https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

Deep Learning-Based Sea Surface Temperature Prediction for Short-term Marine Heatwave Detection in the East China Sea

Zefang Ma[1], Hui Chen[1], Qiyan Ji[1], Lifang Jiang[2,3], Cui Shen[1], Xiayan Lin[1]

1 Marine Science and Technology College, Zhejiang Ocean University, Zhoushan, 316022, China
2 Key Laboratory of Marine Environmental Survey Technology and Application, Ministry of Natural Resources, Guangzhou 510310, China
3 South China Sea Marine Forecast and Hazard Mitigation Center, Ministry of Natural Resources, Guangzhou 510310, China

Correspondence to: Hui Chen (chenhui@zjou.edu.cn) and Qiyan Ji (jiqiyan@zjou.edu.cn)

Abstract. Accurate short-term sea surface temperature (SST) prediction is important for supporting fisheries management and marine heatwave (MHW) detection. However, SST prediction in the East China Sea (ECS) remains challenging due to its complex oceanic and atmospheric processes. In this study, we developed a hybrid SwinTrans-ConvLSTM model that integrates ConvLSTM and Swin Transformer for short-term SST prediction and subsequent MHW detection. The model uses 14 consecutive days of historical data to predict daily SST for the subsequent 7 days. Its predictive performance was evaluated against OSTIA SST analyses over the 2020–2024 period and compared with four other deep learning models, as well as persistence and AR(1) baselines. Across all forecast lead times (1–7 days), SwinTrans-ConvLSTM achieved the lowest pooled MAE and RMSE of 0.5011 °C and 0.7081 °C, respectively. Even at the 7-day lead, its MAE and RMSE were 0.6378 °C and 0.8620 °C, respectively, representing error reductions of 9.8% and 7.2% compared with AR(1). Subsequent MHW detection using the forecast SST fields yielded a higher SEDI for the proposed model (0.6590) than for persistence (0.5845) and AR(1) (0.5794) at a seven-day lead time. Beyond its superior MHW detection performance, the proposed model also yielded the lowest errors in mean MHW intensity, event count, and total MHW days among the five deep learning models for next-day forecasts. Overall, these results highlight the potential of deep learning-based SST forecasts for short-term MHW detection in the ECS and provide a foundation for further research into event-based MHW onset prediction.

1 Introduction

Marine heatwaves (MHWs) are extreme ocean warming events characterized by sea surface temperatures (SSTs) exceeding the local climatological 90th percentile for at least five consecutive days (Hobday et al., 2016). With continued global warming, MHWs have become more frequent, longer-lasting, and more intense (Oliver et al., 2018). These events have caused widespread ecological and socioeconomic impacts, including coral bleaching, benthic habitat degradation, and species redistribution, while also amplifying the vulnerability of coastal fisheries and associated economies (Smale et al., 2019).

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

In recent decades, the East China Sea (ECS) has experienced rapid sea surface temperature (SST) warming exceeding the global ocean average, making it a hotspot for marine heatwaves (MHWs) (Sasaki and Umeda, 2021; Xu et al., 2025). As a dynamically complex marginal shelf sea, ECS is always influenced by multiple water masses and currents, including Changjiang Diluted Water, coastal currents, and the Kuroshio (Hickox et al., 2000). Their interactions with the East Asian monsoon generate pronounced temperature–salinity fronts and active submesoscale processes (Chelton and Xie, 2010). SST variability in ECS is further modulated by local air–sea interactions and large-scale climate modes, such as the western Pacific subtropical high and the Indian Ocean Dipole, which affect surface winds, solar radiation, and air–sea heat fluxes (Tan et al., 2023). These atmospheric and oceanic processes, together with the pronounced spatiotemporal variability, make SST prediction in the ECS particularly challenging.

Accurate SST prediction is essential for reliable marine heatwave (MHW) detection. SST prediction errors can lead to false or missed detections and uncertainties in estimating MHW onset, duration, intensity, and spatial extent. To improve SST prediction, various approaches have been developed. Traditional numerical ocean models simulate SST by solving ocean dynamical equations but are computationally expensive and sensitive to uncertainties in atmospheric forcing, initial conditions, and subgrid-scale parameterizations (Zhao et al., 2026). Statistical methods offer greater computational efficiency but often struggle to capture the non-stationary behaviour and abrupt transitions associated with extreme events (Flint and Flint, 2012). More recently, data-driven approaches have emerged as promising alternatives, offering efficient multi-step SST forecasting and the potential to capture the spatial structures and temporal evolution of MHWs (Welandawe et al., 2025).

Among data-driven approaches, deep learning has shown considerable success in SST prediction (Welandawe et al., 2025; Zhao, 2024). Early studies primarily focused on single-point time-series forecasting using wavelet-based multiscale decomposition and recurrent neural networks (RNN) and long short-term memory networks (LSTM) (Mallat, 1989; Zhang et al., 2017).However, these methods often neglect spatial correlations and advective processes. To address this limitation, convolutional LSTM (ConvLSTM) networks have been developed to jointly capture spatial and temporal features, demonstrating promising performance in short-term SST prediction (Hao et al., 2023; Shi et al., 2015; Xiao et al., 2019). Subsequent studies have incorporated deformable attention to better represent spatial heterogeneity and sharp temperature gradients (Shi et al., 2024; Sun et al., 2024). Nevertheless, convolution-based models remain limited in capturing long-range dependencies and large-scale thermal forcing, particularly for extreme SST events. Transformer architectures offer a potential solution through global self-attention (Vaswani et al., 2017), but their quadratic computational complexity and limited preservation of fine-scale spatial structures pose challenges (Dai et al., 2024; Song et al., 2024). The Swin Transformer addresses computational efficiency through hierarchical shifted-window attention and has been applied to meteorological downscaling and bias correction (Liu et al., 2021; Xiang et al., 2022). However, its original architecture lacks explicit temporal modeling, limiting its direct applicability to spatiotemporal SST forecasting.

Beyond direct SST prediction, data-driven methods have been increasingly applied to MHW characterization and forecasting. Traditional machine learning methods, such as random forests, have been used to estimate MHW characteristics

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

(Giamalaki et al., 2022). In addition, three-dimensional U-Net architectures incorporating multiple physical variables have been developed for MHW intensity estimation and probabilistic forecasting (Bonino et al., 2024; Parasyris et al., 2025; Sun et al., 2023; Xie et al., 2024). Despite these advances, challenges remain in capturing complex spatiotemporal dependencies and accurately predicting extreme SST anomalies. In particular, the potential of short-term SST forecasts to improve threshold-based MHW detection in dynamically complex marginal seas, such as the ECS, remains insufficiently explored. This gap motivates the development of an integrated framework linking short-term SST forecasting with MHW detection.

In this study, we develop a hybrid deep learning framework, SwinTrans-ConvLSTM, for short-term daily SST forecasting and MHW detection in the ECS. The framework integrates ConvLSTM and the Swin Transformer to capture local spatiotemporal features and long-range spatial dependencies, while incorporating vector winds and air–sea temperature differences to represent air–sea interactions. Spatial-gradient regularization, online hard example mining (OHEM), and curriculum learning are employed to improve the prediction of SST spatial structures and extreme thermal anomalies. The predicted SST fields are subsequently used for threshold-based MHW detection following Hobday et al. (2016), enabling evaluation of MHW occurrence and spatiotemporal characteristics.The remainder of this paper is organized as follows. Section 2 describes the data, methodology, and evaluation metrics. Section 3 presents the results, Section 4 discusses the findings, and Section 5 summarizes the main conclusions.

2 Data and Methods

2.1 Study Area

The study area is located in the East China Sea (ECS; 28°–33°N, 120°–125°E), encompassing the Changjiang Estuary, Hangzhou Bay, the Zhoushan Archipelago, and coastal waters off Jiangsu and Shanghai (Fig. 1). The region features complex coastal geometry and variable bathymetry, with hydrographic conditions influenced by the Kuroshio intrusion, Taiwan Warm Current, Zhejiang–Fujian Coastal Current, and Changjiang Diluted Water. These interacting water masses and currents generate pronounced thermal fronts and seasonal SST variability (Hickox et al., 2000). Such spatial heterogeneity presents substantial challenges for high-resolution regional SST forecasting.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

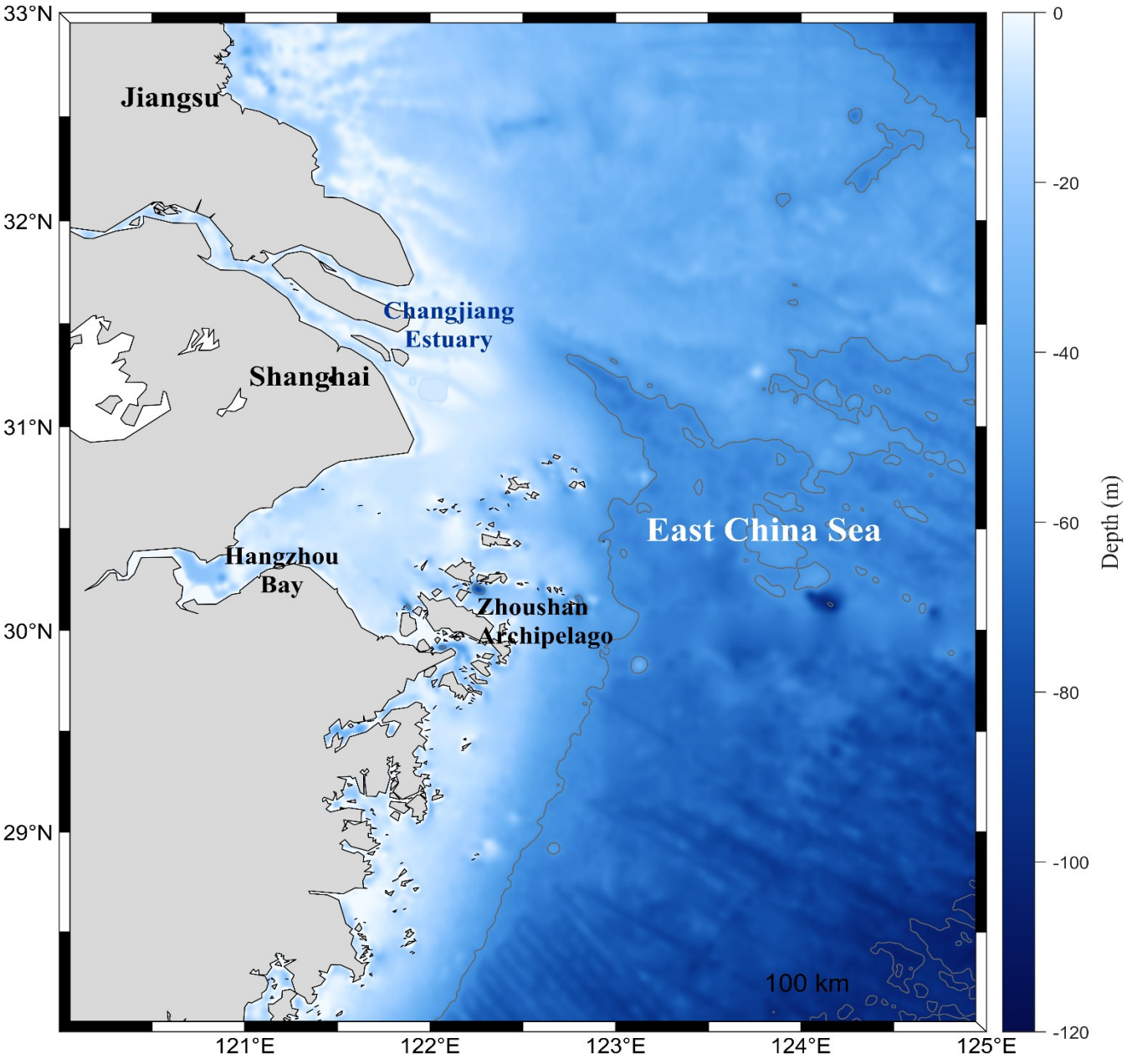


Figure 1. Bathymetric map of the study area in the coastal East China Sea (28–33° N, 120–125° E). Blue shading shows bathymetry (m), grey shading denotes land, and thin grey lines indicate bathymetric contours.

2.2 Data

This study used the Level-4 analyzed SST from the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system, provided by the Copernicus Marine Environment Monitoring Service (CMEMS) (Donlon et al., 2012). The OSTIA product combines satellite observations with in situ measurements to generate gap-free global daily SST fields at a spatial

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

resolution of 0.05° × 0.05°. The OSTIA SST fields were used as the prediction targets and verification references. Sea ice fraction and SST analysis error fields from the OSTIA dataset were also retained as auxiliary predictors.

Atmospheric predictors were obtained from the ECMWF ERA5 reanalysis dataset at hourly temporal resolution and 0.25° × 0.25° spatial resolution (Hersbach et al., 2020). The selected variables included 10-m zonal and meridional winds (u_{10} and v_{10}) and 2-m air temperature (T_{2m}). These variables were selected used for motivated predictors of SST variability.

To ensure spatiotemporal consistency, hourly ERA5 data were averaged daily and bilinearly interpolated onto the 0.05° OSTIA grid. The 1982–2024 dataset was chronologically divided into training (1982–2015), validation (2016–2019), and testing (2020–2024) sets. Climatologies and normalization statistics were derived exclusively from the training set to prevent data leakage. For each forecast origin, only atmospheric predictors available up to that day were used, excluding ERA5 data from the subsequent seven-day forecast period.

