Comparative ocean temperature forecasting using SARIMA, ETS, and LSTM

Seema Gupta[1] & Hakan Lane [2,*]

[1] Hochschule Rhein-Waal University of Applied Sciences, Germany
[2] Johannes Gutenberg University Mainz, Germany

SUMMARY

Accurate ocean temperature forecasting is essential for understanding long–term environmental variability and supporting ecological decision–making. This study evaluates the performance of SARIMA (*Seasonal Autoregressive Integrated Moving Average*), ETS (*Error Trend Seasonality*), and LSTM (*Long Short-Term Memory*) models for predicting ocean temperature using historical time series data from the CalCOFI programme. The dataset was preprocessed using seasonal decomposition and stationarity analysis, and forecasting accuracy was assessed using MAPE (*Mean Absolute Percentage Error*), RMSE (*Root Mean Squared Error*), and MSE (*Mean Squared Error*). The results indicate that LSTM produced the most accurate and stable forecasts overall, achieving lower RMSE and MAPE values at most stations and depths. It effectively captured nonlinear behaviour, seasonal variability, and extreme temperature fluctuations. SARIMA demonstrated a strong capability to model trend and seasonality, while ETS generated smoother forecasts but with comparatively higher error values. Residual diagnostics further confirmed LSTM's superior ability to learn complex temporal dependencies present in ocean temperature data. This study contributes to a comparative evaluation of classical statistical and deep learning models for ecological time series forecasting and provides evidence supporting the application of LSTM for improved oceanographic prediction and environmental monitoring.

KEYWORDS

Ecological time-series, ETS, forecasting, LSTM, SARIMA

ARTICLE TYPE
Preliminary communication

ARTICLE INFO
Received: April 21, 2026
Accepted: May 22, 2026
DOI: 10.62366/crebss.2026.1.002
JEL: C22, C45, Q54

1. Introduction

Ocean temperature is critical in climate studies because it affects marine ecosystems, weather patterns, and global climate systems (Venegas et al., 2023). Accurate forecasting is therefore essential to understand and mitigate the climate-change impacts on marine environments and coastal communities (Singh Bist et al., 2024). Traditional time-series models such as ARIMA, SARIMA, and ETS are widely used across environmental and epidemiological forecasting, including ARIMA for COVID-19 epidemiology (Tomov et al., 2023), SARIMA for environmental

*Corresponding author

epidemiology (Gudžiūnaitė et al., 2023) and river-flow prediction (Tadesse & Dinka, 2017), and SARIMA/ETS for food-borne disease forecasting (Xian et al., 2023). With advances in deep learning, Long Short-Term Memory (LSTM) networks can improve forecasting by capturing complex temporal patterns, as shown in financial time-series prediction (Buczyński et al., 2023). Because ocean temperature records are sequential, temporally dependent, and span multiple decades (1949–2016), time-series models are well suited for forecasting future variability, i.e. ARIMA, SARIMA, and ETS are commonly applied to environmental and climate variables, while LSTM offers a nonlinear alternative that can capture long-memory property. Beyond forecasting, time-series data support trend analysis, pattern discovery, anomaly detection, and forecasting. Accordingly, this study compares SARIMA, ETS, and LSTM for forecasting ocean temperatures at depths between 0 and 500 m using monthly observations from the CalCOFI programme (California Cooperative Oceanic Fisheries Investigations). This is a long-term oceanographic ecosystem monitoring and research programme with the goal of serving data in support of sustainable marine resource management (CalCOFI, 2025).

Therefore, the study objective is to compare SARIMA, ETS, and LSTM using performance measures RMSE, MSE, and MAPE in order to identify the most effective model for monthly ocean temperature forecasting. The relevance of this objective lies in the growing importance of accurate ocean temperature forecasting for climate monitoring, marine ecosystem management, and early warning of phenomena such as marine heatwaves and coral bleaching, all of which depend on reliable monthly-scale predictions. Methodologically, the study addresses a genuine gap: although SARIMA and ETS are well established for seasonal time-series and LSTM has shown superior in capturing complex nonlinear patterns, there is no consensus on which approach performs best for oceanographic data. By benchmarking the three models under identical conditions using complementary accuracy metrics (RMSE, MSE, and MAPE), the study offers an objective, evidence-based comparison, with particular attention to whether ETS improves upon SARIMA and whether LSTM delivers superior predictive accuracy across stations and depths. Its contribution is thus twofold: practically, it identifies the most suitable tool for ocean temperature prediction, and academically, it adds empirical evidence to the debate between traditional statistical methods and machine learning in environmental forecasting, providing a replicable framework for similar studies.

2. Materials and methods

Time-series models capture temporal dependence in sequential observations to identify trend, seasonality, and autocorrelation structures for forecasting (Box & Jenkins, 1976; Hyndman & Athanasopoulos, 2018). A commonly used additive decomposition represents a time-series as:

$$Y_t = T_t + S_t + R_t \tag{1}$$

where Y_t is the observed value, T_t is the trend, S_t is the seasonal component, and R_t is the residual. Stationarity is a fundamental assumption in many statistical time-series models. The Augmented Dickey–Fuller (ADF) test is used to detect unit roots and assess whether differencing is required (Dickey & Fuller, 1979). Seasonal time-series often require seasonal differencing when periodic cycles are present. Seasonal Autoregressive Integrated Moving Average (SARIMA) models extend ARIMA by incorporating seasonal autoregressive and moving-average components and are denoted as $(p, d, q) \times (P, D, Q)_s$, where s represents the seasonal period (Box & Jenkins, 1976). Model selection is commonly performed using the Akaike information criterion AIC (Akaike, 1974).

Error–Trend–Seasonality (ETS) models belong to the exponential smoothing family and represent time-series through combinations of error, trend, and seasonal components within a state-space framework (Hyndman & Athanasopoulos, 2018).

Long Short-Term Memory (LSTM) networks are recurrent neural architectures designed to capture long-term dependencies and nonlinear temporal relationships without strict stationarity assumptions (Hochreiter & Schmidhuber, 1997).

