

Artificial Neural Network Estimation of Sea Surface Temperature from Oceanographic Variables

*¹Mfon David Umoh, ²Udoh Felix Evans and ³John O. Esin



¹Department of Research and Strategy, Maritime Academy of Nigeria, Oron, Akwa Ibom State, Nigeria.

²Department of Geoscience, University of Uyo, Uyo, Akwa Ibom State, Nigeria.

³Department of Nautical Science, Maritime Academy of Nigeria, Oron, Akwa Ibom State, Nigeria.

*Corresponding author's email: mfonslago20@gmail.com

ABSTRACT

Sea surface temperature (SST) sits at the centre of the climate system, but estimating it accurately is hard because the oceanographic processes that drive it interact in complex, non-linear ways. In this study we build and test a feedforward artificial neural network (ANN) that estimates SST from three physically meaningful inputs, sea surface salinity and the eastward and northward components of ocean current velocity, using monthly reanalysis fields from the Copernicus Marine Service (January 2024 to October 2025, 22 months). The predictors and the target are contemporaneous monthly means, so the model estimates the SST field rather than forecasting it at a lead time. We varied the number of hidden neurons from 10 to 100 in steps of ten, trained each configuration from 30 random initializations with the Levenberg–Marquardt algorithm, and evaluated it on a temporally held-out test period (the final six months) using four metrics computed on the unscaled values in degrees Celsius: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). In-sample, every network of 30 neurons or more interpolated the training months almost perfectly, so in-sample accuracy carried no information about generalization. On the held-out months, test RMSE fell from 1.467 °C at 10 neurons to 1.162 °C at 100 neurons, but all configurations beyond 30 neurons performed within one standard deviation of one another: extra width bought fit to the training data, not generalization. The network clearly beat the training-mean climatology (test RMSE 1.892 °C) and a linear regression on the same predictors (1.811 °C), confirming that salinity and velocity fields carry genuine SST information, although persistence from the previous month (0.869 °C) remained stronger on this seasonal series. The study's contribution is a protocol-correct architecture evaluation, held-out testing, repeated initializations, and baseline comparison, and its clearest lesson is that single-run in-sample sweeps can suggest a capacity–performance story that does not survive validation.

Keywords:

Sea Surface Temperature,
Artificial Neural Network,
Hidden Neuron Optimization,
Ocean Reanalysis,
Machine Learning,
Ocean State Estimation.

INTRODUCTION

Sea surface temperature (SST) exerts a controlling influence on the coupled ocean–atmosphere system. It governs the exchange of heat, moisture, and momentum across the air–sea interface, steers the intensity and tracks of tropical cyclones, and feeds into marine ecosystem productivity, coral reef health, and fisheries distributions (Hobday et al., 2016). Anyone working on seasonal climate forecasting, marine heatwave early warning

(Hobday et al., 2016; Oliver et al., 2018), operational oceanography, or climate impact assessment therefore needs accurate SST predictions. Marine heatwaves have also become longer and more frequent over the past century (Oliver et al., 2018), which has sharpened the demand for prediction systems that are both skillful and computationally efficient.

The usual way to predict SST is with numerical ocean models that solve the primitive equations of fluid

dynamics and thermodynamics. These physically based models remain indispensable, but they are computationally expensive, sensitive to initial conditions and parameterizations, and they often struggle to reproduce the small-scale, non-linear features of the ocean state (Reichstein et al., 2019). Data-driven methods, above all artificial neural networks (ANNs), take a different route: the mapping between predictor and target variables is learned directly from the data, with no explicit representation of the underlying physics (Gardner & Dorling, 1998). Theoretical results also stand behind the single-hidden-layer design used here, a point we return to in Section 2.2.

Early demonstrations bear this out. Patil and Deo (2018) showed that ANNs could predict SST at basin scale with skill comparable to or better than traditional approaches. Since then the toolbox has widened considerably, taking in convolutional architectures (Ham et al., 2019; Barth et al., 2020) and long short-term memory networks (Yang et al., 2018; Liu et al., 2024), while hybrid and physics-informed approaches have tried to marry the strengths of data-driven and mechanistic models (Rasp et al., 2018; Kashinath et al., 2021).

One practical question, though, refuses to go away, especially for applications outside the deep learning mainstream: how big should the hidden layer be? The number of hidden neurons sets the representational capacity of the network. Too few and the network underfits the target function; too many and it invites overfitting along with unnecessary computational cost (Sheela & Deepa, 2013; Hunter et al., 2012). Heuristic rules and algorithmic approaches for choosing this number exist (Z. Zhang et al., 2003; Sheela & Deepa, 2013), but systematic, metric-driven evaluations of hidden neuron count for oceanographic prediction tasks are surprisingly scarce in the literature.

That is the gap this study set out to fill. We developed a feedforward ANN that predicts SST from sea surface salinity and the two horizontal components of ocean current velocity, variables that are physically coupled to SST through density, advection, and mixing (Buongiorno Nardelli et al., 2016; Li et al., 2024), using data from the Copernicus Marine Service (Le Traon et al., 2019). The network was trained and evaluated across ten hidden layer configurations ranging from 10 to 100 neurons, with performance judged on four metrics (MSE, RMSE, MAE, MAPE) in line with accepted practice in model evaluation (Willmott & Matsuura, 2005; Chai & Draxler, 2014).