2.3 Input Features and Preprocessing

In addition to historical SST, input features were constructed to represent atmospheric forcing, recent thermal variability, seasonality, and geographic information.The 10-m zonal u_{10} and meridional v_{10} wind components as separate input predictors. These components were retained to capture the influence of atmospheric conditions on SST variability and MHW development (Holbrook et al., 2019).

The air–sea temperature differenceΔT was included to characterize surface thermal contrast associated with sensible heat exchange (Fairall et al., 2003), the ΔT was calculated as:

$$\Delta T(t) = SST(t) - T_{2m}(t) \quad (1)$$

where $\mathrm{SST(t)}$ and $\mathrm{T_{2m}(t)}$ denote SST and 2-m air temperature (°C) on input day (t), respectively. This feature represents thermal contrast rather than an explicit surface heat-flux estimate.

To capture recent atmospheric variability, three day trailing means of u_{10}, v_{10}, and ΔT were calculated as additional input channels. For each input day (t), these means used only days (t-2), (t-1), and (t) to provide smoothed summaries of recent conditions without introducing future information:

$$\frac{q(s) + q(s-1) + q(s-2)}{3}, \qquad q \in \{u_{10}, v_{10}, \Delta T\} \quad (2)$$

Additional predictors included SST anomalies, four wavelet-detail features, longitudinal and latitudinal SST gradients, one-day backward SST differences, sea ice fraction, and SST analysis error. Seasonality and geographic information were represented by sine–cosine annual-cycle encodings, three unit-sphere coordinates, scaled latitude and longitude, and normalized distance to the nearest non-ocean grid point. Together, these variables constituted the complete input feature set.

2.4 Marine Heatwave Identification and Metrics

MHWs were identified at each ocean grid point following Hobday et al. (2016), defined as periods when daily SST exceeded the seasonally varying 90th-percentile threshold for at least five consecutive days. Events separated by gaps of two days or

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

fewer were merged, with intervening gap days included as MHW days, but merging was not permitted across missing data. The climatological baseline was 1982–2011. Daily climatological means and 90th-percentile thresholds were calculated from OSTIA SST using an 11-day centered window and subsequently smoothed with a 31-day moving average (Hobday et al., 2016; Oliver et al., 2018). These reference curves were fixed across all models and forecast lead times and were distinct from the training-period climatology used for SST anomaly construction and normalization.

Four metrics were used to characterize MHWs (Table 1), including mean duration (average event length), mean intensity (average SST anomaly relative to the climatological mean over event days), occurrence frequency (number of events), and total MHW days (sum of event durations).

Table 1. Key metrics of MHWs and their formulations.

Metric	Formula	Unit
Mean duration	$D_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N} D_i$	Days
Mean intensity	$I_{\text{mean}} = \frac{1}{T_D}\sum_{k\in\Omega}\left(T_k - T_{\text{clim},k}\right)$	°C
Frequency	$N_{\text{MHW}} = N$	Events
Total MHW days	$T_D = \sum_{i=1}^{N} D_i$	Days

Notes: N denotes the event count, D_i the duration of event i in days, and Ω the set of event days. T_k and $T_{\text{clim},k}$ denote daily SST and its climatological mean, respectively. All metrics refer to the specified analysis period.

For each forecast lead time, predictions from successive forecast origins were aligned by their valid dates to construct a separate daily SST forecast series. The same fixed climatological mean, percentile threshold, and event-identification criteria were applied to these lead-specific series and the corresponding OSTIA reference series. Event characteristics were evaluated from these series rather than from individual seven-day forecast windows. Continuous SST forecast accuracy and MHW event-day detection were assessed separately using the metrics defined in Sect. 2.7.

2.5 Deep learning models

2.5.1 Architecture of SwinTrans-ConvLSTM

The proposed SwinTrans-ConvLSTM adopts an asymmetric encoder – decoder architecture (Fig. 2), combining ConvLSTM for local spatiotemporal feature extraction with Swin Transformer attention for long-range spatial dependency modeling. As shown in Fig. 2(a), the model takes 14 daily multichannel feature frames (Sect. 2.3) on a 100 × 100 grid as input. A single ConvLSTM layer with 64 hidden channels and 3 × 3 kernels encodes the temporal sequence while preserving spatial structure (Shi et al., 2015).

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

The final hidden state is projected to 96 channels through a 1 × 1 convolution, followed by a depth wise 3 × 3 convolution for positional encoding.The spatial attention module comprises two successive window-based attention blocks with 10 × 10 windows and three and six attention heads, respectively. The first uses regular window attention, while the second employs a five-grid-point cyclic shift to facilitate cross-window interactions. Both blocks incorporate layer normalization, relative positional bias, feed-forward networks with an expansion ratio of four, and residual connections.

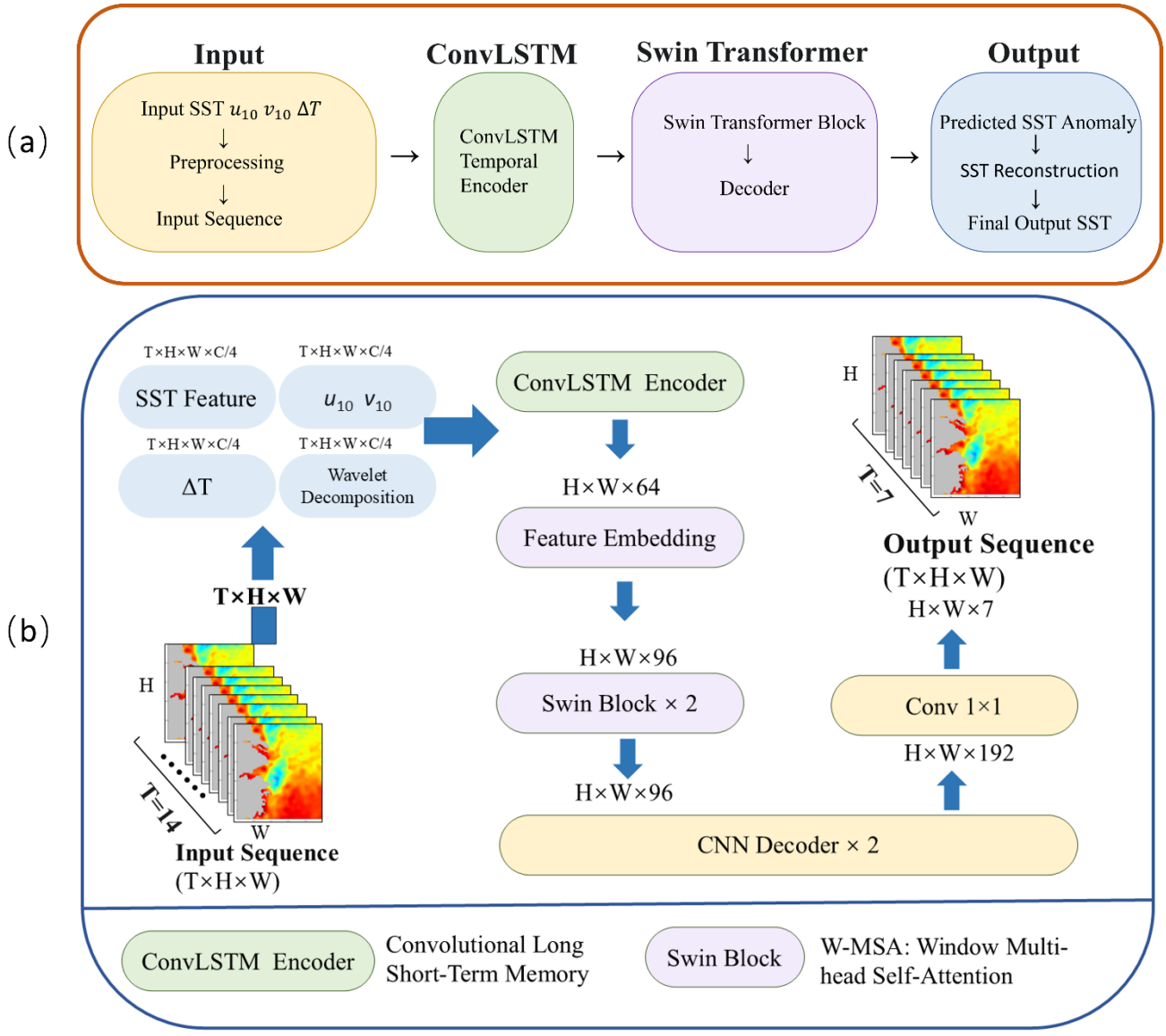


Figure 2. Architecture of the SwinTrans-ConvLSTM model. (a) Schematic overview of the model workflow, illustrating the complete data flow from multi-source input preprocessing, through temporal feature extraction and spatial feature modeling, to final SST reconstruction. (b) Detailed network architecture, showing the internal interactions among individual modules.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

The prediction head consists of a 3 × 3 convolution mapping 96 channels to 192 channels, a GELU activation, and a 1 × 1 convolution producing seven output channels. Each output channel corresponds to one forecast lead time. All seven fields are generated jointly in a single forward pass from the same historical input sequence, without recursively feeding earlier predictions into later forecast steps.

The network outputs standardized SST anomalies. During target preparation, daily SST anomalies are calculated relative to the training-period calendar-day climatology and standardized using their grid-point training means and standard deviations. Absolute SST predictions are reconstructed as

$$\hat{\boldsymbol{T}}_{t+n|t} = \boldsymbol{C}_{\mathrm{tr}}(d(t+n)) + \boldsymbol{\mu}_A + \boldsymbol{\sigma}_A \odot [G_\theta(\boldsymbol{Z}_{t-13:t})]_n, \quad n = 1, \ldots, 7 \tag{3}$$

Here, t denotes the forecast origin; $Z_{t-13:t}$ is the sequence of 14 multichannel input frames; and G_θ is the network with trainable parameters θ. The subscript n selects the output for lead day n. The function $d(t+n)$ identifies the calendar day of the target date, and $\boldsymbol{C}_{\mathrm{tr}}$ is the corresponding SST climatological field estimated from the training period. The fields $\boldsymbol{\mu}_A$ and $\boldsymbol{\sigma}_A$ are the training-period mean and standard deviation of SST anomalies at each grid point, and $\odot$ denotes element-wise multiplication. The resulting $\hat{\boldsymbol{T}}_{t+n|t}$ is the predicted daily SST field in degrees Celsius.

The training objective additionally incorporates spatial-gradient regularization, which compares the magnitudes of adjacent-grid SST differences in the predicted and reference fields. This term is intended to encourage agreement in local spatial variation and is treated as mathematical regularization. The complete training objective is described in Sect. 2.6.1.

2.5.2 Baseline models

To evaluate the effectiveness of the proposed SwinTrans-ConvLSTM model, four representative deep-learning models were selected for comparison: LSTM, ConvLSTM, Transformer, and Swin Transformer. SST persistence and an AR(1) model fitted to SST anomalies forecasts were additionally included to assess predictive skill relative to simple benchmarks that exploit the temporal persistence of SST.

LSTM models temporal dependence at each grid points independently without explicitly considering spatial correlations. ConvLSTM extends this framework by introducing convolutional operations to jointly model spatial and temporal features, but its receptive field is limited by kernel size, restricting large-scale dependency capture (Xiao et al., 2019). Transformer architectures use global self-attention to model long-range dependencies but suffer from high computational cost and potential loss of local structure in high-resolution gridded data (Zhao et al., 2024). The Swin Transformer reduces computational complexity through shifted-window attention while improving spatial dependency modeling; however, its lack of explicit recurrence limits its ability to represent continuous temporal evolution (Dai et al., 2024).

In contrast, the proposed SwinTrans-ConvLSTM integrates ConvLSTM-based temporal evolution modeling with Swin Transformer-based spatial representation, enabling unified learning of local dynamics and large-scale spatial dependencies.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

The persistence baseline retained the raw origin-day SST at all seven leads. The AR(1) baseline was fitted separately at each ocean grid point using daily SST anomalies from 1982–2015. Its calendar-day climatology was calculated from the training period and smoothed using a 31-day cyclic moving average. The intercept and lag-one coefficient were estimated by constrained least squares, with $|\varphi| \leq 0.999$ and at least 365 valid consecutive-day pairs required per grid point. Forecasts were calculated as

$$\begin{aligned} \hat{\mathrm{T}}^{\mathrm{P}}_{t+n|t} &= \mathrm{T}_t, \\ \hat{\mathrm{a}}_{t+n|t} &= \mathrm{c} + \varphi \odot \hat{\mathrm{a}}_{t+n-1|t}, \qquad n = 1, \ldots, 7. \\ \hat{\mathrm{T}}^{\mathrm{AR}(1)}_{t+n|t} &= \mathrm{C}_{\mathrm{B}}\big(d(t+n)\big) + \hat{\mathrm{a}}_{t+n|t}, \end{aligned} \tag{4}$$

Here, T_t denotes raw SST at the forecast origin, and C_{B} is the smoothed training-period climatology used by the AR(1) baseline. Forecast anomalies were initialized as $\hat{\mathrm{a}}_{t|t} = \mathrm{T}_t - \mathrm{C}_{\mathrm{B}}\big(d(t)\big)$. The fields c and φ contain the grid-point intercepts and lag-one coefficients, and $\odot$ denotes element-wise multiplication. Anomalies were propagated recursively without updating them from forecast-period observations. Training and checkpoint-selection procedures for the deep-learning models are described in Sect. 2.6.1.

