM: Exploring multimodal vision-language models for augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 78478–78497, 2025. URL https://proceedings.mlr.press/v267/zhong25a.html.