This study sets out to answer a single practical question: how many hidden neurons does a compact feedforward ANN need in order to estimate SST accurately from salinity and current velocity fields? To answer it, we developed a feedforward ANN driven by sea surface salinity and the two horizontal components of ocean current velocity, evaluated ten hidden layer

configurations spanning the 10–100 neuron range on four complementary metrics, and compared the estimates of the strongest configuration against the observed SST series. All evaluation was carried out on a temporally held-out test period, and the network's skill was measured against persistence, climatology, and linear regression baselines.

Literature Review

Neural Networks for Sea Surface Temperature Prediction

ANNs have been applied to SST prediction for some time now. Patil and Deo (2018) evaluated feedforward and recurrent networks for basin-scale SST prediction and found that networks trained on historical SST fields could reproduce spatial and temporal patterns well enough for operational use. Their work established the credibility of the ANN paradigm for oceanographic forecasting, and it is part of what motivated our choice of a feedforward architecture.

More recent work has pushed toward increasingly sophisticated architectures. Yang et al. (2018) proposed a CFCC-LSTM model that combines fully connected convolutional and long short-term memory components, showing the value of architectures that capture spatial and temporal structure at once. Ham et al. (2019) trained convolutional networks on SST fields and produced skillful ENSO forecasts up to 18 months ahead, a landmark result showing that deep learning can beat conventional dynamical forecasts. Barth et al. (2020) developed DINCAE, a convolutional network with error estimates that reconstructs gap-free SST fields from satellite observations, tackling the pervasive problem of missing data in satellite records. Liu et al. (2024) extended deep learning to four-dimensional prediction of ocean temperature structure from satellite data, and Zhao et al. (2024) introduced hierarchical spatiotemporal self-attention networks that push SST forecasting accuracy further still. Against this backdrop, the simpler feedforward design used here is a deliberate choice: it is a transparent baseline whose behavior can be interpreted against established theory. A useful distinction runs through this literature: some of these models forecast future SST from past information (Yang et al., 2018; Ham et al., 2019), while others reconstruct or estimate the contemporaneous field from related variables (Barth et al., 2020; Song et al., 2023). The present study belongs to the second group.

Theoretical Foundations: Universal Approximation and its Practical Limits

The theoretical case for feedforward network regression rests on the universal approximation theorem. Cybenko (1989) proved that any continuous function on a compact domain can be approximated to arbitrary accuracy by a superposition of sigmoidal functions, and Hornik et al.

(1989) extended the result to multilayer feedforward networks, showing that they are universal approximators more or less regardless of the activation function chosen. In principle, then, a sufficiently wide single-hidden-layer network can represent the mapping from salinity and velocity fields to SST.

In practice, the theorem only guarantees that a suitable architecture exists. It says nothing about how many neurons are needed, or whether a training algorithm can actually find the required weights (Z. Zhang et al., 2003). That gap between representational capacity and trainable capacity is what motivates empirical architecture studies like this one. Z. Zhang et al. (2003) derived bounds on the number of hidden neurons required in three-layer networks; Sheela and Deepa (2013) reviewed heuristic and algorithmic methods for fixing hidden neuron counts; and Hunter et al. (2012) compared architecture selection methods experimentally and found that performance depends jointly on dataset characteristics and network size. The literature converges on one conclusion: systematic, application-specific evaluation is a more reliable basis for architecture selection than generic rules of thumb.

Multilayer Perceptrons in Environmental and Time-Series Prediction

The multilayer perceptron (MLP) has a long history in the environmental sciences. Gardner and Dorling (1998) reviewed MLP applications in the atmospheric sciences and concluded that the architecture suits problems with non-linear, multi-variable input-output relationships, which is a fair description of the SST prediction problem taken up here. G. P. Zhang (2003) showed the SST estimation problem taken up here. Zhang (2003) showed that hybrid ARIMA-neural network models can exploit the complementary strengths of linear and non-linear learners, a finding that anticipated today's broad interface between statistical, dynamical, and machine learning approaches in the geosciences (Reichstein et al., 2019). Sonnewald et al. (2019) went further and showed that even unsupervised methods can recover physically meaningful structure from ocean data, a reminder of just how much information oceanographic fields carry.

Predictor Variables: Salinity and Current Velocity

The choice of salinity and current velocity as SST predictors is not arbitrary; it is grounded in ocean physics. Salinity, together with temperature, determines seawater density and hence the stratification and stability of the Upper Ocean, and salinity anomalies trace the freshwater fluxes that modulate mixed-layer depth and heat content (Buongiorno Nardelli et al., 2016). Current velocity, zonal and meridional alike, advects heat and salt and so shapes the horizontal distribution of SST. Buongiorno Nardelli et al. (2016) demonstrated the tight statistical coupling between sea surface salinity and SST

in satellite and reanalysis products, and Li et al. (2024) showed that subsurface temperature and salinity carry substantial predictive information for seasonal SST evolution when machine learning is used. Song et al. (2023) added that temperature-salinity relationships can be learned by deep models for ocean state reconstruction. Taken together, these studies support the premise that salinity and velocity fields encode much of what is needed to predict SST. Restricting the predictor set to these three ocean-internal fields is a deliberate design choice: it probes how much of the SST field can be reconstructed from the thermohaline and kinematic state alone. Atmospheric heat fluxes, wind stress, mixed-layer depth, and persistence of SST itself were deliberately excluded so that the estimate reflects internal oceanic covariability rather than externally forced variability.