2.6 Experimental Setup

2.6.1 Data preparation and training strategy

Model construction and training were implemented using the PyTorch deep-learning framework with CUDA acceleration on a high-performance computing platform equipped with an NVIDIA GeForce RTX 3090 GPU. Training samples were generated using a sliding-window strategy. The chronological partitions and preprocessing described in Sects. 2.2–2.3 were applied to avoid temporal information leakage, with all seven target dates of each sample strictly confined to a single subset.

Given the long-tailed distribution of marine heatwave occurrences approximately 5 %–10 % (Sun et al., 2024), standard data-driven models optimized solely on global mean errors frequently underestimate extreme values. To address this and preserve the spatial structural fidelity of SST fields, we designed a custom objective function termed the Gradient-Regularized Curriculum Loss. The total loss L_{total} at a given training epoch E is formulated as the sum of a base regression loss L_{base} and a spatial-gradient regularization term L_{grad}:

$$L_{total}(E) = L_{base}(E) + \lambda_{grad} L_{grad} \tag{5}$$

where λ_{grad} is a scaling hyperparameter (empirically set to 1.0) that controls the weight of the regularization term.

The base regression loss first computes the element-wise Huber loss (Huber, 1964) between the reference SST y_i and the predicted SST $\hat{y}_i$. Within each mini-batch, the index i represents a valid combination of forecast sample, lead day, and ocean grid point; missing targets and non-ocean points are excluded. To maintain robustness against extreme outliers while providing stable gradients for small errors, the Huber penalty ℓ_i is defined using $e_i = \hat{y}_i - y_i$:

$$\ell_i = \begin{cases} \frac{1}{2} {e_i}^2, & \text{if } |e_i| \leq \delta \\ \delta\left(|e_i| - \frac{1}{2}\delta\right), & \text{otherwise} \end{cases} \tag{6}$$

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

where δ is the threshold parameter set to 1.0 °C. Furthermore, to increase the contribution of summer prediction errors, a seasonal weighting factor was applied according to each target date:

$$w_i = \begin{cases} 2.0, & \text{if the target date falls in June, July, or August} \\ 1.0, & \text{otherwise} \end{cases} \tag{7}$$

The weighted loss for each valid element i is thus defined as

$$L_i = w_i \ell_i \tag{8}$$

An epoch-dependent curriculum learning strategy was adopted. With epochs numbered from 1, the first 20 epochs minimized the mean weighted loss over all N valid elements in each mini-batch. From epoch 21 onward, online hard example mining (OHEM) selected the highest-loss 20 % of valid elements (Shrivastava et al., 2016). Thus, the base regression loss $L_{\text{base}}(E)$ was formulated as a piecewise function:

$$L_{\text{base}}(E) = \begin{cases} \frac{1}{N}\sum_{i\in\mathcal{V}} L_i\,, & \text{if } 1 \le E \le 20 \\ \frac{1}{k}\sum_{i\in\mathcal{H}_E} L_i\,, & \text{if } E \ge 21 \end{cases} \tag{9}$$

where $N = |\mathcal{V}|$, $k = \max(1, \lfloor 0.2N \rfloor)$, and $\mathcal{H}_E$ is the index set of the selected highest-loss elements. Selection was based on weighted SST errors and could include both MHW and non-MHW conditions. OHEM was applied only to the base regression term; the spatial term continued to use all valid adjacent-grid pairs.

Finally, a spatial-gradient regularization term was included to encourage agreement in local SST variation (Mathieu et al., 2016). It penalized the L1 discrepancy between the magnitudes of adjacent-grid SST differences:

$$L_{\text{grad}} = \sum_{d\in x,y} \frac{1}{|\mathcal{E}_d|} \sum_{(i,j)\in\mathcal{E}_d} \left| |\hat{y}_j - \hat{y}_i| - |y_j - y_i| \right| \tag{10}$$

Here, $\mathcal{E}_d$ denotes valid adjacent-grid pairs in direction d, with both endpoints belonging to the same forecast field and having finite ocean reference values. The differences were computed along the longitudinal and latitudinal grid directions. This term is a mathematical spatial regularizer rather than a physical conservation constraint.

Model optimization used AdamW with a base learning rate of 3×10^{-5}, weight decay of 10^{-5}, and gradient-norm clipping at 1.0. Five epochs of linear learning-rate warm-up were followed by cosine decay. Training used automatic mixed precision and gradient accumulation, with a nominal effective batch size of 16 forecast windows. Mini-batch configurations varied between models, and OHEM selection was performed before gradient accumulation. Training was limited to 120 epochs, with early stopping after 30 epochs without improvement in validation MAE. The checkpoint with the lowest validation MAE was selected for testing.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

2.6.2 Forecasting strategy and evaluation setup

A daily rolling forecasting strategy was applied during testing. At each forecast origin (t), the models used the preceding 14-day input sequence to jointly predict SST for the next seven days, with fixed model parameters and no future predictor information. Predictions were reconstructed as absolute SST using Eq. (3), yielding a tensor of shape$[N_{\text{windows}}, 7, H, W]$, where $N_{windows}$ is the number of forecast windows and H and W are the spatial dimensions. For all deep-learning models, test-time augmentation was performed by averaging predictions from the original and horizontally reflected inputs after restoring their spatial orientation. Averaging was conducted in standardized anomaly space before SST reconstruction. Persistence and AR(1) forecasts were generated using Eq. (4).

SST forecasts were evaluated against the original daily OSTIA fields without temporal filling or wavelet smoothing. Comparisons used common forecast origins and valid ocean grid-point–target pairs. MAE, RMSE, R^2, and PCC were calculated for individual leads and all seven leads (Sect. 2.7), with pooled RMSE computed directly from all retained squared errors. For MHW evaluation, lead-specific daily SST series were constructed by aligning rolling forecasts with their target dates, without averaging across leads. Predicted and reference series were subjected to identical MHW detection criteria (Sect. 2.4). Event duration, mean intensity, occurrence frequency, and total MHW days were compared over the common evaluation period, while detection scores were calculated using matched valid grid-point–date samples.

The architecture and training configuration of SwinTrans-ConvLSTM are summarized in Table 2, with training and checkpoint selection described in Sect. 2.6.1.

Table 2. Model architecture and training configuration of the SwinTrans-ConvLSTM.

Tensor Shape	Input Shape	[B,14,25,100,100]
	Output Shape	[B,7,1,100,100]
Model Architecture	Hidden Channels	64 (ConvLSTM), 96 (Swin)
	Kernel Size	3×3
	Encoder Layers	2
	Num. Heads	[3,6]
	Window Size	10×10
	FFN ratio	4
Training Settings	Max Epochs	120
	Batch Size	4 during Epochs 1–3; 8 from Epoch 4 onwards
	Optimizer	AdamW
	Initial LR	3×10^{-5}

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

LR Scheduler	Cosine Annealing Warm Restarts
Weight Decay	1×10^{-5}
Early Stopping	30 Epochs
OHEM Ratio	Top 20 %

2.7 Evaluation metrics

To quantitatively evaluate the prediction performance of the proposed SwinTrans-ConvLSTM model, four statistical metrics commonly used in SST forecasting studies were adopted. All metrics were calculated over valid ocean grid points in the test set. Predictions were compared with the original daily OSTIA SST reference fields using the evaluation setup described in Sect. 2.6.2.

The mean absolute error (MAE) measures the average absolute difference between predicted and observed SST, providing an intuitive estimate of prediction accuracy.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{11}$$

The root mean square error (RMSE) assigns a squared penalty to large errors and is therefore sensitive to extreme deviations. It is particularly useful for evaluating the model's ability to capture extreme SST variations associated with MHWs (Sun et al., 2024).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{12}$$

The Pearson correlation coefficient (PCC) was used to quantify the linear agreement between predicted and observed SST, reflecting phase consistency across temporal scales.

$$\mathrm{PCC} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})\left(\hat{y}_i - \bar{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{13}$$

The coefficient of determination (R^2) quantifies the proportion of variance explained by the model, with values closer to 1 indicating better performance.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{14}$$

Here, N is the number of valid gridpoint–forecast pairs included in the calculation; y_i and $\hat{y}_i$ denote reference and predicted SST, respectively; and $\bar{y}$ and $\bar{\hat{y}}$ are their corresponding means. Metrics were calculated separately for each lead and pooled across leads 1–7. Pooled PCC describes association across all evaluated pairs and should not be interpreted as a separate

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

assessment of temporal phase agreement at individual grid points. A perfect prediction gives $R^2 = 1$, whereas negative values indicate larger squared errors than a constant prediction equal to the mean reference SST over the evaluated pairs.

For MHW detection, a positive case was a grid-point day belonging to an event identified according to Sect. 2.4. A hit (TP) occurred when both the forecast and the reference classified the day as an event day, a false alarm (FP) when only the forecast did so, a miss (FN) when only the reference did so, and a correct rejection (TN) when neither did so. The counts were pooled over the common valid grid-point and date sample before calculating the detection scores for each forecast lead.

$$POD = H = \frac{TP}{TP+FN}, F = \frac{FP}{FP+TN} \tag{15}$$

The false alarm ratio (FAR) is the proportion of forecast MHW event days that were not classified as event days in the reference.

$$FAR = \frac{FP}{TP+FP} \tag{16}$$

FAR and the false alarm rate F have different denominators and are not interchangeable. To evaluate rare binary MHW event-day detection, the Symmetric Extremal Dependence Index (SEDI) was introduced following Ferro and Stephenson (Ferro and Stephenson, 2011):

$$\mathrm{SEDI} = \frac{\ln F - \ln H + \ln(1-H) - \ln(1-F)}{\ln F + \ln H + \ln(1-H) + \ln(1-F)} \tag{17}$$

SEDI ranges from −1 to 1, with higher values indicating better performance. The expression in Eq. (17) applies when $0 < H < 1$ and $0 < F < 1$. POD, FAR, and SEDI were calculated separately for each lead using the common evaluation samples. These scores assess MHW event-day detection; the evaluation of MHW event characteristics follows the duration-based identification procedure in Sect. 2.4.

3 Results

3.1 SST forecast performance

3.1.1 Overall prediction Accuracy

The SwinTrans-ConvLSTM model was evaluated against four deep-learning benchmarks (LSTM, ConvLSTM, Transformer, and Swin Transformer; Sect. 2.5.2) and two reference methods (persistence and AR(1)). All methods were assessed against OSTIA SST over 2020–2024 using common forecast origins and valid ocean grid points. Table 3 summarizes the pooled performance across forecast leads of 1–7 days.

SwinTrans-ConvLSTM achieved the lowest pooled MAE (0.501 °C) and RMSE (0.708 °C) among all evaluated models, with an R^2 of 0.9881 and a PCC of 0.9941. Compared with persistence, SwinTrans-ConvLSTM reduced MAE and RMSE by 22.4 % and 21.0 %, respectively. The corresponding MAE and RMSE reductions relative to AR(1) were 6.4 % and 4.2 %. Among the four deep-learning benchmarks, Swin Transformer achieved the best performance, with an MAE of 0.5109 °C

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

and an RMSE of 0.7118 °C. Nevertheless, compared with Swin Transformer, SwinTrans-ConvLSTM further reduced MAE and RMSE by 1.9 % and 0.5 %, respectively. Larger improvements were observed relative to ConvLSTM, LSTM, and Transformer, with MAE reductions ranging from 2.5 % to 4.7 % and RMSE reductions from 2.8 % to 3.3 %.

The pooled metrics demonstrate the strong overall SST forecasting performance of SwinTrans-ConvLSTM. The high PCC of persistence (0.9904) also highlights the importance of evaluating forecast accuracy using both correlation- and error-based metrics. To provide a more comprehensive assessment, forecast performance across different lead times, spatial regions, and seasons is further examined in Sects. 3.2–3.4. Extreme-temperature prediction and MHW detection performance are evaluated separately in Sect. 3.5.

Table 3. SST forecast performance on the test set, pooled across forecast leads of 1–7 days.

Model	MAE (°C)	RMSE (°C)	R^2	PCC
SwinTrans-ConvLSTM	0.501	0.708	0.9881	0.9941
ConvLSTM	0.511	0.712	0.9879	0.9941
LSTM	0.514	0.729	0.9874	0.9938
Swin-Transformer	0.522	0.731	0.9873	0.9938
Transformer	0.526	0.732	0.9872	0.9937
Persistence	0.535	0.739	0.987	0.994
AR(1)	0.646	0.897	0.9809	0.9904

3.1.2 Lead-Time Dependence of Prediction Errors

Figure 3 and Table 4 show the dependence of SST forecast errors on lead time over the 2020–2024 test period. Both MAE and RMSE increased with lead time for all seven methods. For SwinTrans-ConvLSTM, MAE increased from 0.2878 °C at a 1-day lead to 0.6378 °C at a 7-day lead, while RMSE increased from 0.4361 to 0.8620 °C. The corresponding increases were 0.3500 and 0.4259 °C, compared with 0.6614 and 0.8042 °C for persistence and 0.4340 and 0.5325 °C for AR(1), respectively.