Forecast accuracy is evaluated using Mean Squared Error, Root Mean Squared Error, and Mean Absolute Percentage Error (Hyndman & Koehler, 2006). These metrics are selected for their complementary properties: MSE and RMSE emphasize large errors and retain the original measurement units, whereas MAPE offers a scale-independent percentage that enables comparison across stations and depths.

Data were obtained from the CalCOFI database CalCOFI (2025), comprising 864,863 discrete water samples collected at different depths and locations from 1949 to the present, including measurements of temperature, salinity, dissolved oxygen, chlorophyll *a*, nutrients, and other variables. Each record corresponds to a bottle closure during a CTD–rosette cast at a specific depth and station. Temperature (`T_degC`) represents *in situ* water temperature measured at the recorded depth (`Depthm`).

For temperature forecasting, the extracted variables were `Cst_Cnt`, `Btl_Cnt`, `Sta_ID`, `Depth_ID`, `Depthm`, and `T_degC`. Table 1 describes the raw CalCOFI bottle database, of which station, depth, and water temperature are the primary variables used in this study. For the temperature time-series, observations of temperature data were aggregated to monthly intervals using the monthly mean temperature (month-start resampling), and missing values were replaced by linear interpolation. The full dataset contained 113 stations and 74 variables; three stations were randomly selected and analysed at 0 m (surface) and 500 m (deep water), producing six time series (three stations × two depths). Because temperature declined steeply until approximately 500 m and was relatively stable thereafter, forecasting focused on these two representative depth levels. The six subsets are summarised in Table 2.

Table 1. *Database description*

Feature	Description	Units
Cst_Cnt	Auto-numbered cast count – all casts consecutively numbered (1 is the first station)	n.a.
Btl_Cnt	Auto-numbered bottle count – all bottles ever sampled, consecutively numbered	n.a.
Sta_ID	CalCOFI line and station	n.a.
Depth_ID	[Century][YY][MM][ShipCode][CastType][Julian Day]-[CastTime]-[Line][Sta][Depth][Bottle]-[Rec_Ind]	n.a.
Depthm	Depth in metres	metre
T_degC	Temperature of water	°C

Source: CalCOFI Bottle Database https://calcofi.org/data/oceanographic-data/bottle-database

Table 2. *Ocean temperature data shape*

Station name	Station	Depth (metre)	Data shape
090.0 037.0	1a	0	328×3
	1b	500	323×3
090.0 070.0	2a	0	316×3
	2b	500	313×3
090.0 045.0	3a	0	330×3
	3b	500	316×3

Note: columns used are time, depth, and temperature.

Preprocessing comprised monthly resampling, missing-value handling, and seasonality identification. After downloading, the data were assessed for seasonality, trend, and distributional behaviour. Missing/irregular values were corrected using interpolation (Lepot et al., 2017) to ensure modelling readiness (Usmani et al., 2024), and because stations follow a uniform sampling methodology, the workflow used for one station was treated as representative of all. Seasonal decomposition was applied to the monthly temperature series (especially 500 m) using an additive form because seasonal amplitude was approximately constant over time (Hyndman & Athanasopoulos, 2018).

Stationarity was evaluated using the ADF test (Dickey & Fuller, 1979). An ADF statistic of -5.41, below the 1%, 5%, and 10% critical values with an approximately zero $p-$value, rejected non-stationarity. Thus, the series is stationary in its non-seasonal component, but the annual cycle still requires seasonal differencing ($D = 1$) for SARIMA to capture periodic behaviour.

SARIMA model is commonly used for time-series exhibiting trend, seasonality, and autocorrelation (Box & Jenkins, 1976). Prior to modelling, data were cleaned, missing values interpolated, and observations resampled to monthly frequency (Castel & Burr, 2021; Hyndman & Athanasopoulos, 2018). Initial model identification used autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. For Station 1b, the ACF showed slow decay with significant spikes at multiples of 12, indicating strong seasonality, while the PACF exhibited a sharp cutoff after lag 1, suggesting an autoregressive order of $p = 1$. The ACF further suggested $q = 1 - 2$. Based on this inspection, a manual SARIMA$(1, 1, 1) \times (1, 1, 1)_{12}$ specification was proposed. To reduce subjectivity and allow consistent optimization across series, AutoSARIMA was also applied. AutoSARIMA evaluates combinations of (p, d, q, P, D, Q) and selects the model that minimizes the AIC information criterion (Akaike, 1974; Hyndman & Khandakar, 2008):

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{2}$$

where k is the number of parameters and L is the maximised likelihood. According to Table 3 SARIMA$(3, 1, 2) \times (2, 0, 1)_{12}$ was selected, achieving a lower AIC than the manual model (-1260.55 versus -1195.02). Forecasting employed an 80–20 train–test split, with one-step-ahead predictions generated using autoregressive, moving-average, differencing, and seasonal components.

The ETS model, originating in the late 1950s, is part of the exponential-smoothing family and represents time series through combinations of error, trend, and seasonal components within a state-space framework (Hyndman & Athanasopoulos, 2018). While ARIMA-type models are valued for simplicity, they may struggle with complex temporal structures (Fatima & Rahimi, 2024). Unlike SARIMA, ETS does not require differencing or strict stationarity assumptions, directly modelling level, trend, and seasonality. This makes ETS suitable for series with irregular trends or multiplicative seasonal behaviour, where SARIMA performance may degrade (Gao, 2025).

In this study, AutoETS evaluated additive and multiplicative combinations of error, trend, and seasonality, including damped trends, selecting the specification with the lowest AIC (Hyndman & Athanasopoulos, 2018). As shown in **??**, the optimal model for the main station series had no trend, additive seasonality, and no damping (AIC=-3070.46), while alternatives incorporating trends or multiplicative seasonality produced higher AIC values and were rejected.

LSTM networks are recurrent neural architectures designed to capture long-term temporal dependencies and nonlinear dynamics in time-series data through gated memory cells, mitigating the vanishing-gradient problem (Hochreiter & Schmidhuber, 1997). They are particularly suitable for temperature series with long-term variability and complex seasonality (Zhang et al., 2023) and can generalise across nonlinear and nonstationary behaviours that challenge SARIMA and ETS models (Zhang et al., 2023).