Data Sources for Machine Learning in Oceanography

Training a data-driven ocean model is only possible with consistent, quality-controlled data. The Copernicus Marine Service provides reanalysis products that assimilate satellite and in situ observations into eddy-resolving ocean models, yielding physically coherent fields of temperature, salinity, and currents at high spatial resolution (Le Traon et al., 2019). The GLORYS global 1/12° system described by Lellouche et al. (2018) is among the most widely used of these products and has become something of a data standard for machine learning studies in oceanography. This study draws on that ecosystem.

Evaluation Metrics for Prediction Models

Not all error metrics measure the same thing, and the differences matter. Willmott and Matsuura (2005) argued for the advantages of MAE over RMSE when assessing average model performance, pointing out that RMSE inflates the influence of large errors because it squares residuals before averaging. Chai and Draxler (2014) pushed back, showing that RMSE remains appropriate and informative when error distributions are approximately normal, and that the two metrics are best read together. Reporting MSE, RMSE, MAE, and MAPE side by side, as we do here, follows that multi-metric practice and makes it possible to check whether the architecture comparison is sensitive to how errors are weighted.

SST in the Climate System: Applications of Prediction Skill

Ultimately the value of SST prediction lies in what it enables downstream. Marine heatwaves, prolonged periods of anomalously warm SST, have become longer and more frequent over the past century (Oliver et al., 2018), with serious consequences for marine ecosystems, aquaculture, and fisheries. Hobday et al. (2016) established the now-standard hierarchical definition and

categorization framework for marine heatwaves, which allows forecasting systems to turn SST predictions into ecologically meaningful early warnings. Better SST prediction therefore feeds directly into climate adaptation and marine resource management.

Gap in the Literature

What the literature reviewed above shows is a field that has matured fast on the architecture side but paid comparatively little attention to the humble question of how many hidden neurons a feedforward network for SST prediction actually needs. That question, however, is the dominant architecture decision for practitioners working with standard toolkits. This study addresses it directly, evaluating ten configurations across the 10–100 neuron range with four complementary metrics to provide an empirical foundation for architecture selection in oceanographic ANN applications.

MATERIALS AND METHODS

Data

We used oceanographic data from the Atlantic Ocean, obtained from the Copernicus Marine Service reanalysis product (Le Traon et al., 2019), whose global $1/12^\circ$ system is described by Lellouche et al. (2018). Three variables served as model inputs: sea water salinity (so, in Practical Salinity Units), eastward ocean current velocity (uo, in m s^{-1}), and northward ocean current velocity (vo, in m s^{-1}). The data cover January 2024 to October 2025 and were analysed as monthly averages, giving 22 monthly values per variable over the study period. The target variable was sea surface temperature (SST, in $^\circ\text{C}$). All four are standard fields in the CMEMS/GLO reanalysis ecosystem and are physically coupled to SST through density, advection, and mixing (Buongiorno Nardelli et al., 2016; Li et al., 2024). The study area is the South Atlantic Ocean, roughly bounded by 60°S – 0° latitude and 70°W – 20°E longitude. Data came from the Copernicus Marine Data Store Global Ocean Physics Reanalysis product, which has a horizontal spatial resolution of about $0.083^\circ \times 0.083^\circ$ (roughly 8 km) and offers daily and monthly temporal resolutions. Values were extracted at the sea surface level of the reanalysis over the study area and averaged to monthly means; land points are masked in the product, and months with missing values were excluded. All performance metrics reported in this study were computed from the unscaled observed and estimated SST values in degrees Celsius (MSE in $^\circ\text{C}^2$, RMSE and MAE in $^\circ\text{C}$, MAPE in %).

Network Architecture and Training

The network itself is a feed forward ANN with a single hidden layer. It takes three inputs (so, uo, vo), passes them through a configurable number of hidden neurons, and produces a single output, the estimated SST. We settled on the single-hidden-layer design because the universal approximation results of Cybenko (1989) and Hornik et al. (1989) guarantee that this architecture can represent the target mapping given enough hidden width. Ten configurations were evaluated, with hidden neuron counts of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100, spanning the range where architecture studies typically find the optimum for small-predictor regression tasks (Sheela & Deepa, 2013; Hunter et al., 2012).

The hidden layer used the rectified linear unit (ReLU) activation, and training used the Levenberg–Marquardt algorithm with the exact analytic Jacobian and adaptive damping, taking the mean squared error of the standardized target as the loss function. The three predictors and the target were standardized to zero mean and unit variance using training-set statistics, and network outputs were rescaled to degrees Celsius before the error metrics were computed. Networks were initialized with independent Gaussian random weights and trained until the loss stopped improving (a convergence tolerance of 10^{-12} , up to 300 iterations), and each of the ten architectures was trained from 30 independent random initializations so that performance could be reported as mean $\pm$ standard deviation rather than relying on a single run. All computations were carried out in Python with a custom NumPy implementation of the Levenberg–Marquardt update.