The ranking of the methods varied with lead time and error metric. At the 1-day lead, AR(1) achieved the lowest MAE and RMSE, at 0.2732 and 0.3966 °C, respectively. SwinTrans-ConvLSTM achieved the lowest MAE at every lead from 2 to 7 days. For RMSE, Swin-Transformer achieved the lowest values at leads of 2, 3, 4, and 6 days, whereas SwinTrans-ConvLSTM achieved the lowest values at leads of 5 and 7 days. The two models had nearly identical RMSE at the 6-day lead, at 0.8255 °C for SwinTrans-ConvLSTM and 0.8253 °C for Swin-Transformer.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

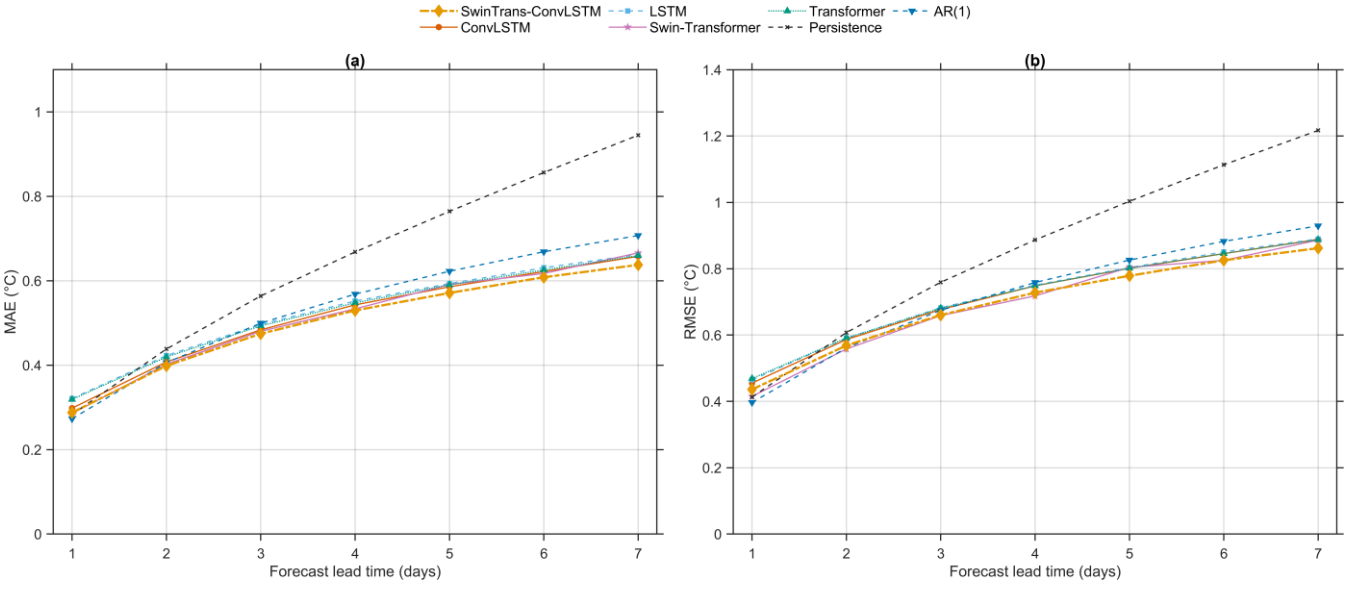


Figure 3. Lead-time dependence of SST forecast errors for the seven methods over the 2020–2024 test period: (a) MAE and (b) RMSE. Metrics were calculated using common forecast origins, ocean grid points, and valid prediction–reference pairs across the seven methods.

At the 7-day lead, SwinTrans-ConvLSTM had the lowest MAE and RMSE among all seven methods. Its MAE of 0.6378 °C and RMSE of 0.8620 °C represented reductions of 32.5 % and 29.2 %, respectively, relative to persistence (0.9449 and 1.2170 °C). Relative to AR(1) (0.7072 and 0.9292 °C), the corresponding reductions were 9.8 % and 7.2 %. The reductions relative to Swin-Transformer were smaller, at 4.2 % for MAE and 2.7 % for RMSE. These comparisons show that the error differences at the longest lead were substantially larger relative to persistence than relative to the standalone Swin-Transformer.

Table 4. SST forecast errors at leads of 1–7 days over the 2020–2024 test period.

Model	1-DAY		2-DAY		3-DAY		4-DAY		5-DAY		6-DAY		7-DAY	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
SwinTrans-ConvLSTM	0.288	0.465	0.399	0.569	0.475	0.660	0.530	0.728	0.571	0.779	0.608	0.826	0.638	0.862
ConvLSTM	0.298	0.415	0.408	0.587	0.484	0.677	0.543	0.748	0.586	0.801	0.622	0.846	0.658	0.888
LSTM	0.321	0.469	0.423	0.591	0.497	0.682	0.552	0.751	0.593	0.803	0.631	0.851	0.662	0.890
Swin-Transformer	0.286	0.413	0.402	0.558	0.480	0.659	0.534	0.718	0.591	0.804	0.617	0.825	0.666	0.886
Transformer	0.319	0.397	0.420	0.591	0.493	0.681	0.548	0.748	0.590	0.802	0.626	0.849	0.660	0.888
Persistence	0.283	0.465	0.438	0.608	0.564	0.759	0.668	0.887	0.765	1.003	0.856	1.113	0.945	1.217

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

Model	1-DAY		2-DAY		3-DAY		4-DAY		5-DAY		6-DAY		7-DAY	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
AR(1)	0.273	0.415	0.406	0.562	0.500	0.675	0.568	0.759	0.623	0.826	0.669	0.883	0.707	0.929

Figure 4 presents the signed error distributions of the five deep-learning models, with errors defined as predicted SST minus OSTIA SST. Error distributions broadened with increasing forecast lead time, accompanied by declining proportions of predictions within fixed error bounds. For SwinTrans-ConvLSTM, the proportions of errors within ±0.1, ±0.3, and ±0.5 °C decreased from 29.3 %, 66.9 %, and 83.6 % at day 1 to 11.8 %, 33.3 %, and 51.4 % at 7-day lead, respectively. At day 7, SwinTrans-ConvLSTM retained a higher proportion of errors within ±0.5 °C was 49.0 %–50.1 %. Together, Figure. 3 and 4 demonstrate the competitive short-term forecasting performance of SwinTrans-ConvLSTM, particularly at longer forecast leads.

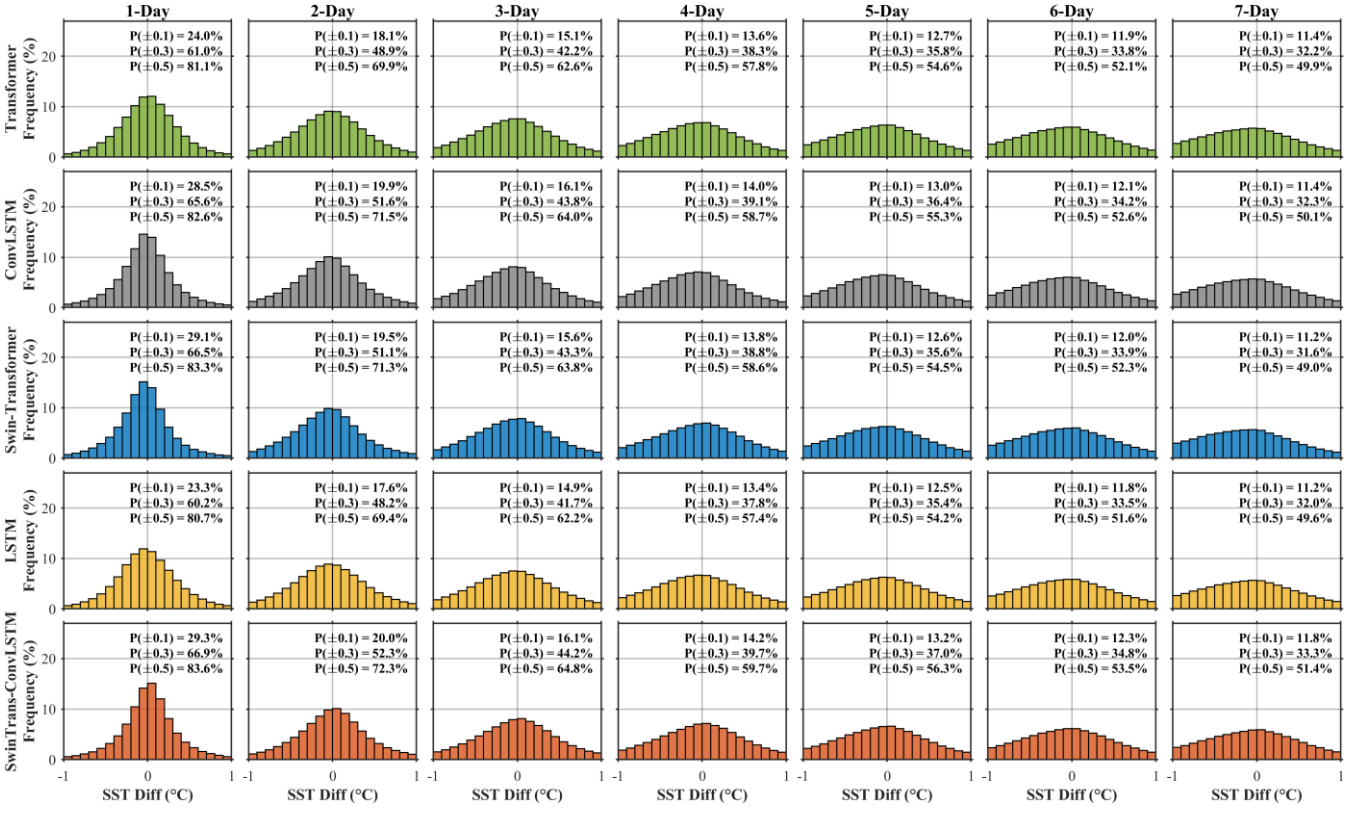


Figure 4. SST forecast error distributions for the five deep-learning models at leads of 1–7 days over the 2020–2024 test period. Rows represent models and columns represent forecast leads. Errors are defined as predicted SST minus daily OSTIA SST. Bars show percentages per 0.10 °C bin. Annotations indicate the percentages of forecasts with absolute errors no greater than 0.1, 0.3, and 0.5 °C.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

Overall, the results in Fig. 3 to Fig. 4 indicate that the SwinTrans-ConvLSTM model achieves superior predictive performance compared with benchmark models. It also demonstrates improved stability in long-term forecasting, reduced error accumulation, and better control of large prediction deviations. These results highlight the effectiveness of the proposed hybrid spatiotemporal architecture for complex ocean forecasting tasks.

3.1.3 Spatial and Temporal Variability in Prediction Errors

Figures 5 and 6 present the spatial distributions of MAE and RMSE, respectively, for the five deep-learning models at forecast leads of 1–7 days over the 2020–2024 test period. Both metrics were calculated at each ocean grid point using the corresponding daily OSTIA SST as the reference, with consistent colour scales across panels within each figure.

The five models exhibited broadly similar spatial error patterns, with errors generally increasing as forecast lead time extended. Lower MAE and RMSE were observed over much of the southeastern offshore region, whereas higher errors were concentrated in the northern and western coastal waters, particularly at leads of 5–7 days. RMSE showed spatial patterns consistent with MAE but placed greater emphasis on larger individual prediction errors. These results reveal substantial spatial variability in forecast accuracy and indicate that error growth with lead time differs across the study area.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