Prior to training, temperature values were scaled to the [0, 1] range to stabilise optimisation (Al-Sharafi et al., 2023). A sliding-window approach transformed the series into supervised learning sequences, where past observations formed inputs and the subsequent value served as the target (Kong et al., 2024; Zhang et al., 2023), reshaped to (samples, timesteps, features). The network consisted of a single LSTM layer followed by a dense output layer for one-step-ahead prediction, trained using the Adam optimizer with MSE loss. An 80:20 train–test split was applied, with 20–50 epochs and batch size = 16. Training loss decreased smoothly from 0.1691 at epoch 1 to 0.0039 at epoch 20, indicating stable convergence; predictions were subsequently inverse-transformed to the original °C scale.

The choice of a single LSTM layer with a moderate number of units was motivated by the need to balance model complexity and generalisation capability. Deeper architectures may capture more complex hierarchical representations but also increase the risk of overfitting, particularly for relatively small datasets such as the station-specific temperature series used in this study. The selected configuration provided sufficient capacity to model nonlinear temporal dependencies while maintaining computational efficiency.

Hyperparameter selection was guided by empirical testing and prior literature. The number of epochs (20–50) was chosen to allow the model to converge while avoiding excessive training that could lead to overfitting. The batch size of 16 ensured stable gradient updates without significantly increasing computational cost. The Adam optimizer was selected due to its adaptive learning-rate properties and proven effectiveness in time-series forecasting applications.

To ensure consistency and comparability across models, the same train–test split (80:20) was applied to all series. Model performance was evaluated on the test set using one-step-ahead forecasting, ensuring that predictions were based only on past information. This approach reflects realistic forecasting scenarios and avoids data leakage.

Finally, inverse transformation of scaled predictions was performed to return values to the original temperature scale, allowing direct comparison of model outputs with observed data and enabling meaningful interpretation of forecast errors.

Residual analysis was used to assess whether SARIMA, ETS, and LSTM adequately captured temporal structure. An appropriate model should produce residuals that are random, uncorrelated, and approximately normally distributed (Ljung & Box, 1978). Residual plots were inspected for remaining trends or seasonality, while histograms and Q–Q plots assessed normality (Caña et al., 2021; Liu & Li, 2023). Residual autocorrelation was examined using the ACF, where spikes outside confidence bounds indicated remaining structure (Hyndman & Athanasopoulos, 2018).

Model performance was evaluated using MSE, RMSE, and MAPE (Hyndman & Koehler, 2006; Chai & Draxler, 2014):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\text{actual} - \text{predicted})^2 \tag{3}$$

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{4}$$

$$\text{MAPE} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{\text{actual} - \text{predicted}}{\text{actual}}\right| \tag{5}$$

where actual denotes the observed value, predicted denotes the forecasted value generated by the model, N represents the total number of observations.

MSE and RMSE measure absolute forecast error (RMSE in °C), while MAPE provides a scale-independent percentage error. Lower values indicate better performance. Because observed temperatures were strictly positive across all stations and depths, division-by-zero issues did not arise, and MAPE was well-defined for all series.

3. Empirical results

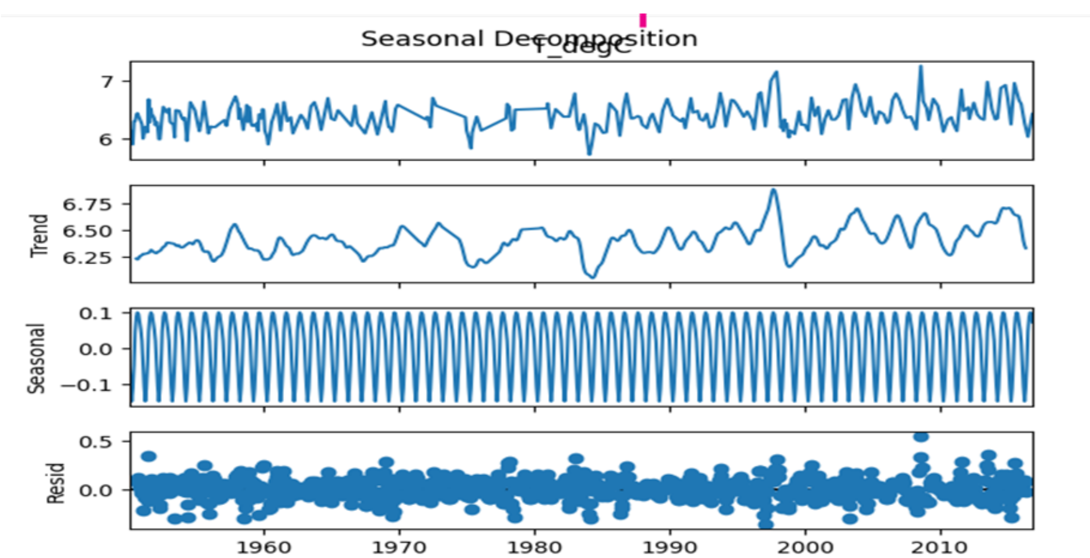

Figure 1. *Seasonal decomposition of ocean temperature*

Figure 1 shows the seasonal decomposition of temperature data: the top plot displays the original time series, the second shows the trend, the third the seasonal component, and the fourth the residual component. Seasonal decomposition of the monthly temperature series showed that the trend component captured long-term fluctuations over the study period, while the seasonal component displayed a stable annual cycle with a small amplitude of approximately ±0.1 °C. Residuals were largely random with occasional anomalies, indicating that the trend and seasonal components captured most systematic variability (Hyndman & Athanasopoulos, 2018).

The ADF test statistic (−5.41) with a near-zero $p-$value suggested the absence of a non-seasonal unit root (Dickey & Fuller, 1979). However, the pronounced annual cycle observed in the decomposition justified the use of a seasonal model with period 12 and seasonal differencing for SARIMA.