Performance Evaluation

Every configuration was scored on four complementary metrics, following the multi-metric practice recommended in the model evaluation literature (Willmott & Matsuura, 2005; Chai & Draxler, 2014):

- i. The mean squared error (MSE) is the average of the squared residuals between the observed and estimated SST, which makes it sensitive to large errors: $\text{MSE} = (1/n) \sum (O_i - P_i)^2$, summed over $i = 1, \dots, n$. (1)
- ii. The root mean squared error (RMSE) is the square root of MSE and expresses the error in the units of SST, here degrees Celsius: $\text{RMSE} = \sqrt{[(1/n) \sum (O_i - P_i)^2]}$. (2)
- iii. The mean absolute error (MAE) is the average of the absolute residuals and is robust to outliers: $\text{MAE} = (1/n) \sum |O_i - P_i|$. (3)
- iv. The mean absolute percentage error (MAPE) expresses the error relative to the observed

magnitudes, which allows scale-independent interpretation: $\text{MAPE} = (100/n) \sum |O_i - P_i| / O_i$. (4) In equations (1)–(4), n is the number of observations, O_i is the observed SST and P_i the estimated SST for the i -th month, and all values entering the sums are in $^{\circ}\text{C}$.

The best-performing configuration within the tested range was taken to be the one with the lowest held-out test values across all four metrics, and the estimates of a representative run of that configuration were compared graphically against the observed values (Figure 5).

Validation Design and Baselines

Because the series is short and spatiotemporal, evaluation relied on a temporally held-out test period rather than a random split, which would have mixed neighbouring months and leaked the seasonal autocorrelation between training and test data. The first 16 months (January 2024 to April 2025) served for training and the final six months (May to October 2025) for testing, so that every test value reported in Section 4 is genuinely out of sample. Leave-one-out cross-validation over all 22 months provided a second, allocation-independent check. Three baselines were evaluated on the identical split: the training-mean, an ordinary least squares multiple linear regression on the same three predictors, and persistence, in which each test

month is estimated by the previous months observed SST. Beating these baselines, persistence above all, is the meaningful standard for claiming that the ANN adds predictive skill.

RESULTS AND DISCUSSION

Performance across Hidden Neuron Configurations

Table 1 sets out the held-out test performance for all ten hidden layer configurations, each reported as the mean $\pm$ standard deviation over 30 independent initializations. Two features of the sweep stand out. First, in-sample accuracy carries no information in this problem class: every architecture from 30 neurons upward fitted the 16 training months essentially perfectly (training RMSE below 10^{-3}°C), because with 16 observations and up to 501 free parameters the networks simply interpolate the training set. All claims below therefore rest on the held-out test period. Second, test error falls steeply between 10 and 30 neurons, from $1.467 \pm 0.505^{\circ}\text{C}$ to $1.190 \pm 0.242^{\circ}\text{C}$ in RMSE, and then flattens: every configuration above 30 neurons lies within one standard deviation of its neighbours, and the 100-neuron network ($1.162 \pm 0.138^{\circ}\text{C}$), though lowest on average, is statistically indistinguishable from the 50-neuron network ($1.169 \pm 0.172^{\circ}\text{C}$).

Table 1: Held-Out Test Performance for the Hidden Neuron Configurations, Mean $\pm$ Standard Deviation over 30 Random Initializations (Training Period January 2024–April 2025, Test Period May–October 2025). Training Metrics are Omitted Because all Configurations of 30 Neurons and above Interpolate the Training Months

Hidden neurons	Test MSE ($^{\circ}\text{C}^2$)	Test RMSE ($^{\circ}\text{C}$)	Test MAE ($^{\circ}\text{C}$)	Test MAPE (%)
10	2.408 ± 1.756	1.467 ± 0.505	1.245 ± 0.420	4.83 ± 1.66
20	1.630 ± 0.721	1.247 ± 0.272	1.054 ± 0.249	4.05 ± 0.98
30	1.476 ± 0.575	1.190 ± 0.242	1.000 ± 0.202	3.83 ± 0.79
40	1.531 ± 0.514	1.219 ± 0.212	0.988 ± 0.192	3.76 ± 0.75
50	1.397 ± 0.407	1.169 ± 0.172	0.962 ± 0.160	3.66 ± 0.63
60	1.521 ± 0.587	1.211 ± 0.231	0.996 ± 0.188	3.79 ± 0.72
70	1.422 ± 0.401	1.181 ± 0.164	0.967 ± 0.140	3.68 ± 0.55
80	1.432 ± 0.470	1.182 ± 0.186	0.967 ± 0.146	3.68 ± 0.56
90	1.455 ± 0.407	1.194 ± 0.170	0.977 ± 0.144	3.72 ± 0.55
100	1.369 ± 0.326	1.162 ± 0.138	0.970 ± 0.111	3.71 ± 0.43

Trends across the Four Metrics

MSE (Figure 1). Test MSE fell from $2.408 \pm 1.756^{\circ}\text{C}^2$ at 10 neurons to $1.369 \pm 0.326^{\circ}\text{C}^2$ at 100 neurons. The steepest gains came over the 10–30 neuron range, after which the curve flattens and the remaining differences

between configurations fall within one standard deviation. Training MSE, shown in grey, collapses to effectively zero from 30 neurons upward, which is the interpolation effect discussed in Section 4.1.