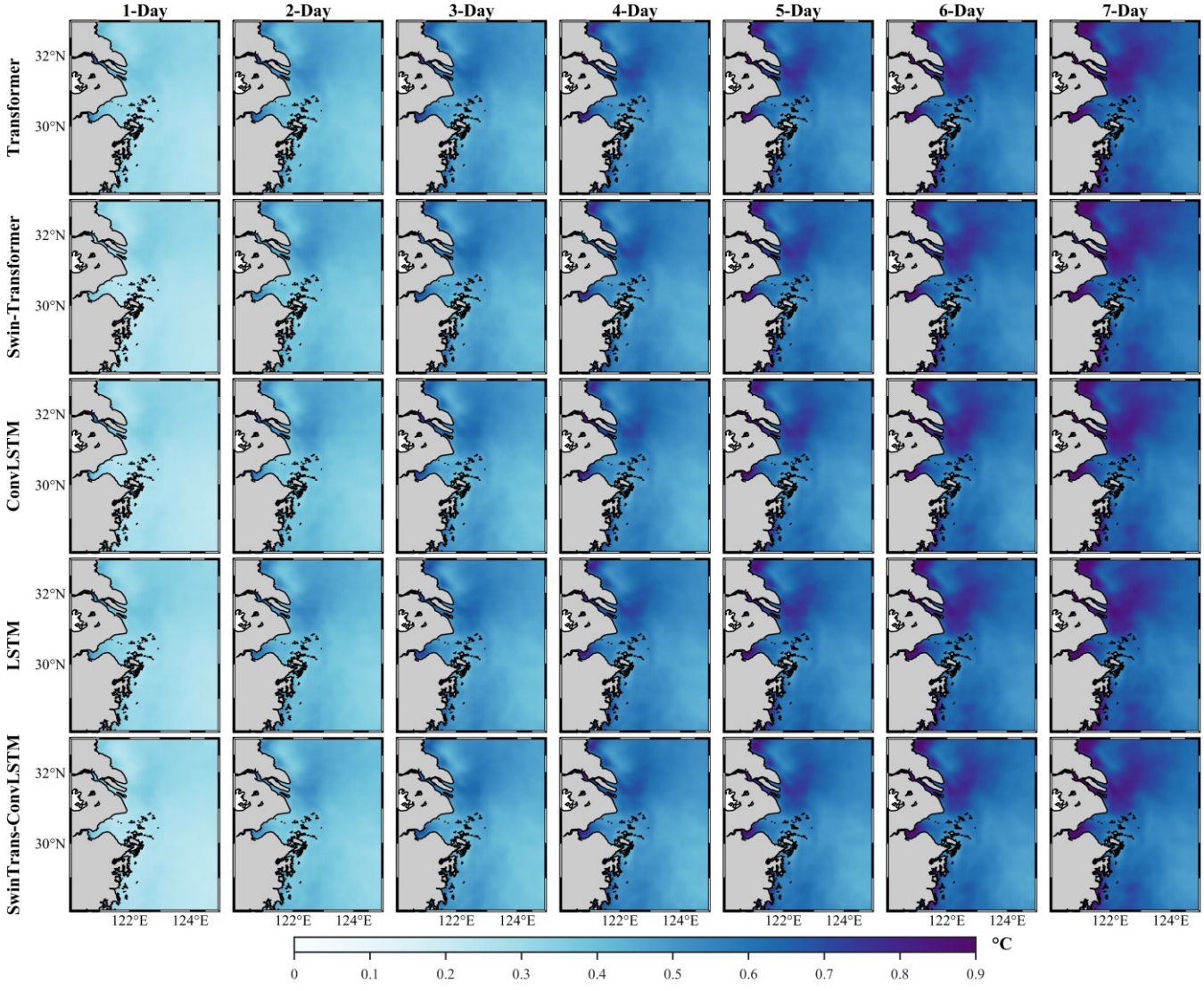


Figure 5. Spatial distributions of SST MAE for the five deep-learning models at forecast leads of 1–7 days over the 2020–2024 test period. Rows represent models and columns represent forecast leads. MAE was calculated at each ocean grid point across forecast origins.

SwinTrans-ConvLSTM achieved the lowest pooled MAE and RMSE among the evaluated methods (Table 3), with spatial differences in model performance varying across regions and forecast leads. The shared regions of elevated error suggest that the northern and coastal waters remain challenging for all five models. These spatial results complement the overall and lead-specific evaluations in Tables 3 and 4 and identify areas where further improvements in SST forecasting are needed.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

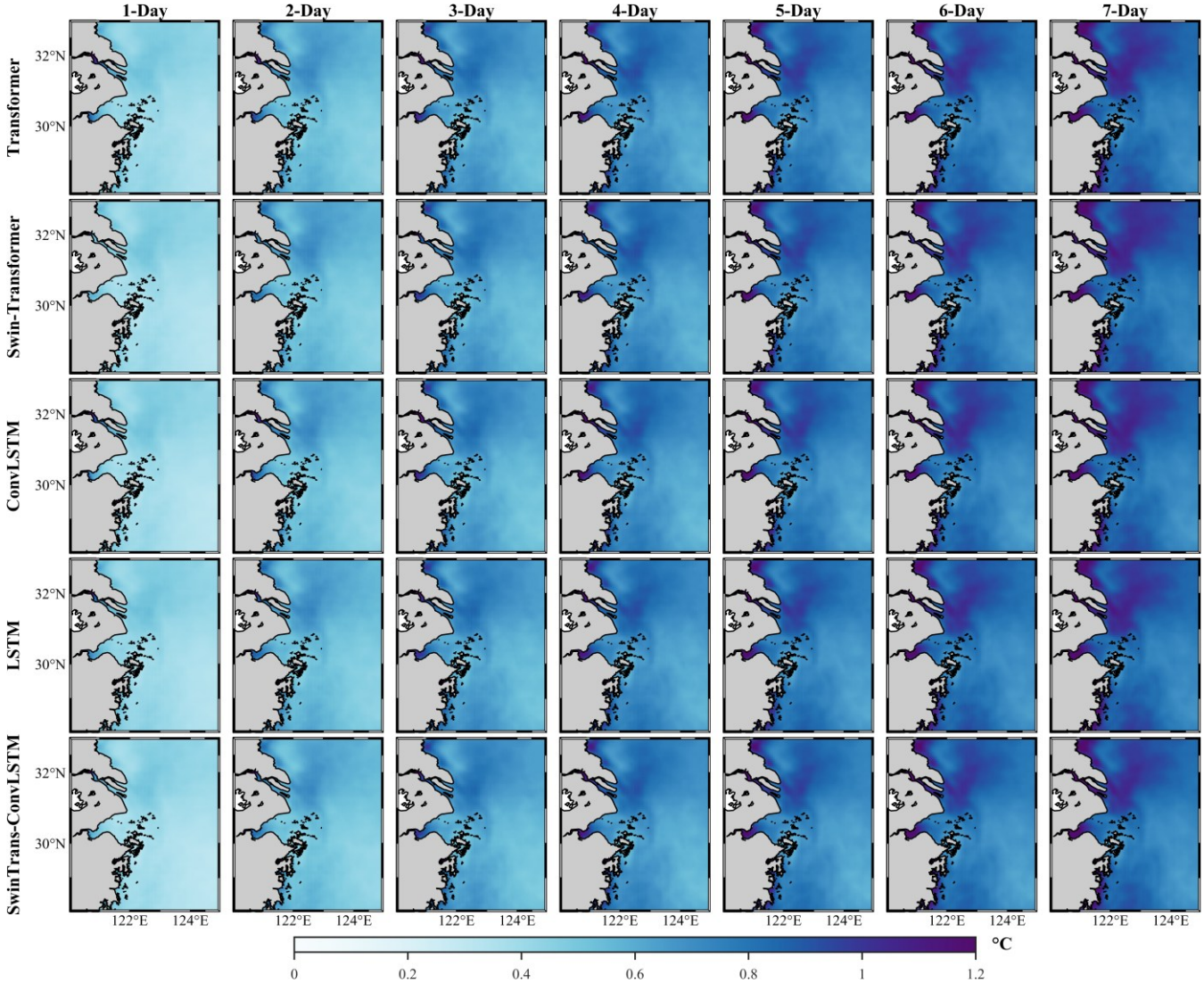


Figure 6. Spatial distribution of SST forecast RMSE over the 2020–2024 test period. Rows show the five deep-learning models, and columns show forecast leads of 1–7 days. RMSE was calculated across forecast origins at each ocean grid point.

Figure 7 presents the monthly MAE and RMSE of the seven forecasting methods over the 2020–2024 test period, pooled across forecast leads of 1–7 days according to the target month. The five deep-learning models exhibited similar monthly error patterns, with the highest MAE and RMSE in August and the lowest in October. For SwinTrans-ConvLSTM, MAE ranged from 0.40 °C in October to 0.607 °C in August, and RMSE ranged from 0.570 to 0.835 °C. Errors generally increased in April, declined in May and June, and rose again in July and August.

SwinTrans-ConvLSTM achieved the lowest MAE in eight months (January–March, June, and August–November) and the lowest RMSE in seven months (January–March and August–November). Swin Transformer performed best in terms of MAE in April, May, and July and RMSE from April to July, whereas ConvLSTM achieved the lowest MAE and RMSE in

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

December. These results demonstrate the strong monthly prediction performance of SwinTrans-ConvLSTM, with model rankings varying across months and error metrics.

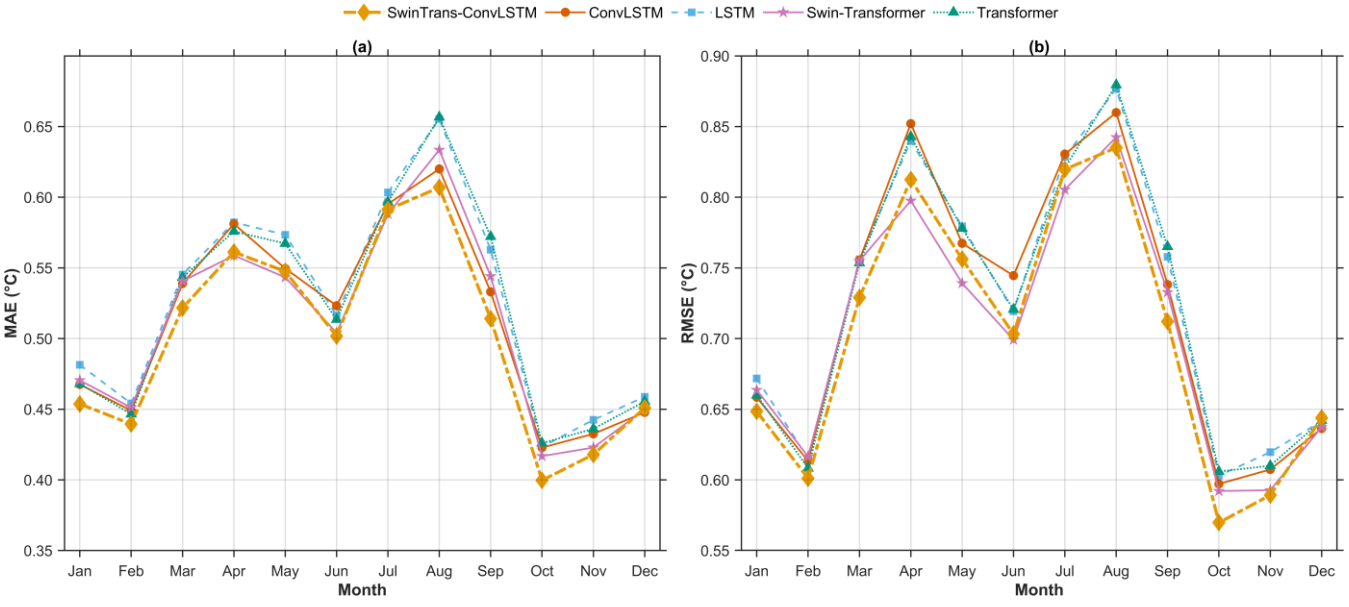


Figure 7. Monthly variation in SST forecast errors for the five deep-learning models over the 2020–2024 test period: (a) MAE and (b) RMSE. Errors were grouped by the calendar month of the target date and pooled across forecast leads of 1–7 days.

3.2 MHW Detection Based on Predicted SST

3.2.1 MHW intensity and related SST warm-anomaly patterns

MHW detection based on predicted SST was evaluated using the reference climatology, percentile threshold, and event criteria described in Sect. 2.4. For lead-specific analyses, forecasts from successive initialization dates were aligned by target date to construct daily SST series at each lead. The spatial analysis in Fig. 10 instead used all seven forecast fields from a single initialization.

Figure 8 illustrates the 1-day forecasts of regional mean SST in the nearshore waters of the ECS (121–122°E, 29–30°N) during 2022. The model reproduced the broad seasonal evolution and the prolonged summer warm conditions relative to the regional climatology and 90th-percentile threshold. Differences remained during rapid warming episodes in spring and at several summer SST peaks, for which the predicted temperatures were lower than the observations. The shaded intervals identify threshold exceedances belonging to model-diagnosed regional MHWs. This regional example illustrates the evolution of warm conditions, while the following grid-point analyses assess intensity, event-day detection and event characteristics.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

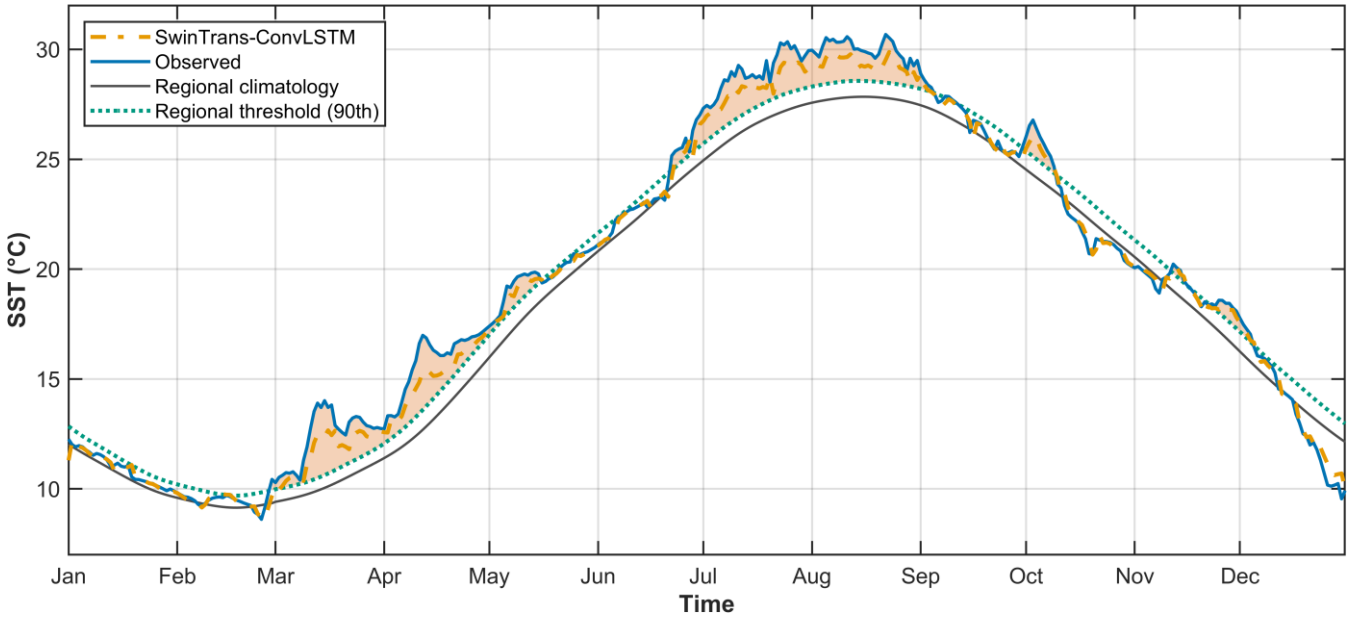


Figure 8. Regional mean SST and diagnosed MHW conditions in the nearshore ECS (121–122°E, 29–30°N) during 2022 at a fixed forecast lead of 1 day. Blue and dashed orange curves show daily OSTIA SST and SwinTrans-ConvLSTM forecasts, respectively. The black curve and dotted green curve show the regional climatological mean and 90th-percentile threshold estimated from the 1982–2011 reference period. The regional threshold was estimated from the regional mean SST series. Light orange shading denotes predicted SST above the threshold on model-diagnosed regional MHW event days. The forecast curve combines successive forecast origins at the same 1-day lead.

