

Authors' Response to Referee #1

We sincerely thank the referee for the thorough and constructive evaluation of our manuscript. We recognize the referee's concerns regarding the framing of our scientific contribution, the use of the term "physics-informed", the physical interpretation of the input features, and the temporal configuration of the atmospheric predictors. In response, we have audited the temporal sourcing of the specified atmospheric inputs. We have reconsidered the scope of our scientific claims and will revise the manuscript to distinguish the methodological contribution, the empirical forecasting results, and the physical interpretations supported by the experiments. Below, the referee's comments are reproduced in black, and our point-by-point responses are presented in blue.

Referee comment

This manuscript develops a SwinTrans-ConvLSTM framework for 1–7-day SST forecasting and subsequent marine heatwave (MHW) identification in the East China Sea. The reported prediction accuracy is impressive, and the chronological train–validation–test split is appropriate. However, in its current version, I do not find the manuscript to provide sufficiently strong scientific or methodological novelty to justify publication. The claimed "physics-informed" contribution is substantially overstated, the physical interpretation of the learned processes is not adequately demonstrated, and several aspects of the experimental design and extreme-event verification raise important concerns. In particular, the study mainly represents an integration and optimization of existing deep-learning components rather than a major methodological advance, while the MHW analysis remains largely diagnostic rather than revealing new physical or physical–ecological mechanisms. I therefore recommend rejection.

Response

We appreciate the referee's assessment and would like to clarify the contribution and scope of our study. The work develops and evaluates a hybrid spatiotemporal framework, SwinTrans-ConvLSTM, for 1–7-day SST forecasting at 0.05° resolution in the East China Sea, supplemented by analyses of SST anomalies and daily threshold exceedances relevant to marine heatwaves. The architecture combines ConvLSTM-based temporal encoding with Swin Transformer blocks for spatial feature learning. Its empirical evaluation includes comparisons with four deep-learning

reference models—LSTM, ConvLSTM, Transformer, and Swin Transformer—using aggregate and lead-time-specific metrics (Tables 3 and 4), together with an input-ablation study comparing configurations that incorporate different combinations of wind components and air–sea temperature contrast (Table 5).

We acknowledge that combining established neural-network components does not, by itself, constitute a fundamental architectural advance. Accordingly, we will frame the contribution in terms of the development and systematic regional evaluation of this specific hybrid forecasting framework, rather than claiming a fundamentally new deep-learning architecture.

At the same time, we recognize that the use of “physics-informed” and several mechanistic descriptions could overstate what a data-driven forecasting framework directly proves. In the revised manuscript, we have refined the terminology to “multivariate input features” and “spatial-gradient regularization”. We also distinguish continuous SST forecast accuracy, SST-anomaly statistics, and daily threshold-exceedance performance, and clarify the scope of our temporal audit, as detailed in Comment 4.

We acknowledge that the present study does not establish a new physical mechanism governing MHW development or evaluate ecological responses and physical–ecological interactions. We have therefore limited our interpretation to the predictive results and analyses actually supported by our experiments. We respectfully request that the revised manuscript be considered in light of these clarifications and the specific responses below.

Major comments

(1) The manuscript is presented as a physics-guided study of marine heatwave prediction, but the actual scientific contribution is primarily improved SST forecasting. MHWs are subsequently diagnosed from the predicted SST using a conventional percentile-based definition. No ecological variables, ecosystem responses, biological processes, or physical–ecological interactions are explicitly analyzed. More importantly, the manuscript does not identify a new physical mechanism controlling MHW development in the East China Sea. The discussions regarding the Western Pacific Subtropical High, air–sea heat exchange, Kuroshio intrusion, monsoon forcing, stratification, and coastal fronts mainly rely on mechanisms already established in previous studies.

Therefore, the manuscript currently provides limited new process-level understanding of MHW dynamics.