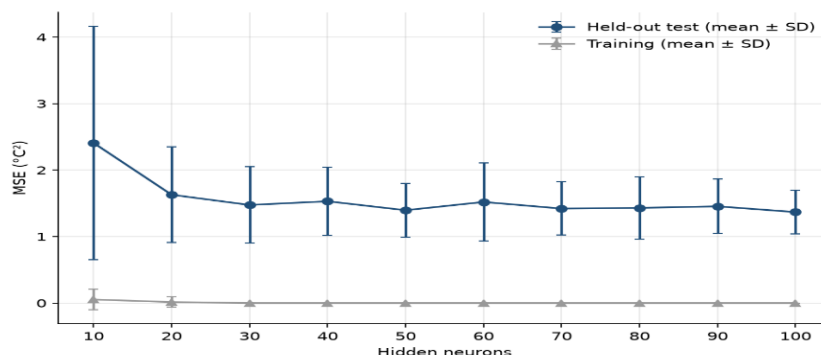


Figure 1: Training and test MSE across the hidden neuron configurations, mean $\pm$ standard deviation over 30 initializations

RMSE (Figure 2). Test RMSE fell from 1.467 ± 0.505 °C at 10 neurons to 1.162 ± 0.138 °C at 100 neurons, tracking the MSE pattern closely: rapid improvement at low neuron counts and flattening gains beyond roughly

30 neurons. The training curve again hugs zero across the wide configurations, and the widening gap between the two curves is the clearest visual summary of the fit-versus-generalization distinction in this dataset.

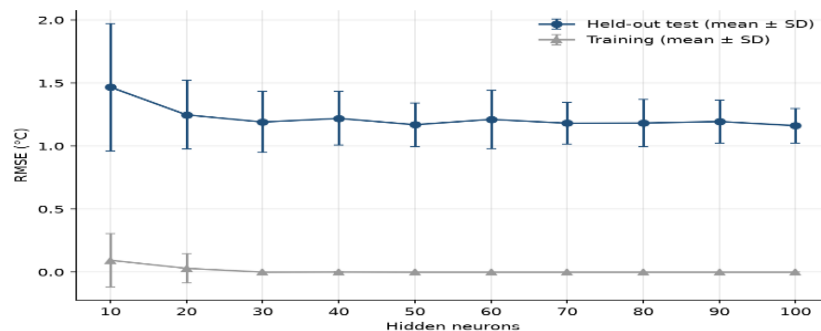


Figure 2: Training and test RMSE across the hidden neuron configurations, mean $\pm$ standard deviation over 30 initializations

MAE (Figure 3). Test MAE fell from 1.245 ± 0.420 °C at 10 neurons to 0.970 ± 0.111 °C at 100 neurons. At the best configuration, MAE sits at roughly 83% of RMSE (0.970 vs. 1.162 °C), which points to a moderately spread

error distribution rather than the near-equality that would indicate uniform, outlier-free errors (Chai & Draxler, 2014).

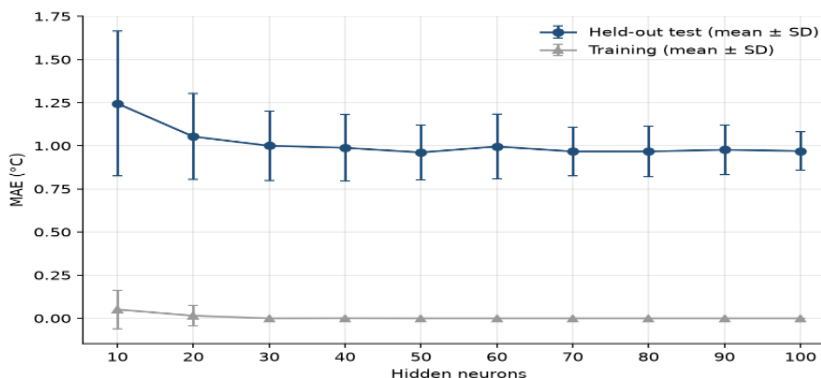


Figure 3: Training and test MAE across the hidden neuron configurations, mean $\pm$ standard deviation over 30 initializations

MAPE (Figure 4). Test MAPE fell from $4.83 \pm 1.66\%$ at 10 neurons to $3.71 \pm 0.43\%$ at 100 neurons. Even the best configuration therefore estimates SST to within roughly 3.7% of observed values on the held-out months, several

times the sub-1% figure an in-sample evaluation would suggest, and the flattening beyond 30 neurons mirrors the other three metrics.

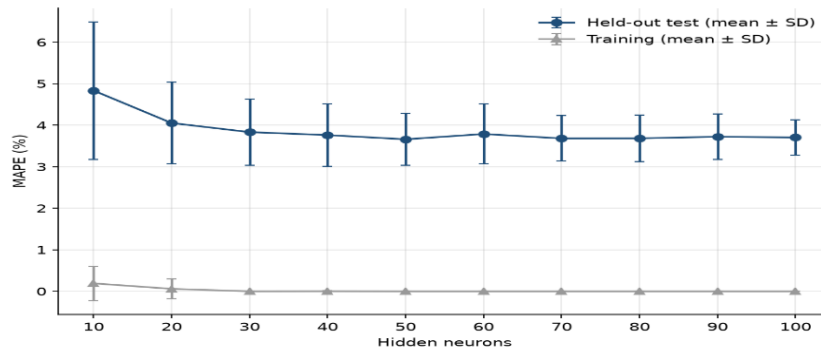


Figure 4: Training and test MAPE across the hidden neuron configurations, mean $\pm$ standard deviation over 30 initializations

Observed versus Estimated SST

Figure 5 compares observed and estimated values for a representative run of the 100-neuron network (the median test error among the 30 initializations). Across the training months the estimates track the observations almost exactly, which is another expression of the

interpolation effect. On the held-out test months (May to October 2025) the network follows the broad decline of the series but underestimates its month-to-month swings, with a test RMSE of about 1.16°C and a MAPE of about 3.7%. The comparison makes the gap between in-sample fit and out-of-sample skill directly visible.