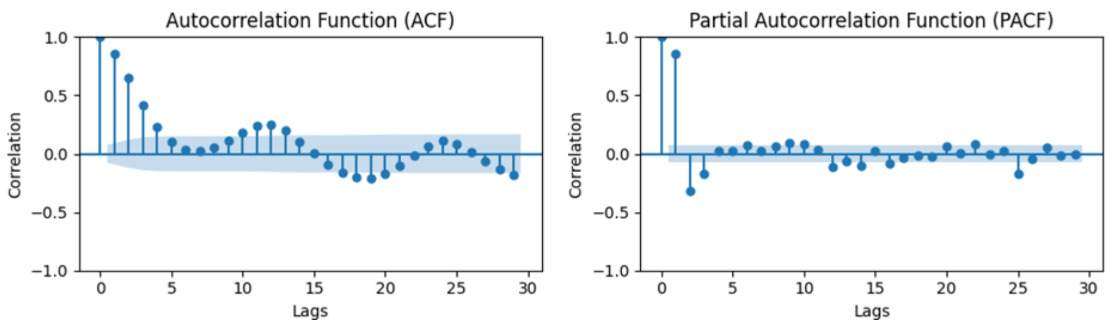

Figure 2. *ACF and PACF of monthly temperature of Station 1b*

Figure 2 illustrates the ACF and PACF plots used to guide SARIMA parameter selection. The ACF displays strong autocorrelation across multiple lags, with significant seasonal peaks confirming the need for seasonal differencing (D). The PACF shows a sharp cutoff after lag 1, indicating an autoregressive term of $p = 1$, while significant ACF spikes at lags 1 and 2 suggest a moving-average component ($q = 1 - 2$). The 12-month seasonal pattern evident in the ACF is consistent with the monthly structure of the dataset.

Table 3. *AIC results of manual SARIMA and automatic SARIMA*

Model	Specification	AIC
Manual	SARIMA(1, 1, 1) × $(1, 1, 1)_{12}$	−1195.02
Automatic	SARIMA(3, 1, 2) × $(2, 0, 1)_{12}$	−1260.55

Source: authors calculation based on CalCOFI data.

As shown in Table 3, the automatic model achieved a lower AIC (−1260.55) than the manual model (−1195.02) and was therefore retained for further analysis.

Residual diagnostics were conducted using graphical analysis of training and test residuals. Residual plots showed that errors were centred around zero and exhibited no clear remaining trend or seasonal structure. Histogram and Q–Q plot inspection indicated approximate normality with slightly heavier tails, while the residual ACF displayed only minor spikes at a few lags, suggesting that most temporal dependence had been adequately captured. Overall, the AutoSARIMA model reproduced the dominant seasonal and trend components of the temperature series and was retained for comparative evaluation.

The best-performing ETS model had no trend, additive seasonality, and no damped trend, indicating stable and repeating seasonal patterns without a strong long-term trend (AIC=−3070.46). Models including additive trend or multiplicative seasonality produced higher AIC values and were therefore rejected. ETS forecasts were smoother than SARIMA, capturing overall seasonality but underestimating sharp peaks and troughs.

The LSTM model trained for 20 epochs showed a monotonic reduction in loss from 0.1691 to 0.0039, indicating stable convergence without severe overfitting and good computational efficiency. Residual diagnostics indicated that LSTM residuals were smaller and more tightly clustered around zero than those of ETS, with reduced autocorrelation at short lags, indicating improved capture of temporal structure. Histograms indicated greater symmetry, Q–Q plots showed fewer outliers, and residual ACF plots exhibited substantially reduced autocorrelation, particularly at lags 3–5. Increasing training to 50 epochs further reduced residual magnitude and improved randomness, although minor autocorrelation persisted. Overall, LSTM residuals were closest to white noise.

Because performance trends were consistent across all stations (Table 5, Table 6, and Table 7), Station 1b is presented as a representative example for graphical comparison. This station was selected to illustrate model behaviour under deep-water conditions with moderate variability. Quantitative results for all locations are provided in Table 5, Table 6, and Table 7.

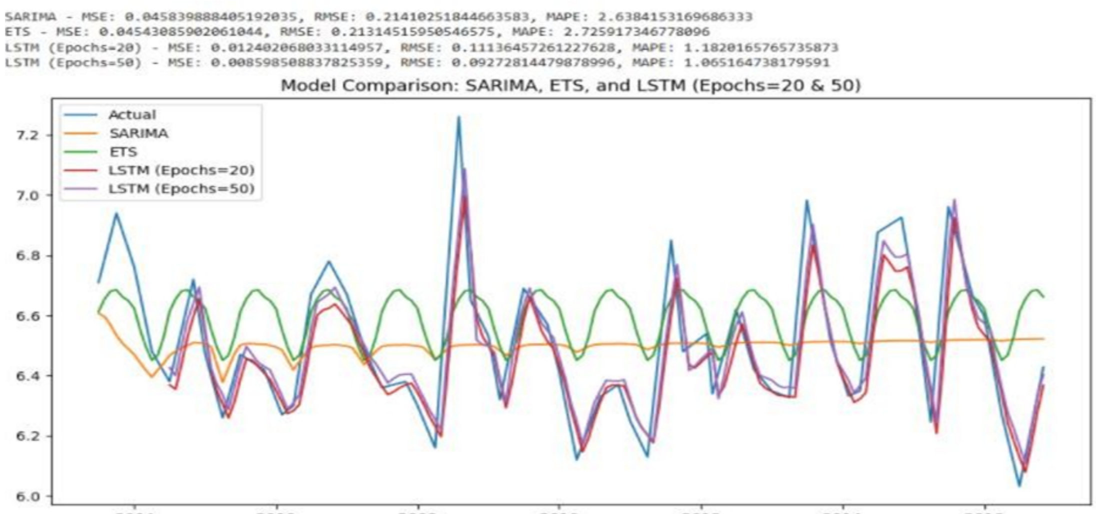

Figure 3. *Comparison of the actual temperature series at Station 1b*

Figure 3 compares observed temperatures at Station 1b with forecasts from SARIMA, ETS, LSTM (20 epochs), and LSTM (50 epochs), with evaluation metrics summarised in Table 4. SARIMA followed seasonal patterns but deviated at sharp peaks and troughs (MSE = 0.046, RMSE = 0.214, MAPE = 2.64%). ETS produced smoother forecasts with similar performance (MSE = 0.045, RMSE = 0.213, MAPE = 2.73%). LSTM (20 epochs) tracked observations more closely, particularly at extremes (MSE = 0.012, RMSE = 0.111, MAPE = 1.18%), while LSTM (50 epochs) achieved the best performance overall (MSE = 0.009, RMSE = 0.093, MAPE = 1.07%). These results show that LSTM, especially with 50 epochs, provided the most accurate forecasts for Station 1b.