Response

We thank the referee for this comment and would like to clarify the methodological scope of our study. Forecasting SST and subsequently identifying MHWs from the predicted SST has precedent in data-driven MHW forecasting research, including Bonino et al. (2024) and Xie et al. (2024). The present study should therefore be understood as a regional SST forecasting study with subsequent MHW-relevant analyses, rather than as a study proposing a new physical mechanism of MHW development.

Under the definition of Hobday et al. (2016), MHW identification requires daily SST to exceed a seasonally varying 90th-percentile threshold for at least five consecutive days. Errors in predicted SST can therefore affect event detection, timing, duration, and intensity. This relationship motivates SST-based analyses relevant to MHWs, but improvements in aggregate SST accuracy alone do not establish improved MHW event forecast skill.

We acknowledge that the present study focuses on SST forecasting and subsequent analyses of SST anomalies and daily threshold exceedances, without examining ecological variables, biological responses, or physical–ecological interactions. It also does not establish a new physical mechanism governing MHW development in the East China Sea. Discussions of the Western Pacific Subtropical High, air–sea heat exchange, Kuroshio intrusion, monsoon forcing, stratification, and coastal fronts will therefore be presented as regional background supported by previous studies. The reported spatial and seasonal forecast-error patterns will not be interpreted as evidence that these processes causally determine the model’s predictive performance. We will clearly distinguish established regional oceanographic knowledge from conclusions directly supported by our experiments.

In the revised manuscript, we will clarify the contribution of our work and distinguish the model’s SST forecasting performance from the scope of the subsequent MHW analyses. Specifically, we will revise Sects. 3 and 4 to distinguish established regional oceanographic context from conclusions supported by our experiments. In Sect. 3.5, we will clarify the evaluated SST-anomaly quantities, sample selection, and averaging procedures. Forecast-performance claims will be

limited to the quantities actually evaluated, and daily threshold-exceedance skill will be discussed separately from continuous SST accuracy and SST-anomaly statistics.

Unless event-level verification has been conducted, we will explicitly acknowledge that the present analyses do not establish full MHW event forecasting skill in terms of onset, duration, and intensity. We will not infer improved event-level prediction solely from improvements in aggregate SST metrics.

(2) The manuscript repeatedly describes the framework and loss function as “physics-informed”. However, the model does not explicitly impose governing physical equations, heat conservation, mixed-layer heat-budget constraints, advection–diffusion equations, or dynamically consistent momentum/thermodynamic balances. The wind components and air–sea temperature difference are simply added as predictor variables, while the gradient loss is essentially a spatial-structure regularization term rather than a physical conservation constraint. Therefore, the method would be more appropriately described as physics-guided or physics-aware feature engineering.

Response

We thank the referee for this important distinction regarding terminology. We agree that our use of “physics-informed” overstates the physical constraints imposed by the present framework. Neither the network architecture nor the training objective explicitly enforces governing physical equations, heat conservation, mixed-layer heat-budget balances, or advection–diffusion equations. Instead, the 10-m wind components (u_{10} , v_{10}) and the air–sea temperature difference (ΔT) are incorporated at the input stage as physically motivated predictors, selected for their relevance to near-surface wind conditions and air–sea thermal contrast. Their inclusion does not impose the corresponding physical relationships on the network.

Likewise, we agree that the spatial gradient loss L_{grad} is a mathematical spatial-structure regularization term rather than a physical conservation constraint. In the implementation, this term penalizes the L1 discrepancy between the magnitudes of adjacent-grid finite differences in the predicted and reference fields along the two horizontal directions. It is intended to encourage agreement in local spatial variation and discourage excessive smoothing. This regularization does not enforce heat conservation, mixed-layer heat-budget balances, or other governing physical equations.