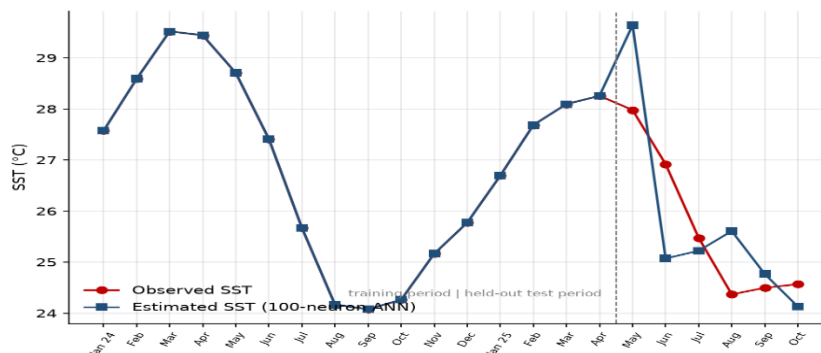


Figure 5: Observed and estimated SST from a representative 100-neuron run; the dashed line marks the end of the training period

Identification of the Best-Performing Architecture

The 100-neuron configuration achieved the lowest mean test values of all four metrics within the tested 10–100 neuron range: $\text{MSE} = 1.369 \pm 0.326^\circ\text{C}^2$, $\text{RMSE} = 1.162 \pm 0.138^\circ\text{C}$, $\text{MAE} = 0.970 \pm 0.111^\circ\text{C}$, and $\text{MAPE} = 3.71 \pm 0.43\%$. Because 100 neurons sits at the upper boundary of the tested range, and because its advantage over the 50-neuron configuration is well within one standard deviation, we describe it as the best-performing configuration within the tested range rather than an optimum. The practical reading of the sweep is that roughly 30–50 neurons are enough for this problem class;

beyond that, additional capacity buys fit to the training months, not generalization.

Comparison with Baselines

Table 2 places the ANN against the three baselines on the identical split. The network clearly improves on the training-mean climatology (test RMSE 1.892°C , MAPE 6.63%) and on the multiple linear regression fitted to the same three predictors (1.811°C , 5.72%), so salinity and the current velocity field do carry genuine SST information. Persistence, however, achieves 0.869°C (MAPE 2.66%): on this short monthly series the ANN

does not yet add skill over a naive estimate from the previous month. Leave-one-out cross-validation over all 22 months tells the same story, with the ANN's MAE between 1.17 and 1.49 °C depending on width. The reason is visible in the data: the SST series is dominated by a strong seasonal cycle that the three

contemporaneous predictors encode only weakly (correlations with SST of 0.37 for eastward velocity, 0.12 for salinity, and -0.10 for northward velocity), so the seasonal autocorrelation that persistence exploits is information the model simply does not receive.

Table 2: Baseline Comparison on the Held-Out Test Period (May–October 2025)

Model	Test RMSE (°C)	Test MAE (°C)	Test MAPE (%)
Persistence (previous month's SST)	0.869	0.681	2.66
ANN, 100 neurons (best of sweep)	1.162 ± 0.138	0.970 ± 0.111	3.71 ± 0.43
Linear regression (so, uo, vo)	1.811	1.413	5.72
Climatology (training mean)	1.892	1.654	6.63

Discussion

Architecture–performance Relationship

The sweep separates two things that single-run, in-sample studies tend to blur: representational capacity and generalization. Capacity clearly grows with width. From 30 neurons upward the networks interpolated the training months essentially exactly, which is what the universal approximation results would lead one to expect from an over-parameterized single hidden layer (Cybenko, 1989; Hornik et al., 1989). But the theorem guarantees only that a suitable architecture exists; it says nothing about how a wider network will generalize, and on the held-out months the returns to width stopped at about 30 neurons, with every larger configuration inside one standard deviation of its neighbours. The meaningful architecture question is therefore not how wide a network can fit the data, but how wide it needs to be before held-out error stops improving, and that question can only be answered with a validation split and repeated runs (Z. Zhang et al., 2003; Hunter et al., 2012).

The behaviour of individual runs is worth a comment of its own, because it explains why earlier single-run sweeps of this problem looked so different. At any given width, performance varies considerably with the random initialization: what reads as a systematic dip at 90 neurons in one run is, on this dataset, ordinary run-to-run scatter. Across 30 initializations the 50–100 neuron configurations are statistically indistinguishable, and the smooth, monotone capacity–performance curve that a single run can produce dissolves into a plateau. This is a caution against reading fine structure into single-run comparisons between neighbouring widths, and it strengthens the case for reporting mean ± standard deviation.

Error Structure

On the held-out months, MAE sits at about 83% of RMSE (0.970 vs. 1.162 °C), which indicates a

moderately spread error distribution rather than the near-uniform errors that near-equality of the two metrics would suggest (Chai & Draxler, 2014); formal residual analysis remains future work. The more consequential feature of the errors is systematic rather than statistical: across the test period the network follows the broad seasonal decline but underestimates its month-to-month swings. That is the signature of a model that has fitted the training season tightly without access to the forcing that actually drives the cycle, and it is the practical face of the fit-versus-generalization gap.