Table 4. *Model results of Station 1b based on MSE, RMSE, and MAPE*

Model	MSE (°C)	RMSE (°C)	MAPE (%)
SARIMA	0.046	0.214	2.64
ETS	0.045	0.213	2.73
LSTM (50 units, 20 epochs)	0.012	0.111	1.18
LSTM (50 units, 50 epochs)	0.009	0.093	1.07

Source: authors calculation based on CalCOFI data.

The evaluation metrics MSE, RMSE, and MAPE across stations are reported in Table 5, Table 6, and Table 7.

Table 5. *Accuracy measures with respect to different models applied to Stations 1a and 1b*

Station	Model	Settings	MSE (°C)	RMSE (°C)	MAPE (%)
1a	SARIMA	Auto	1.22	1.10	4.68
1a	ETS	Auto	1.06	1.03	4.64
1a	LSTM	(50 units, 50 epochs)	0.46	0.68	3.07
1b	SARIMA	Auto	0.05	0.21	2.64
1b	ETS	Auto	0.05	0.21	2.73
1b	LSTM	(50 units, 50 epochs)	0.01	0.09	1.07

Source: authors calculation based on CalCOFI data.

Table 6. *Accuracy measures with respect to different models applied to Stations 2a and 2b*

Station	Model	Settings	MSE (°C)	RMSE (°C)	MAPE (%)
2a	SARIMA	Auto	1.76	1.33	6.07
2a	ETS	Auto	1.76	1.33	6.09
2a	LSTM	(50 units, 50 epochs)	0.26	0.51	2.56
2b	SARIMA	Auto	0.04	0.21	3.02
2b	ETS	Auto	0.04	0.19	2.65
2b	LSTM	(50 units, 50 epochs)	0.01	0.09	1.18

Source: authors calculation based on CalCOFI data.

Table 7. *Accuracy measures with respect to different models applied to Stations 3a and 3b*

Station	Model	Settings	MSE (°C)	RMSE (°C)	MAPE (%)
3a	SARIMA	Auto	1.65	1.28	6.45
3a	ETS	Auto	2.35	1.53	8.32
3a	LSTM	(50 units, 50 epochs)	0.46	0.68	3.27
3b	SARIMA	Auto	0.05	0.23	2.77
3b	ETS	Auto	0.03	0.17	2.12
3b	LSTM	(50 units, 50 epochs)	0.01	0.07	0.83

Source: authors calculation based on CalCOFI data.

Table 5, Table 6, and Table 7 summarise forecasting performance across all stations and depths. The Long Short-Term Memory (LSTM) model consistently achieved the lowest MSE, RMSE, and MAPE at every station, confirming its superior ability to capture nonlinear and seasonal temperature dynamics. Compared with SARIMA and ETS, LSTM reduced MSE substantially at both surface (a) and deep-water (b) stations. For example, at Station 2a, MSE decreased from 1.76 (SARIMA/ETS) to 0.26 with LSTM, while at Station 3b, MSE decreased from 0.05 (SARIMA) and 0.03 (ETS) to 0.01.

Surface stations (1a, 2a, 3a) exhibited higher forecasting errors than their corresponding deep-water stations, indicating greater variability at 0 m compared to 500 m depth. Station 3a presented the most challenging forecasting conditions, where ETS produced the highest MAPE (8.32%), whereas LSTM reduced the error to 3.27%, demonstrating improved robustness under higher variability.

Across deep-water stations (1b, 2b, 3b), overall error magnitudes were considerably lower, with LSTM achieving MAPE values below 1.2% in all cases and as low as 0.83% at Station 3b. While ETS slightly outperformed SARIMA at some deep stations (e.g., 3b), both statistical models were consistently outperformed by LSTM.

These results confirm that deep learning provides substantial performance gains over traditional statistical approaches for ocean-temperature forecasting across multiple spatial locations and depths.