Following the referee’s suggestion, we have revised the terminology to describe the input design as “multivariate input features” and the gradient term as “spatial-gradient regularization”. We will revise the terminology and associated interpretations throughout the entire article. Specifically, Sect. 2.3 will be renamed from “Data Preprocessing with physical constraints” to “Multivariate Input Features and Data Preprocessing”; the objective function in Sect. 2.6.1 will be renamed from “Physics-Informed Curriculum Loss” to “Gradient-Regularized Curriculum Loss”; and Sect. 4.1 will be renamed from “Predictive Gains from Physics-Informed Variables” to “Evaluation of Multivariate Input Features”. These revisions will clarify the motivation for selecting related-physical variables as inputs and the mathematical role of regularization in the existing method.

(3) Several sections infer that the model captures processes such as Ekman transport, wind-driven mixing, warm-water advection, Kuroshio intrusion, mixed-layer thermal memory, and air–sea heat exchange simply because wind vectors, ΔT , or a three-day moving average are included as inputs. These interpretations are plausible hypotheses, but they are not demonstrated by the experiments. For example, including u_{10} and v_{10} does not by itself prove that the network has learned Ekman transport or vertical mixing. Similarly, interpreting a three-day moving average as a direct analogue of mixed-layer thermal inertia is not physically validated. If process understanding is intended as a major contribution, additional diagnostics such as heat-budget analysis, feature attribution, perturbation/counterfactual experiments, or process-conditioned evaluation would be required. Otherwise, these mechanistic claims should be substantially weakened.

Response

We appreciate this comment on the distinction between the physical motivation for selecting input variables and the mechanisms demonstrated by the experiments. The input-ablation experiments in Table 5 compare SST prediction errors across configurations containing different combinations of wind vectors (u_{10} , v_{10}) and air–sea temperature contrast ($\Delta T = SST - T_{2m}$). These comparisons assess predictive performance, but cannot demonstrate that the neural network has learned specific physical processes such as Ekman transport, vertical mixing, Kuroshio intrusion, or mixed-layer heat storage. Following the referee’s suggestion, we will substantially weaken the corresponding mechanistic claims and distinguish the physical rationale for feature selection from the empirical conclusions supported by forecast verification.

For the wind inputs (u_{10} and v_{10}), we will state that the vector components retain information about wind magnitude and direction relevant to near-surface wind conditions. We will remove claims that their inclusion imposes a dynamical constraint or demonstrates the learning of Ekman transport, advection, or vertical mixing. Differences associated with the wind input group will be described as empirical performance differences between the tested configurations, without attributing them to a specific physical mechanism.

For the air–sea temperature difference ($\Delta T = \text{SST} - T_{2m}$), we will describe it as a measure of surface thermal contrast. Its selection is physically motivated by the role of air–sea temperature differences in bulk sensible-heat-flux formulations (Fairall et al., 2003). The temperature contrast alone does not constitute a heat-flux estimate or impose a thermodynamic budget constraint. We will revise the corresponding interpretations accordingly.

For the three-day trailing means, we will describe them as empirical temporal smoothing features that provide a smoothed summary of recent wind conditions and air–sea thermal contrast. We will remove the description of this averaging operation as a validated mathematical analogue of mixed-layer thermal inertia.

We will revise the feature-engineering description in Sect. 2.3, the architecture interpretation in Sect. 2.5.1, the input-ablation discussion in Sect. 4.1, and the Conclusions. Literature on regional physical processes will support the motivation for predictor selection and provide regional context, while conclusions about model performance will be limited to the evidence provided by forecast verification and input comparisons.