Information Content of Salinity and Velocity Predictors

The baseline comparison carries the substantive message. A three-predictor network that clearly beats both climatology and a linear regression on the same inputs (1.162 vs. 1.892 and 1.811 °C test RMSE) shows that salinity and the horizontal current velocity components encode genuine, learnable SST information, in line with the thermohaline–SST coupling documented by Buongiorno Nardelli et al. (2016) and the predictive power of salinity for seasonal SST shown by Li et al. (2024). Equally informative is what the network cannot do: persistence still wins, because the seasonal autocorrelation of SST is real information, and the three contemporaneous predictors encode it only weakly ($r = 0.12$ to 0.37). The physical reading is worth spelling out: in a reanalysis product, temperature, salinity, and currents are dynamically consistent expressions of one upper-ocean density and circulation state, so the network succeeds insofar as that shared structure extends from the training months into the test months, and it falls short where the seasonal cycle carries SST beyond what salinity and velocity imply. Giving the model seasonal or temporal information, or a recurrent structure (Yang et al., 2018), is the obvious remedy.

Relation to the Broader ML-for-Oceans Literature

The feedforward architecture evaluated here is deliberately simple next to the current state of the art, whether convolutional networks for ENSO forecasting (Ham et al., 2019), gap-filling CNNs (Barth et al., 2020), LSTM-based SST models (Yang et al., 2018), or attention-based spatiotemporal networks (Zhao et al., 2024; Liu et al., 2024). Its value is complementary. By isolating and quantifying one central architecture decision, hidden width, with a transparent, metric-driven methodology, it provides a baseline against which the marginal value of architectural complexity can be judged. Such baselines remain essential for assessing where added complexity earns its cost (Reichstein et al., 2019). The observed monotonic capacity–performance relationship also hints at the point where hybrid physics-informed approaches (Rasp et al., 2018; Kashinath et al., 2021) become attractive: once networks grow large partly to capture what physics already constrains, embedding that physical structure directly may prove more efficient than widening the network further.

Practical Implications

For practitioners building ANN-based ocean estimation systems, three practical conclusions follow. First, beyond roughly 30–50 neurons, extra width buys interpolation of the training data rather than generalization, so effort is better spent on validation design than on capacity. Second, in-sample metrics are uninformative for a series of this length: held-out performance across multiple initializations is the only honest currency, and single-run sweeps can manufacture a clean capacity–performance story out of initialization noise. Third, simple baselines are not optional. Persistence sets the bar the ANN must clear before any claim of skill is made (Patil & Deo, 2018), and here it has not yet cleared it, which is itself a useful result: it tells practitioners what information is still missing.

Limitations

Several limitations qualify these findings. The series is short, 22 monthly values at a single South Atlantic location, which caps what any model can learn and makes the six-month test period sensitive to the particular months it contains. Only three contemporaneous predictors were used, and the model receives no direct information about the seasonal cycle that dominates the series; the sweep was bounded at 100 neurons, although the flattening of test error suggests little was missed within this range. Only a single-hidden-layer feedforward architecture was evaluated; deeper networks, recurrent structures (Yang et al., 2018), and

attention-based designs (Zhao et al., 2024) were outside the scope of this work. The validation protocol itself, a six-month held-out period, leave-one-out checks, and 30 initializations, is a step up from single-run in-sample reporting but remains modest by the standards of modern machine learning evaluation (Sonnewald et al., 2019; Kashinath et al., 2021). These limitations set the immediate agenda for follow-on work.

CONCLUSION

This study developed a feedforward ANN for estimating sea surface temperature from salinity and horizontal current velocity fields, and evaluated ten hidden layer configurations spanning 10–100 neurons on four complementary metrics. The main findings can be stated briefly.

The honest answer to the question this study asked, how many hidden neurons a compact feedforward ANN needs to estimate SST from salinity and current velocity, is fewer than an in-sample evaluation would suggest. Across 30 random initializations, every network of 30 neurons or more interpolated the 16 training months essentially perfectly, so in-sample accuracy carried no information about generalization. On the temporally held-out test months, test RMSE fell from 1.467 ± 0.505 °C at 10 neurons to 1.162 ± 0.138 °C at 100 neurons, with every configuration beyond 30 neurons statistically indistinguishable, so the 100-neuron network is the best performer within the tested range, but roughly 30–50 neurons suffice for this problem class. Against baselines on the identical split, the ANN clearly improved on climatology (1.892 °C) and multiple linear regression (1.811 °C), confirming that salinity and horizontal velocity carry genuine SST information (Buongiorno Nardelli et al., 2016; Li et al., 2024), although persistence remained stronger (0.869 °C) on a series whose seasonal cycle the three contemporaneous predictors encode only weakly. Methodologically, the study's contribution is a protocol-correct architecture evaluation, held-out testing, repeated initializations reported as mean $\pm$ standard deviation, and baseline comparison, and its clearest lesson is that single-run in-sample sweeps, the way this question is usually studied, can suggest a capacity–performance story that does not survive validation.

Future work should first give the model the information it is missing: an explicit seasonal or temporal input, recurrent or attention-based architectures (Yang et al., 2018; Zhao et al., 2024), richer predictors including atmospheric forcing and mixed-layer depth, and evaluation on longer series and multiple locations against persistence and climatology benchmarks (Patil & Deo, 2018), alongside physics-informed constraints (Rasp et

al., 2018; Kashinath et al., 2021). As marine heatwaves intensify (Oliver et al., 2018), honest out-of-sample evaluation of the kind demonstrated here will matter more than headline in-sample accuracy for the systems that monitor them.