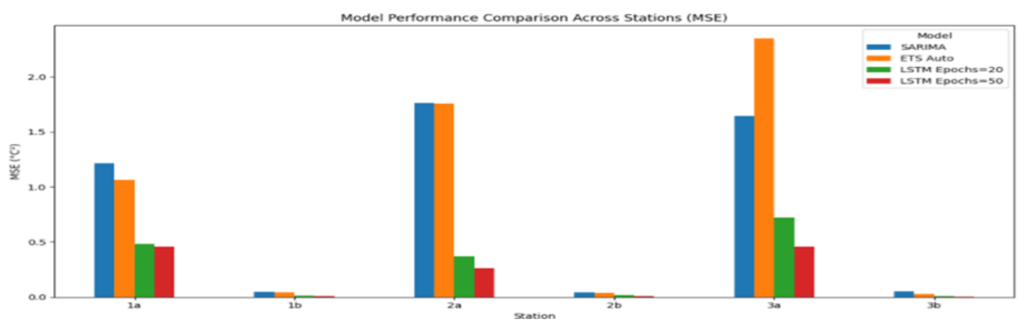

Figure 4. *Model comparison across Stations based on MSE (°C)*

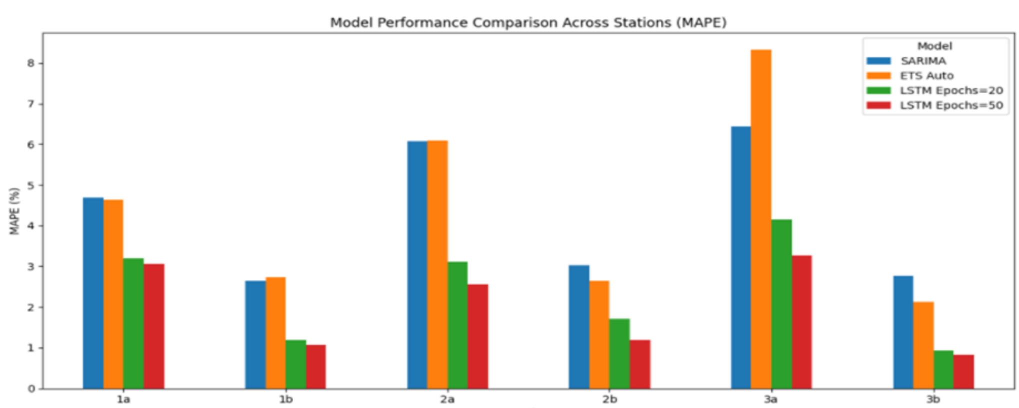

Figure 5. *Model comparison across Stations based on MAPE (%)*

Table 5–Table 7 and Figure 4–Figure 5 summaries performance across all stations and depths, where "a" denotes the surface and "b" 500 m depth. AutoETS generally outperformed AutoSARIMA in MSE and MAPE at Stations 1a, 1b, 2a, 2b, and 3b, with Station 3a

as an exception, where SARIMA performed better. LSTM (50 epochs) consistently achieved the lowest MSE, RMSE, and MAPE across all stations and depths, demonstrating superior pattern capture. LSTM (20 epochs) outperformed SARIMA and ETS at most stations but remained slightly weaker than the 50-epoch model. RMSE trends mirrored MSE, with lower errors at deeper stations (1b and 3b), likely due to reduced variability at 500 m. For MAPE, ETS generally improved upon SARIMA, but LSTM (50 epochs) consistently produced the lowest values, indicating the most precise predictions under both surface and deep-water conditions, particularly at the more challenging Station 3a.

4. Discussion

In this study, LSTM outperformed SARIMA and ETS in forecasting monthly ocean-temperature time series across the analysed stations and depths. This result is consistent with the theoretical capability of LSTM networks to model sequential data with long-term temporal dependencies, as the architecture was originally developed to address limitations such as vanishing gradients in recurrent neural networks (Hochreiter & Schmidhuber, 1997). Previous studies have also reported the usefulness of LSTM-based approaches for temperature forecasting, where decomposition-based Bi-LSTM models improved prediction accuracy by separating trend, seasonal, and residual components (Zhang et al., 2023). In the present study, the observed seasonal structure and nonlinear variability in ocean-temperature records provide a reasonable explanation for the stronger performance of LSTM compared with SARIMA and ETS. More broadly, ocean-temperature prediction is important because marine temperature variability is closely linked to ocean warming and ecological change (Venegas et al., 2023).

The consistent reduction in MSE, RMSE, and MAPE achieved by LSTM across most stations in this study indicates that deep-learning models can provide robust predictions when ocean-temperature series contain nonlinear variability and temporal dependence. Previous studies have similarly demonstrated the suitability of LSTM for temperature and environmental forecasting applications, where the model showed strong capability in learning temporal patterns from sequential data (Liu & Li, 2023; Salman et al., 2018). In the present study, the improved performance of the 50-epoch LSTM compared with the 20-epoch configuration further suggests that sufficient training helped the model better learn the temporal structure of the ocean-temperature series.

The comparative performance of SARIMA and ETS also provides useful insights. ETS generally produced smoother forecasts and achieved competitive performance under relatively stable conditions, whereas SARIMA demonstrated stronger capability in capturing explicit seasonal and autoregressive structures, particularly where periodic patterns were pronounced (Hyndman & Athanasopoulos, 2018). In the present study, both models showed reduced performance under stations with greater variability and more complex temperature fluctuations, which may explain the higher forecasting errors observed at some surface locations. Similar comparative behaviour between SARIMA- and ETS-based forecasting approaches has been reported across different application domains, where model performance depended on the characteristics and variability of the time series (Gao, 2025; Xian et al., 2023; Kuan, 2022).

An important observation from this study is the difference in forecasting performance between surface (0 m) and deep-water (500 m) stations. Surface stations exhibited higher variability and consequently higher prediction errors, likely reflecting the stronger influence of seasonal and environmental variability on near-surface ocean conditions. In contrast, deep-water stations exhibited more stable temperature profiles, resulting in consistently lower RMSE and MAPE values at 500 m depth across all forecasting models.

Overall, the results demonstrate that while SARIMA and ETS remain valuable for interpretable and computationally efficient forecasting, LSTM provides a more flexible framework for modelling complex ecological time series. In the present study, LSTM maintained strong predictive performance across different stations and depths, suggesting improved capability in handling nonlinear and variable ocean-temperature dynamics. The consistently lower MSE, RMSE, and MAPE values achieved by LSTM further support its suitability for ocean-temperature forecasting under varying environmental conditions. These findings suggest that deep-learning approaches may provide advantages for large-scale oceanographic prediction systems and environmental monitoring applications.

Automatic SARIMA achieved a lower AIC than the manually specified SARIMA model, indicating that automated order selection can improve modelling efficiency and support more consistent parameter selection across multiple time series (Arlt & Trčka, 2021). Similarly, the automatic ETS framework selected models using AIC-based optimisation, demonstrating the usefulness of automated model-selection procedures for identifying suitable forecasting configurations across different datasets.

In this study, ETS generally outperformed SARIMA at most stations, particularly under relatively stable conditions, although SARIMA occasionally achieved lower errors at certain locations such as Station 3a. Similar comparative behaviour between ETS- and SARIMA-based forecasting approaches has been reported in previous forecasting studies, where model performance varied depending on the variability and structure of the time series (Gao, 2025; Xian et al., 2023; Kuan, 2022). Mixed forecasting performance between ETS and SARIMA has also been observed in unemployment prediction studies, where SARIMA occasionally outperformed ETS under specific temporal conditions (Davidescu et al., 2021).