To assess MHW peak intensity across forecast lead times, observed and predicted events were compared at individual ocean grid points over 2020–2024 (Figure 9). At each lead, each observed event was matched to the first temporally overlapping predicted event at the same grid point. Peak intensity was defined as the maximum daily SST anomaly relative to the climatological mean during an event. The Pearson correlation coefficient (PCC) decreased from 0.906 at a 1-day lead to 0.842 at 3 days and 0.649 at 7 days. The corresponding root mean square errors (RMSEs) increased from 0.393 to 0.661 and 1.219 °C, and mean absolute errors (MAEs) from 0.194 to 0.474 and 0.993 °C, respectively. At the 7-day lead, SwinTrans-ConvLSTM tended to underestimate the largest observed peaks.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

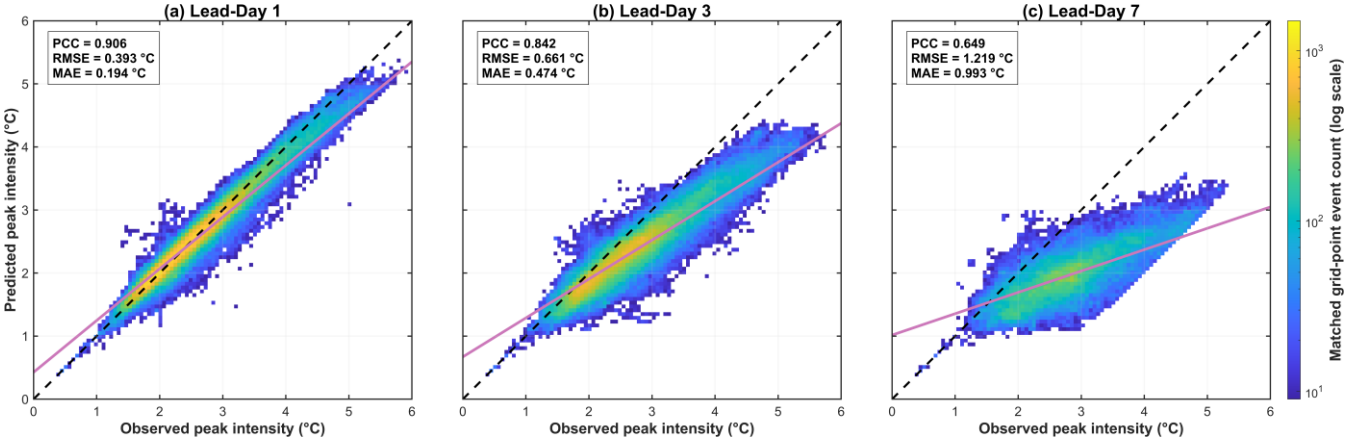


Figure 9. Observed versus SwinTrans-ConvLSTM predicted MHW peak intensity at leads of (a) 1, (b) 3 and (c) 7 days for 2020–2024. Colours show log-scaled counts of matched grid-point event pairs; the dashed line is 1:1 and the magenta line is a linear fit. Statistics are conditional on matched events.

Figure 10 compares the observed and predicted spatial distribution of MHW intensity for a selected seven-day case, 23–29 June 2021, using a single forecast issued on 22 June 2021 (leads 1–7). The mapped quantity is the mean SST anomaly over threshold-exceedance days within this period, which differs from the full-event peak intensity in Fig. 9. All five models reproduced the broad pattern of stronger anomalies in the northern part of the domain and weaker, spatially variable warm features farther south, but they underestimated the extent of the strongest anomalies (≥3 °C) in the north. Transformer, Swin-Transformer and SwinTrans-ConvLSTM also produced a spurious warm spot near the Yangtze Estuary (about 122.5°E, 31.3°N) that is absent from the observations. SwinTrans-ConvLSTM nevertheless had the highest spatial PCC (0.969) and the lowest MAE (0.193 °C) and RMSE (0.267 °C). The benchmark PCCs ranged from 0.957 to 0.964, MAEs from 0.270 to 0.321 °C and RMSEs from 0.356 to 0.450 °C, so the MAE and RMSE of SwinTrans-ConvLSTM were about 10% lower than those of the best benchmark (Transformer).

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

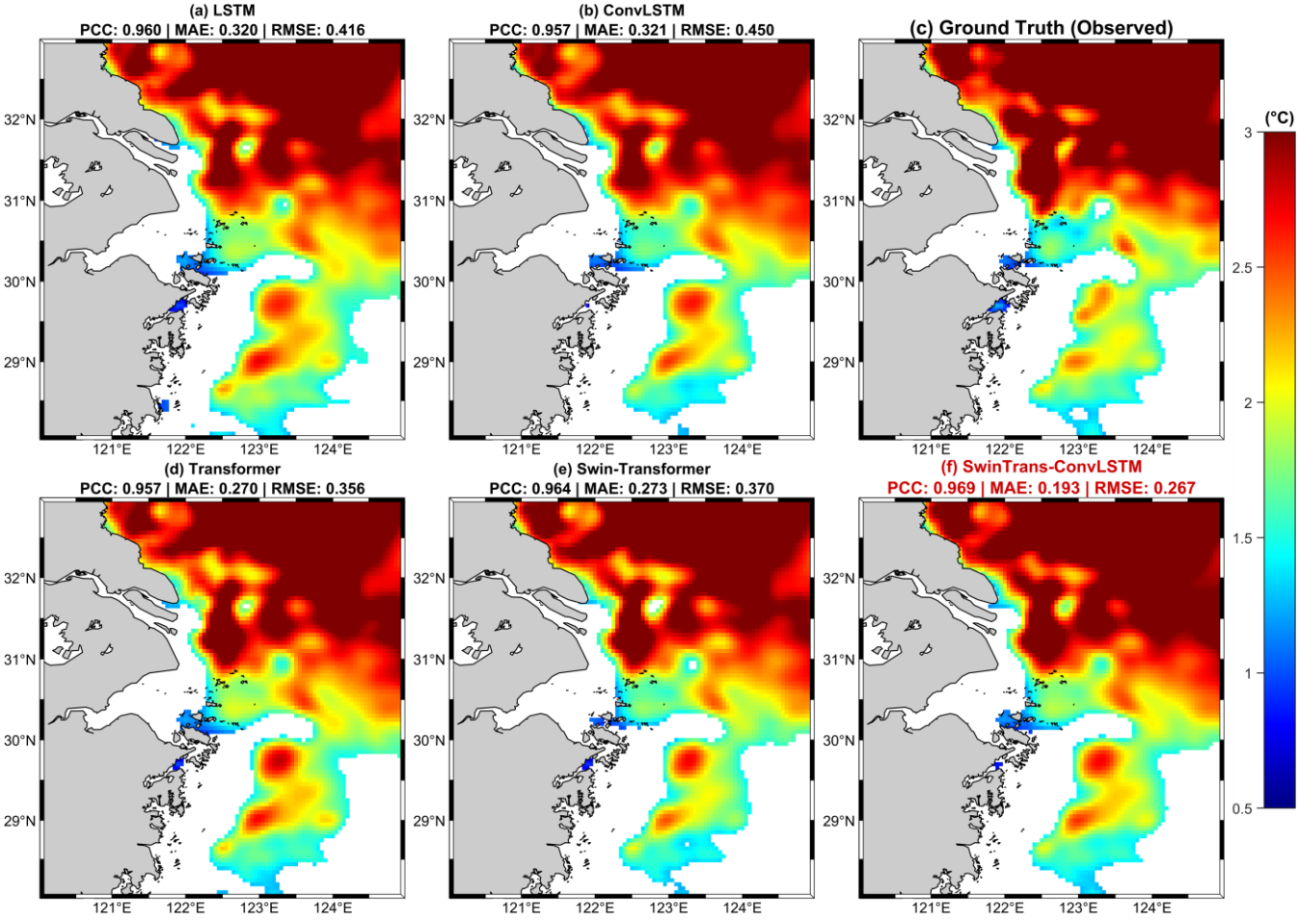


Figure 10. Spatial comparison of mean MHW intensity over the seven days 23–29 June 2021: (a) LSTM, (b) ConvLSTM, (c) observed OSTIA SST, (d) Transformer, (e) Swin-Transformer and (f) SwinTrans-ConvLSTM. Intensity is the SST anomaly relative to the 1982–2011 climatological mean, averaged over days on which SST exceeds the 90th-percentile threshold. PCC, MAE and RMSE compare each model with the observed field. Grey, land; white, no threshold exceedance.

3.2.2 MHW Event-Day Detection Performance

To assess MHW event-day detection skill, the symmetric extremal dependence index (SEDI) was calculated for all seven methods at forecast leads of 1–7 days over 2020–2024 (Fig. 11), using pooled contingency counts from common valid grid-point days as described in Sect. 2.7.

For SwinTrans-ConvLSTM, SEDI decreased from 0.9511 at a 1-day lead to 0.6590 at 7 days. Persistence achieved the highest 1-day SEDI (0.9564), followed by AR(1) (0.9526), whereas SwinTrans-ConvLSTM ranked first among all seven methods at leads of 2–7 days. At the 7-day lead, its SEDI exceeded those of persistence (0.5845) and AR(1) (0.5794) by 0.0745 and 0.0796, respectively.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

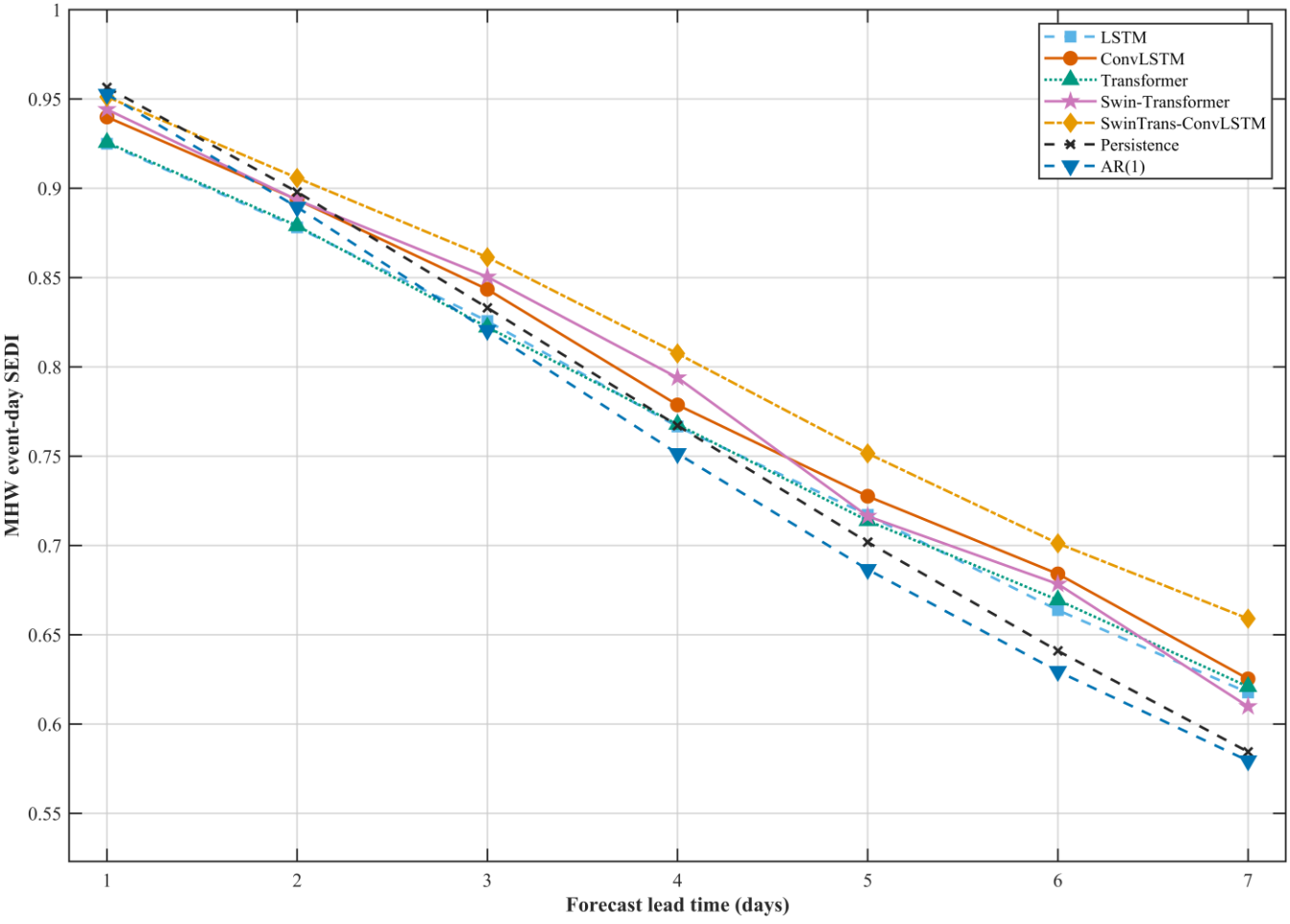


Figure 11. MHW event-day SEDI at forecast leads of 1–7 days for five deep-learning models, persistence and AR(1), evaluated over the common valid 2020–2024 grid-point-day sample. Observed and predicted MHW event-day labels were derived using the same climatology, threshold and duration/gap criteria (Sect. 2.4), and SEDI was calculated as described in Sect. 2.7. Higher values indicate better event-day detection skill.