(4) There is a fundamental inconsistency between the forecasting setup and the physical interpretation of forecast-error growth. According to the experimental design, the model uses information from the preceding 14 days to predict SST for the subsequent 7 days. The manuscript therefore needs to state very clearly whether atmospheric variables during the forecast period are available to the model. If only past SST and past ERA5 atmospheric variables are used, then future atmospheric forcing over lead days 1–7 is not provided. In that case, statements attributing the degradation of long-lead forecasts to the “accumulation of uncertainty in atmospheric forcing and boundary conditions” are not logically consistent with the actual forecasting configuration. The more direct limitation would be the absence of future atmospheric forcing itself. Conversely,

if ERA5 atmospheric variables from the forecast period are used, then the experiment is not a true operational forecast because future reanalysis information would be available to the model, resulting in information leakage or an unrealistic hindcast advantage. The authors must explicitly clarify whether atmospheric predictors are past observations/reanalysis only or include future atmospheric fields. If future forcing is required, an operational experiment should instead use genuine numerical weather prediction forecasts. This issue directly affects the interpretation and practical validity of the claimed 7-day MHW forecasting capability.

Response

Thank you for identifying the ambiguity in the atmospheric input description. At a forecast origin t , the 14 input frames are dated $t - 13$ through t , and the model produces the seven lead-time outputs for $t + 1$ through $t + 7$ jointly. ERA5 values dated within the forecast period are not supplied as additional atmospheric inputs. The predictions are produced directly rather than by feeding a predicted SST field back into the model to generate the next lead time.

The atmospheric variables used as input are the 10-m wind components and 2-m air temperature. The air–sea temperature contrast is calculated from SST and air temperature with the same daily timestamp. The three-day means of $u10$, $v10$, and ΔT are trailing means. For an input date s , a full three-day mean uses $s - 2$, $s - 1$, and s . It therefore does not require an atmospheric value after the forecast origin. These derived features can extend the underlying atmospheric history by two days beyond the first input-frame date, which we will distinguish from the 14-frame input length.

We have audited the temporal sourcing of the atmospheric inputs in the merged daily dataset over the 7,289 ocean grid points used by the model. The audit covered the training period (1982–2015), validation period (2016–2019), and test period (2020–2024). Across these periods, $u10$, $v10$, $T2m$, and the derived daily air–sea temperature contrast contained no missing or infinite values on the ocean grid points. The three-day averages were calculated using trailing windows that included only the corresponding input day and preceding days. For every forecast sample, the source-date check for these four inputs and their three trailing-mean features found no values dated after the forecast origin. These checks confirm that forecast-period ERA5 values were neither supplied through the atmospheric input windows nor introduced by missing-value filling.

We agree that the original explanation of lead-time error growth in terms of accumulated uncertainty in atmospheric forcing and boundary conditions is not supported by this configuration. No future atmospheric forcing is prescribed, and no explicit ocean boundary forcing is supplied to the network. We have therefore removed the claim that error growth is caused by the accumulation of uncertainty in future atmospheric forcing and boundary conditions. The increase in forecast error with lead time does not, by itself, quantify the contribution of the absence of future atmospheric forcing, so we do not present this limitation as the sole demonstrated cause of error growth.

We will describe the experiments as retrospective forecasts using historical reanalysis predictors. The study does not demonstrate that all predictors would be available at an operational forecast issue time. An operational implementation would require documented input availability and latency. Providing future numerical weather prediction fields would be a different configuration requiring separate training and verification; such fields are not inputs in the present experiments.

We will make the input dates and the direct seven-output strategy explicit in the Data and forecasting-setup descriptions. We will revise the explanations of lead-time performance in the Results and Discussion, including Sect. 4.2, and add the reanalysis-availability and future-forcing limitations to Sect. 4.3. A diagram distinguishing input-frame dates, trailing-mean source dates, and output dates will be added if it improves clarity.

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 Science*, 20, 417–432, <https://doi.org/10.5194/os-20-417-2024>, 2024.
- 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, *Journal of Climate*, 16, 571–591, [https://doi.org/10.1175/1520-0442\(2003\)016%3C0571:BPOASF%3E2.0.CO;2](https://doi.org/10.1175/1520-0442(2003)016%3C0571:BPOASF%3E2.0.CO;2), 2003.

Hobday, A. J., Alexander, L. V., Perkins, S. E., Smale, D. A., Straub, S. C., Oliver, E. C. J., Benthuisen, 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.

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.