Finally, LSTM, particularly with 50 training epochs, achieved the strongest forecasting performance across most stations and depths in this study. The lower MSE, RMSE, and MAPE values obtained by LSTM suggest that deep-learning approaches may provide advantages in capturing nonlinear variability and long-term temporal dependencies in ocean-temperature data. Similar advantages of LSTM-based forecasting models have been reported in temperature, weather, and energy forecasting applications, where deep-learning approaches demonstrated strong capability in learning complex temporal patterns from sequential data (Liu & Li, 2023; Salman et al., 2018; Bilgili & Pinar, 2023). In the present study, the strong performance of LSTM at more variable stations, such as Station 3a, further highlights its robustness under complex environmental conditions.

5. Conclusion

This study compared traditional statistical models (SARIMA and ETS) with a deep-learning model (LSTM) for ocean-temperature forecasting at multiple stations and depths in the CalCOFI region. Automatic SARIMA improved on manual SARIMA in terms of AIC, and ETS generally outperformed SARIMA across stations. However, neither matched the performance of LSTM. Particularly, LSTM with 50 training epochs, consistently achieved the lowest MSE, RMSE, and MAPE, supporting its ability to capture nonlinear patterns and long-term dependencies in ecological time series. SARIMA and ETS remain useful for simpler datasets or situations where computational cost and interpretability are critical, whereas LSTM emerges as the most robust option for complex ocean-temperature prediction in the present study. These findings underline the advantages of deep learning for ecological forecasting and highlight its potential to advance predictive modelling in environmental applications. They also emphasise the importance of selecting appropriate modelling approaches for long-term environmental monitoring and climate-related decision-making.

The forecasting framework developed in this study may support detailed analysis of ocean-temperature behaviour with depth, including long-term trends, seasonal variability, and temporal patterns in the CalCOFI region and similar marine monitoring programmes (O'Donncha et al., 2022). Accurate forecasting at both surface (0 m) and deep-water (500 m) stations may also assist marine ecosystem assessments and climate-related environmental studies (Wei & Guan, 2022). Furthermore, LSTM-based forecasting approaches may contribute to early-warning systems by helping identify anomalous or extreme temperature conditions in marine environments (Bonino et al., 2024). Integration of these forecasting models into AI-driven environmental decision-support systems could additionally support planning and optimisation in coastal resource management and environmental monitoring.

The datasets used were relatively small, so AutoSARIMA and AutoETS procedures were efficient; however, for larger datasets and higher model orders, parameter estimation and model selection may become more computationally demanding. Strategies for choosing regular and seasonal differencing orders, especially in more complex series, still require refinement.

The present work does not fully address real-world complexities such as extensive data transformations, outlier treatment, or structural breaks, all of which may influence performance and limit generalisation to other regions or higher-frequency data.

Future research should extend this analysis to larger and more diverse datasets to better understand model behaviour under different temporal and spatial conditions. Further optimisation of LSTM hyperparameters, such as batch size, number of layers, and epochs, could improve performance and stability.

Hybrid models that combine SARIMA or ETS with LSTM or other deep-learning architectures are a promising direction. Such approaches could exploit the interpretability of statistical models together with the nonlinear learning capacity of deep networks to produce more robust ocean-temperature forecasts.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. doi: 10.1109/TAC.1974.1100705

Al-Sharafi, M. A., Al-Emran, M., Al-Kabi, M. N. & Shaalan, K. (Eds.). (2023). *Proceedings of the 2nd International Conference on Emerging Technologies and Intelligent Systems (ICETIS 2022), Volume 1*. Springer. doi: 10.1007/978-3-031-25274-7

Arlt, J. & Trčka, P. (2021). Automatic SARIMA modeling and forecast accuracy. *Communications in Statistics – Simulation and Computation*, 50(10), 2949–2970. doi: 10.1080/03610918.2019.1618471

Bilgili, M. & Pinar, E. (2023). Gross electricity consumption forecasting using LSTM and SARIMA approaches: A case study of Türkiye. *Energy*, 284, 128575. doi: 10.1016/j.energy.2023.128575

Bonino, G., Galimberti, G., Masina, S., McAdam, R. & Clementi, E. (2024). Machine learning methods to predict sea surface temperature and marine heatwave occurrence: A case study of the Mediterranean Sea. *Ocean Science*, 20(2), 417–432. doi: 10.5194/os-20-417-2024

Box, G. E. P. & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. Holden-Day.

Buczyński, M., Chlebus, M., Kopczewska, K. & Zajenkowski, M. (2023). Financial time series models—comprehensive review of deep learning approaches and practical recommendations. *Engineering Proceedings*, 39(1), 79. doi: 10.3390/engproc2023039079

CalCOFI. (2025). *California Cooperative Oceanic Fisheries Investigations*. https://calcofi.org

Caña, J. P. L., Herrero, J. G. & López, J. M. M. (2021). Forecasting nonlinear systems with LSTM: Analysis and comparison with EKF. *Sensors*, 21(5), 1805. doi: 10.3390/s21051805

Castel, S. & Burr, W. S. (2021). Assessing statistical performance of time series interpolators. *Engineering Proceedings*, 5(1), 57. doi: 10.3390/engproc2021005057

Chai, T. & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250. doi: 10.5194/gmd-7-1247-2014

Davidescu, A. A., Apostu, S. A. & Paul, A. (2021). Comparative analysis of different univariate forecasting methods in modelling and predicting the Romanian unemployment rate for the period 2021–2022. *Entropy*, 23(3), 325. doi: 10.3390/e23030325

Dickey, D. A. & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366), 427–431. doi: 10.1080/01621459.1979.10482531

Fatima, S. S. W. & Rahimi, A. (2024). A review of time-series forecasting algorithms for industrial manufacturing systems. *Machines*, 12(6), 380. doi: 10.3390/machines12060380

Gao, Y. (2025). A comparative study of ARIMA and ETS models for time series forecasting. *Advances in Economics, Management and Political Sciences*, 149(1), 196–201. doi: 10.54254/2754-1169/2024.19252

Gudžiūnaitė, S., Shabani, Z., Weitensfelder, L. & Moshammer, H. (2023). Time series analysis in environmental epidemiology: Challenges and considerations. *International Journal of Occupational Medicine and Environmental Health*, 36(6), 704–716. doi: 10.13075/ijomeh.1896.02237

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hyndman, R. J. & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts. https://otexts.com/fpp2/