The SEDI comparison evaluates whether the forecasts identified the correct MHW days. Table 5 complements this assessment by comparing how well the five deep-learning models reproduced mean duration, mean intensity, event count and total MHW days at the 1-day lead. Table 5 extends the evaluation to the four MHW characteristics introduced in the methods: mean duration, mean intensity, event count and total MHW days. The table reports 1-day forecasts over the common eligible 2020–2024 sample. Each characteristic was calculated separately for the observed and predicted event catalogues at each grid point; the absolute differences were then averaged spatially. Event count denotes the number of complete eligible events over this period, rather than an annual occurrence rate. Mean intensity is weighted by event duration. Mean-duration and mean-intensity errors use a common event-positive grid-point sample, while event-count and total-day errors retain eligible zero-event cases.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

SwinTrans-ConvLSTM achieved mean-duration, mean-intensity, event-count and total-MHW-day MAEs of 1.6617 days, 0.0644 °C, 2.1478 events and 34.5661 days, respectively. It had the lowest errors for mean intensity, event count and total MHW days among the five models. Relative to the lowest benchmark MAE for each of these characteristics, the reductions were 20.4 %, 18.4 % and 18.3 %, respectively.

Table 5. Grid-point mean absolute errors of MHW characteristics at a 1-day forecast lead over the common eligible 2020–2024 test-period sample.

Model	Mean duration MAE (days)	Mean intensity MAE (°C)	Event count MAE (events)	Total MHW days MAE (days)
LSTM	1.9487	0.0951	3.1786	42.3337
ConvLSTM	2.0571	0.0809	2.6334	48.5619
Transformer	1.9733	0.1381	3.3431	47.7118
Swin-Transformer	1.4983	0.0878	2.9543	59.5884
SwinTrans-ConvLSTM	1.6617	0.0644	2.1478	34.5661

3.2.3 MHW Event Characteristics

To examine the spatial distribution of MHW characteristic errors, Figure 12 presents the four error maps for SwinTrans-ConvLSTM at the 1-day lead, using the same evaluation sample as Table 5. Mean-duration and mean-intensity errors were higher at several coastal and estuarine locations, whereas event-count and total-MHW-day errors also extended across offshore and southern areas.

These patterns indicate spatial differences in the reproduction of MHW intensity, duration, and occurrence. Spatial smoothing and interpolation were applied for visualization only; all MAEs were calculated from the original grid-point values. The maps illustrate error distributions for SwinTrans-ConvLSTM without establishing a spatial ranking among models or identifying the underlying physical causes.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

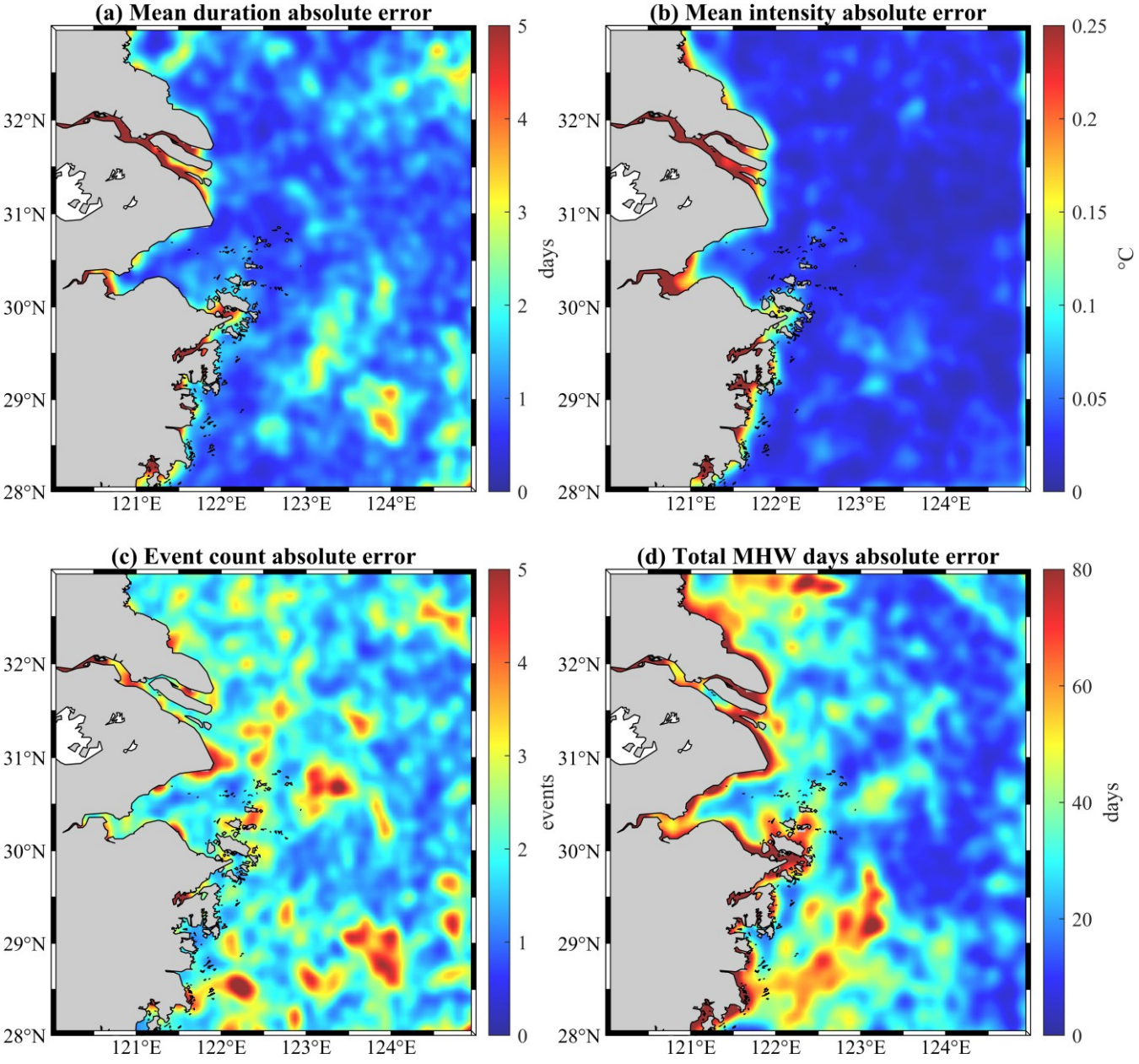


Figure 12. Absolute errors between SwinTrans-ConvLSTM and OSTIA SST in grid-point MHW characteristics, at a 1-day forecast lead over the common eligible 2020–2024 sample: (a) mean duration, (b) duration-weighted mean intensity, (c) event count and (d) total MHW days. Event count and total days refer to complete eligible events over the evaluation period, not annual rates. Grey denotes land; white ocean cells are excluded or have undefined characteristics.

Overall, SwinTrans-ConvLSTM reproduced the main intensity and spatial patterns of MHW-related warm anomalies and maintained MHW event-day detection skill through the 7-day forecast lead. At the 1-day lead, it also achieved the

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

lowest errors among the five deep-learning models for mean intensity, event count, and total MHW days. These results demonstrate the potential of SST-based forecasts to support short-term MHW monitoring in the East China Sea.

4 Discussion

The SST and MHW evaluations in Sect. 3 demonstrate how short-term SST forecasts can support MHW detection in the East China Sea. This section discusses the contribution of physically motivated input features, the implications for short-term MHW detection, and the limitations relevant to potential early-warning applications.

4.1 Contribution of Thermal and Wind Input Features

An ablation study examined the contributions of thermal and wind input features to SwinTrans-ConvLSTM SST forecast accuracy (Table 6). Four configurations shared the same model architecture, training objective, and evaluation procedure: a reference configuration with common SST-based inputs and three variants incorporating thermal predictors, wind predictors, or both.

The thermal predictors comprised the air–sea temperature difference, its 3-day trailing mean, and 2 m air temperature, motivated by the role of air–sea temperature differences in sensible heat exchange (Fairall et al., 2003). Their inclusion reduced pooled MAE by 2.38% relative to the reference configuration. The wind predictors comprised the two 10 m wind components and their 3-day trailing means, reflecting SST–surface wind coupling (Chelton and Xie, 2010), and reduced pooled MAE by 1.67%.

Combining both predictor groups produced the lowest MAE (0.5011 °C) and RMSE (0.7081 °C), representing reductions of 2.75% and 2.56%, respectively, relative to the reference configuration. This full configuration was used in Sect. 3. The 3-day trailing means were calculated from each input day and its two preceding days. Because the ablation tested these averages within predictor groups, their individual contributions could not be isolated. The results indicate that combining physically motivated thermal and wind predictors improved SST forecast accuracy in the present experimental setup.

Table 6. Model experimental results under different input variable combinations.

Group	Input variables	MAE (°C)	RMSE (°C)	MAE improvement
Baseline	SST	0.5153	0.7267	0.00 %
Control	SST+ΔT	0.5031	0.7096	2.38 %
Control	SST+ Wind	0.5067	0.7146	1.67 %
Experimental	SST+ΔT+ Wind	0.5011	0.7081	2.75 %

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

4.2 Implications for Short-Term MHW Detection

The MHW evaluation extends the assessment of SST forecast accuracy to the detection and characterization of extreme warm events. As SwinTrans-ConvLSTM predicts seven daily SST fields using only historical SST and atmospheric inputs, its event-day detection performance reflects the information available at forecast initialization, without atmospheric forcing from the forecast period.

The comparison with persistence and AR(1) highlights the dependence of MHW detection skill on forecast lead time (Fig. 11). Persistence achieved the highest 1-day SEDI (0.956), close to that of SwinTrans-ConvLSTM (0.951), indicating that the most recent SST provides a strong reference for next-day detection. At the 7-day lead, SwinTrans-ConvLSTM achieved a higher SEDI (0.659) than persistence (0.584) and AR(1) (0.579). Persistence detected a larger proportion of observed MHW days (65.4% versus 52.9%) but also had a substantially higher false alarm rate (22.1% versus 6.8%). AR(1) produced fewer false alarms (2.7%) but detected only 34.9% of observed MHW days. These differences suggest that the advantage of SwinTrans-ConvLSTM at longer leads lies in a more favourable balance between event-day detection and false alarms rather than a higher hit rate alone.

MHW event-day detection and intensity prediction represent distinct aspects of MHW forecasting. A predicted event may satisfy the duration and threshold criteria without accurately reproducing its peak intensity. The weakening agreement between observed and predicted event peaks at longer leads, particularly the underestimation of stronger peaks (Fig. 9), indicates that successful event-day detection does not guarantee accurate estimates of event severity. Moreover, the matched-event intensity analysis excludes events without a temporal match, whereas the event-day evaluation accounts for both missed MHW days and false alarms. The two assessments therefore provide complementary information.

MHW event characteristics offer an additional measure of forecast performance beyond individual MHW days. At the 1-day lead, SwinTrans-ConvLSTM achieved the lowest errors in mean intensity, event count, and total MHW days among the five deep-learning models, with Swin Transformer performing best for mean duration (Table 5). Event count and total MHW days describe different aspects of MHW occurrence and persistence, so their joint evaluation provides a broader assessment than either metric alone. However, agreement in aggregate event characteristics does not necessarily imply accurate timing or duration of individual events. These findings support the potential use of SST forecasts for short-term MHW monitoring in the East China Sea. Operational early-warning performance remains to be established through initialization-based evaluation of individual alerts, including onset lead time, missed events, and false alarms.

4.3 Limitations and Future Perspectives

SST forecast errors were relatively large in some coastal and estuarine areas, but the present experiments cannot isolate the effects of coastal grid representation, atmospheric input resolution, or processes such as river discharge and tidal mixing. Future comparisons of alternative coastal representations and input resolutions under consistent evaluation conditions could help identify sources of these errors and potential improvements.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

The model used only historical inputs available at initialization, without atmospheric information for the subsequent seven days. Future work could assess the value of numerical weather forecasts available at initialization by comparing otherwise identical configurations over common dates and lead times. This would help determine whether forecast atmospheric forcing improves SST accuracy and MHW event-day detection, particularly at longer leads. Subsurface temperature and mixed-layer depth were not explicitly included as inputs. Incorporating these variables, where suitable data are available, could help evaluate whether upper-ocean information improves predictions of MHW intensity and duration. Because the ablation study in Sect. 4.1 focused on SST errors, potential benefits for MHW detection and event characteristics would require separate assessment.