Hyndman, R. J. & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3), 1–22. doi: 10.18637/jss.v027.i03

Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. doi: 10.1016/j.ijforecast.2006.03.001

Kong, Y., Zhou, T., Liang, Y., Sun, P., Wang, Z., Nie, Y., Zohren, S. & Wen, Q. (2024). Unlocking the power of LSTM for long term time series forecasting. *arXiv*. doi: 10.48550/arXiv.2408.10006

Kuan, M. M. (2022). Applying SARIMA, ETS, and hybrid models for prediction of tuberculosis incidence rate in Taiwan. *PeerJ*, 10, e13117. doi: 10.7717/peerj.13117

Lepot, M., Aubin, J. B. & Clemens, F. (2017). Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, 9(10), 796. doi: 10.3390/w9100796

Liu, T. & Li, X. (2023). Research on temperature prediction based on LSTM. *Academic Journal of Science and Technology*, 6(3), 54–55. doi: 10.54097/ajst.v6i3.10325

Ljung, G. M. & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297–303. doi: 10.1093/biomet/65.2.297

O'Donncha, F., Hu, Y., Palmes, P., Burke, M., Filgueira, R. & Grant, J. (2022). A spatio-temporal LSTM model to forecast across multiple temporal and spatial scales. *Ecological Informatics*, 69, 101687. doi: 10.1016/j.ecoinf.2022.101687

Salman, A. G., Heryadi, Y., Abdurahman, E. & Suparta, W. (2018). Weather forecasting using merged long short-term memory (LSTM) and autoregressive integrated moving average (ARIMA) model. *Journal of Computer Science*, 14(7), 930–938. doi: 10.3844/jcssp.2018.930.938

Singh Bist, A., Rawat, B., Joshi, Y., Aini, Q., Santoso, N. P. L. & Kusumawardhani, D. A. R. (2024). Harnessing deep learning for accurate climate change predictions. In *3rd International Conference on Creative Communication and Innovative Technology*. IEEE. https://ieeexplore.ieee.org/document/10701199

Tadesse, K. B. & Dinka, M. O. (2017). Application of SARIMA model to forecasting monthly flows in Waterval River, South Africa. *Journal of Water and Land Development*, 35, 229–236. doi: 10.1515/jwld-2017-0088

Tomov, L., Chervenkov, L., Miteva, D. G., Batselova, H. & Velikova, T. (2023). Applications of time series analysis in epidemiology: Literature review and our experience during COVID-19 pandemic. *World Journal of Clinical Cases*, 11(29), 6974–6983. doi: 10.12998/wjcc.v11.i29.6974

Usmani, M., Memon, Z. A., Zulfiqar, A. & Qureshi, R. (2024). Preptimize: Automation of time series data preprocessing and forecasting. *Algorithms*, 17(8), 332. doi: 10.3390/a17080332

Venegas, R. M., Acevedo, J. & Treml, E. A. (2023). Three decades of ocean warming impacts on marine ecosystems: A review and perspective. *Deep Sea Research Part II: Topical Studies in Oceanography*, 212, 105318. doi: 10.1016/j.dsr2.2023.105318

Wei, L. & Guan, L. (2022). Seven-day sea surface temperature prediction using a 3DConv-LSTM model. *Frontiers in Marine Science*, 9, 905848. doi: 10.3389/fmars.2022.905848

Xian, X., Wang, L., Wu, X., Tang, X., Zhai, X., Yu, R., Qu, L. & Ye, M. (2023). Comparison of SARIMA model, Holt-Winters model and ETS model in predicting the incidence of foodborne disease. *BMC Infectious Diseases*, 23(1), 803. doi: 10.1186/s12879-023-08799-4

Zhang, K., Huo, X. & Shao, K. (2023). Temperature time series prediction model based on time series decomposition and Bi-LSTM network. *Mathematics*, 11(9), 2060. doi: 10.3390/math11092060

Usporedba predviđanja temperature oceana korištenjem SARIMA, ETS i LSTM modela

VRSTA ČLANKA
Prethodno priopćenje

INFORMACIJE O ČLANKU
Primljeno: 21. travnja 2026.
Prihvaćeno: 22. svibnja 2026.
DOI: 10.62366/crebss.2026.1.002
JEL: C22, C45, Q54

SAŽETAK
Točno predviđanje temperature oceana ključno je za razumijevanje dugoročne okolišne varijabilnosti i podršku ekološkom odlučivanju. Ovaj rad ocjenjuje uspješnost modela SARIMA (*sezonski autoregresijski integrirani model pomičnih prosjeka*), ETS (*model pogreške, trenda i sezonalnosti*) i LSTM (*model duge kratkoročne memorije*) u predviđanju temperature oceana korištenjem povijesnih vremenskih nizova podataka iz programa CalCOFI. Skup podataka pripremljen je primjenom sezonske dekompozicije i analize stacionarnosti, a točnost predviđanja ocijenjena je pomoću pokazatelja MAPE (*srednja apsolutna postotna pogreška*), RMSE (*korijen srednje kvadratne pogreške*) i MSE (*srednja kvadratna pogreška*). Rezultati pokazuju da je LSTM ukupno dao najtočnija i najstabilnija predviđanja, ostvarujući niže vrijednosti RMSE i MAPE na većini postaja i dubina. Model je uspješno zabilježio nelinearno ponašanje, sezonsku varijabilnost i ekstremne temperaturne fluktuacije. SARIMA je pokazala snažnu sposobnost modeliranja trenda i sezonalnosti, dok je ETS generirao glađa predviđanja, ali s razmjerno višim vrijednostima pogreške. Dijagnostika reziduala dodatno je potvrdila superiornu sposobnost LSTM–a da nauči složene vremenske ovisnosti prisutne u podacima o temperaturi oceana. Ovaj rad doprinosi usporednoj ocjeni klasičnih statističkih modela i modela dubokog učenja za predviđanje ekoloških vremenskih nizova te pruža dokaze koji podupiru primjenu LSTM–a za poboljšano oceanografsko predviđanje i praćenje okoliša.

KLJUČNE RIJEČI
Ekološki vremenski nizovi, ETS, predviđanje, LSTM, SARIMA