The model was developed and evaluated in the East China Sea, and its transferability to other coastal regions remains uncertain. Testing the trained model in independent regions, with and without local retraining, would help assess its generalizability. Training cost, inference time, and memory requirements should also be quantified to evaluate the computational trade-offs relative to simpler forecasting methods. Finally, operational applicability requires evaluation of MHW alerts issued at successive forecast initializations, including onset lead time, missed events, and false alarms under real-time data and computational constraints.

5 Conclusions

This study developed SwinTrans-ConvLSTM for 1–7-day SST prediction in the East China Sea and evaluated its application to short-term MHW detection over 2020–2024. Using 14 days of historical SST and atmospheric inputs, the model jointly predicts seven daily SST fields without requiring atmospheric data from the forecast period.

SwinTrans-ConvLSTM achieved the lowest pooled SST MAE (0.5011 °C) and RMSE (0.7081 °C) among the seven methods evaluated. At the 7-day lead, its MAE and RMSE were 0.6378 °C and 0.8620 °C, respectively, representing reductions of 9.8% and 7.2% relative to AR(1). The input ablation further showed that combining thermal and wind features reduced pooled MAE and RMSE by 2.75% and 2.56%, respectively, compared with the SST-based reference configuration.

For MHW event-day detection, SwinTrans-ConvLSTM achieved the highest SEDI among all seven methods at leads of 2–7 days, reaching 0.659 at day 7 compared with 0.584 for persistence and 0.579 for AR(1). Agreement in matched-event peak intensity decreased with lead time, with the Pearson correlation coefficient declining from 0.906 at day 1 to 0.649 at day 7. At the 1-day lead, the model also achieved the lowest errors among the five deep-learning models in mean intensity, event count, and total MHW days. These findings highlight the complementary roles of event-day detection and event-characteristic evaluation.

Overall, the results demonstrate the potential of SST-based forecasts to support short-term MHW monitoring in the East China Sea. Further evaluation of event-onset warning lead times, missed events, and false alarms across independent periods and coastal regions is needed to establish their applicability to operational early warning.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

Code and data availability

All datasets used in this study are publicly available and properly cited in the text. The OSTIA SST dataset is available from the Copernicus Marine Service (CMEMS), and the ERA5 atmospheric reanalysis data is provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). The source code and minimal reproducible examples for the SwinTrans-ConvLSTM model are archived on GitHub (https://github.com/ZefangMa/MHW-Prediction-SwinTrans).

Author contributions

ZM developed the project conceptualization and methodology, wrote the software, curated the dataset, conducted the formal analysis and validation, and produced the visualizations. HC and QJ provided supervision, managed project administration, and contributed to the formal analysis. ZM prepared the original manuscript. HC, QJ, LJ, SC, and XL contributed to the review and editing of the manuscript.

Competing interests

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

Disclaimer

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

The authors would like to acknowledge the Copernicus Marine Environment Monitoring Service (CMEMS) for providing the OSTIA sea surface temperature dataset, and the European Centre for Medium-Range Weather Forecasts (ECMWF) for providing the ERA5 atmospheric reanalysis data used in this study.

Financial support

This research has been supported by the National Key Research and Development Program of China (Grant No. 2023YFD2401904).

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

References

Bonino, G., Galimberti, G., Masina, S., McAdam, R., and Clementi, E.: Machine learning methods to predict sea surface temperature and marine heatwave occurrence: a case study of the mediterranean sea, Ocean Sci., 20, 417–432, https://doi.org/10.5194/os-20-417-2024, 2024.

Chelton, D. and Xie, S.-P.: Coupled ocean-atmosphere interaction at oceanic mesoscales, Oceanog., 23, 52–69, https://doi.org/10.5670/oceanog.2010.05, 2010.

Dai, H., He, Z., Wei, G., Lei, F., Zhang, X., Zhang, W., and Shang, S.: Long-term prediction of sea surface temperature by temporal embedding transformer with attention distilling and partial stacked connection, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, 17, 4280–4293, https://doi.org/10.1109/JSTARS.2024.3357191, 2024.

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., and Wimmer, W.: The operational sea surface temperature and sea ice analysis (OSTIA) system, Remote Sensing of Environment, 116, 140–158, https://doi.org/10.1016/j.rse.2010.10.017, 2012.

Fairall, C. W., Bradley, E. F., Hare, J. E., Grachev, A. A., and Edson, J. B.: Bulk parameterization of air–sea fluxes: updates and verification for the COARE algorithm, J. Climate, 16, 571–591, https://doi.org/10.1175/1520-0442(2003)016%3C0571:BPOASF%3E2.0.CO;2, 2003.

Ferro, C. A. T. and Stephenson, D. B.: Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events, Weather and Forecasting, 26, 699–713, https://doi.org/10.1175/WAF-D-10-05030.1, 2011.

Flint, L. E. and Flint, A. L.: Downscaling future climate scenarios to fine scales for hydrologic and ecological modeling and analysis, Ecol Process, 1, 2, https://doi.org/10.1186/2192-1709-1-2, 2012.

Giamalaki, K., Beaulieu, C., and Prochaska, J. X.: Assessing predictability of marine heatwaves with random forests, Geophysical Research Letters, 49, e2022GL099069, https://doi.org/10.1029/2022GL099069, 2022.

Hao, P., Li, S., Song, J., and Gao, Y.: Prediction of sea surface temperature in the south china sea based on deep learning, Remote Sensing, 15, 1656, https://doi.org/10.3390/rs15061656, 2023.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hickox, R., Belkin, I., Cornillon, P., and Shan, Z.: Climatology and seasonal variability of ocean fronts in the east china, yellow and bohai seas from satellite SST data, Geophysical Research Letters, 27, 2945–2948, https://doi.org/10.1029/1999GL011223, 2000.

Hobday, A. J., Alexander, L. V., Perkins, S. E., Smale, D. A., Straub, S. C., Oliver, E. C. J., Benthuysen, J. A., Burrows, M. T., Donat, M. G., Feng, M., Holbrook, N. J., Moore, P. J., Scannell, H. A., Sen Gupta, A., and Wernberg, T.: A hierarchical approach to defining marine heatwaves, Progress in Oceanography, 141, 227–238, https://doi.org/10.1016/j.pocean.2015.12.014, 2016.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.

Holbrook, N. J., Scannell, H. A., Sen Gupta, A., Benthuysen, J. A., Feng, M., Oliver, E. C. J., Alexander, L. V., Burrows, M. T., Donat, M. G., Hobday, A. J., Moore, P. J., Perkins-Kirkpatrick, S. E., Smale, D. A., Straub, S. C., and Wernberg, T.: A global assessment of marine heatwaves and their drivers, Nat Commun, 10, 2624, https://doi.org/10.1038/s41467-019-10206-z, 2019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin transformer: hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992–10002, https://doi.org/10.1109/ICCV48922.2021.00986, 2021.

Mallat, S. G.: A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Machine Intell., 11, 674–693, https://doi.org/10.1109/34.192463, 1989.

Mathieu, M., Couprie, C., and LeCun, Y.: Deep multi-scale video prediction beyond mean square error, https://doi.org/10.48550/arXiv.1511.05440, 26 February 2016., 2016.

Oliver, E. C. J., Donat, M. G., Burrows, M. T., Moore, P. J., Smale, D. A., Alexander, L. V., Benthuysen, J. A., Feng, M., Sen Gupta, A., Hobday, A. J., Holbrook, N. J., Perkins-Kirkpatrick, S. E., Scannell, H. A., Straub, S. C., and Wernberg, T.: Longer and more frequent marine heatwaves over the past century, Nat Commun, 9, 1324, https://doi.org/10.1038/s41467-018-03732-9, 2018.

Parasyris, A., Metheniti, V., Kampanis, N., and Darmaraki, S.: Marine heatwaves in the mediterranean sea: a convolutional neural network study for extreme event prediction, https://doi.org/10.5194/egusphere-2024-4003, 9 January 2025., 2025.

Sasaki, Y. N. and Umeda, C.: Rapid warming of sea surface temperature along the kuroshio and the china coast in the east china sea during the twentieth century, Journal of Climate, 34, 4803–4815, https://doi.org/10.1175/JCLI-D-20-0421.1, 2021.

Shi, B., Ge, C., Lin, H., Xu, Y., Tan, Q., Peng, Y., and He, H.: Sea surface temperature prediction using ConvLSTM-based model with deformable attention, Remote Sensing, 16, 4126, https://doi.org/10.3390/rs16224126, 2024.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., and Woo, W.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting, https://doi.org/10.48550/arXiv.1506.04214, 19 September 2015., 2015.

Shrivastava, A., Gupta, A., and Girshick, R.: Training region-based object detectors with online hard example mining, https://doi.org/10.48550/arXiv.1604.03540, 12 April 2016., 2016.

Smale, D. A., Wernberg, T., Oliver, E. C. J., Thomsen, M., Harvey, B. P., Straub, S. C., Burrows, M. T., Alexander, L. V., Benthuysen, J. A., Donat, M. G., Feng, M., Hobday, A. J., Holbrook, N. J., Perkins-Kirkpatrick, S. E., Scannell, H. A., Sen Gupta, A., Payne, B. L., and Moore, P. J.: Marine heatwaves threaten global biodiversity and the provision of ecosystem services, Nat. Clim. Chang., 9, 306–312, https://doi.org/10.1038/s41558-019-0412-1, 2019.

Song, J., Zhang, X., and Jiang, W.: Sea surface temperature prediction in china sea based on SAM-LSTM approach, AJESE, 8, 14–22, https://doi.org/10.11648/j.ajese.20240802.11, 2024.

Sun, D., Jing, Z., and Liu, H.: Deep learning improves sub-seasonal marine heatwave forecast, Environ. Res. Lett., 19, 064035, https://doi.org/10.1088/1748-9326/ad4616, 2024.

Sun, W., Zhou, S., Yang, J., Gao, X., Ji, J., and Dong, C.: Artificial intelligence forecasting of marine heatwaves in the south china sea using a combined U-net and ConvLSTM system, Remote Sensing, 15, 4068, https://doi.org/10.3390/rs15164068, 2023.

https://doi.org/10.5194/egusphere-2026-7232
Preprint. Discussion started: 26 November 2026
© Author(s) 2026. CC BY 4.0 License.



Tan, H.-J., Cai, R.-S., Bai, D.-P., Hilmi, K., and Tonbol, K.: Causes of 2022 summer marine heatwave in the east china seas, Advances in Climate Change Research, 14, 633–641, https://doi.org/10.1016/j.accre.2023.08.010, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention is all you need, https://doi.org/10.48550/arXiv.1706.03762, 12 June 2017., 2017.

Welandawe, S., Priyadarshana, Y. H. P. P., Senanayake, N., and Silva, E. N. S.: Machine learning techniques for marine heatwave prediction: a comprehensive review, Intell. Mar. Technol. Syst., 3, 28, https://doi.org/10.1007/s44295-025-00076-1, 2025.

Xiang, L., Guan, J., Xiang, J., Zhang, L., and Zhang, F.: Spatiotemporal model based on transformer for bias correction and temporal downscaling of forecasts, Front. Environ. Sci., 10, 1039764, https://doi.org/10.3389/fenvs.2022.1039764, 2022.

Xiao, C., Chen, N., Hu, C., Wang, K., Xu, Z., Cai, Y., Xu, L., Chen, Z., and Gong, J.: A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data, Environmental Modelling & Software, 120, 104502, https://doi.org/10.1016/j.envsoft.2019.104502, 2019.

Xie, B., Qi, J., Yang, S., Sun, G., Feng, Z., Yin, B., and Wang, W.: Sea surface temperature and marine heat wave predictions in the south china sea: a 3D U-net deep learning model integrating multi-source data, Atmosphere, 15, 86, https://doi.org/10.3390/atmos15010086, 2024.

Xu, W., Guo, B., Mantravadi, V. S., Xu, Z., Wan, C., Sabuyi, J. S., and Xu, Z.: Spatial–temporal characteristics and trend analysis of marine heatwaves in the east china sea based on deep learning, Water, 17, 3076, https://doi.org/10.3390/w17213076, 2025.

Zhang, Q., Wang, H., Dong, J., Zhong, G., and Sun, X.: Prediction of sea surface temperature using long short-term memory, IEEE Geoscience and Remote Sensing Letters, 14, 1745–1749, https://doi.org/10.1109/LGRS.2017.2733548, 2017.

Zhao, B., Zhang, T., Wang, Y., Lin, P., Liu, H., Huang, P., Huang, W., Wang, P., and Li, Y.: Combined LFS and ConvLSTM to forecast marine heatwaves: a case study, Atmospheric and Oceanic Science Letters, 19, 100690, https://doi.org/10.1016/j.aosl.2025.100690, 2026.

Zhao, X.: Deep learning for ocean temperature forecasting: a survey, 2024.

Zhao, X., Qi, J., Yu, Y., and Zhou, L.: Deep learning for ocean temperature forecasting: a survey, Intell. Mar. Technol. Syst., 2, 28, https://doi.org/10.1007/s44295-024-00042-3, 2024.