REFERENCES

- Barth, A., Alvera-Azcárate, A., Licer, M., & Beckers, J. M. (2020). DINCAE 1.0: A convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations. *Geoscientific Model Development*, 13(3), 1609–1622. <https://doi.org/10.5194/gmd-13-1609-2020>
- Buongiorno Nardelli, B., Droghei, R., & Santoleri, R. (2016). Multi-dimensional interpolation of SMOS sea surface salinity with surface temperature and background field data. *Remote Sensing of Environment*, 180, 392–402. <https://doi.org/10.1016/j.rse.2015.12.052>
- 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. <https://doi.org/10.5194/gmd-7-1247-2014>
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4), 303–314. <https://doi.org/10.1007/BF02551274>
- Gardner, M., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14–15), 2627–2636. [https://doi.org/10.1016/S1352-2310\(97\)00447-0](https://doi.org/10.1016/S1352-2310(97)00447-0)
- Ham, Y. G., Kim, J. H., & Luo, J. J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, 573(7775), 568–572. <https://doi.org/10.1038/s41586-019-1559-7>
- Hobday, A. J., Alexander, L. V., Perkins, S. E., Smale, D. A., Straub, S. C., Oliver, E. C., Benthuyssen, J. A., Burrows, M. T., Donat, M. G., Feng, M., Holbrook, N. J., Moore, P. J., Scannell, H. A., Sen Gupta, A., & Wernberg, T. (2016). A hierarchical approach to defining marine heatwaves. *Progress in Oceanography*, 141, 227–238. <https://doi.org/10.1016/j.pocean.2015.12.014>
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
- Hunter, D., Yu, H., Pukish, M. S., Kolbusz, J., & Wilamowski, B. M. (2012). Selection of proper neural network sizes and architectures—A comparative study. *IEEE Transactions on Industrial Informatics*, 8(2), 228–240. <https://doi.org/10.1109/TII.2012.2187914>
- Kashinath, K., Mustafa, M., Albert, A., Wu, J. L., Jiang, C., Esmailzadeh, S., Azzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., ... Prabhat. (2021). Physics-informed machine learning: Case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 20200093. <https://doi.org/10.1098/rsta.2020.0093>
- Lellouche, J. M., Greiner, E., Le Galloudec, O., Garric, G., Regnier, C., Drevillon, M., Benkiran, M., Testut, C. E., Bourdalle-Badie, R., Gasparin, F., Hernandez, O., Levier, B., Drillet, Y., Remy, E., & Le Traon, P. Y. (2018). Recent updates to the Copernicus Marine Service global ocean monitoring and forecasting real-time 1/12° high-resolution system. *Ocean Science*, 14(5), 1093–1126. <https://doi.org/10.5194/os-14-1093-2018>
- Le Traon, P. Y., Reppucci, A., Alvarez Fanjul, E., Aouf, L., Behrens, A., Belmonte, M., Bentamy, A., Bertino, L., Brando, V. E., Bricaud, C., Drevillon, M., Drillet, Y., Dunsthorp, S., Garric, G., Greiner, E., Gutknecht, E., Aaboe, S., & Zacharioudaki, A. (2019). From observation to information and users: The Copernicus Marine Service perspective. *Frontiers in Marine Science*, 6, 234. <https://doi.org/10.3389/fmars.2019.00234>
- Li, H., Xu, F., & Sun, J. (2024). Prediction of seasonal sea surface temperature based on temperature and salinity of subsurface ocean using machine learning. *International Journal of Climatology*, 44(5), 1820–1836. <https://doi.org/10.1002/joc.8384>
- Liu, Y., Tang, Y., Zhang, X., & Wang, L. (2024). Predicting temporal and spatial 4-D ocean temperature using satellite data based on a novel deep learning model. *Ocean Modelling*, 188, 102333. <https://doi.org/10.1016/j.ocemod.2024.102333>
- Oliver, E. C. J., Donat, M. G., Burrows, M. T., Moore, P. J., Smale, D. A., Alexander, L. V., Benthuyssen, J. A., Feng, M., Sen Gupta, A., Hobday, A. J., Holbrook, N. J., Perkins-Kirkpatrick, S. E., Scannell, H. A., Straub, S. C., & Wernberg, T. (2018). Longer and more frequent marine heatwaves over the past century. *Nature Communications*, 9(1), 1324. <https://doi.org/10.1038/s41467-018-03732-9>

- Patil, K., & Deo, M. C. (2018). Basin-scale prediction of sea surface temperature with artificial neural networks. *Journal of Atmospheric and Oceanic Technology*, 35(7), 1441–1455. <https://doi.org/10.1175/JTECH-D-17-0217.1>
- Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39), 9684–9689. <https://doi.org/10.1073/pnas.1810286115>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Sheela, K. G., & Deepa, S. N. (2013). Review on methods to fix number of hidden neurons in neural networks. *Mathematical Problems in Engineering*, 2013, 1–11. <https://doi.org/10.1155/2013/425740>
- Sonniewald, M., Wunsch, C., & Heimbach, P. (2019). Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science*, 6(5), 784–794. <https://doi.org/10.1029/2018EA000519>
- Song, X., Zhang, Y., & Chen, G. (2023). Reconstructing 3D ocean subsurface salinity (OSS) from T–S mapping via a data-driven deep learning model. *Ocean Modelling*, 185, 102232. <https://doi.org/10.1016/j.ocemod.2023.102232>
- Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean squared error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79–82. <https://doi.org/10.3354/cr030079>
- Yang, Y., Dong, J., Sun, X., Lima, E., Mu, Q., & Wang, X. (2018). A CFCC-LSTM model for sea surface temperature prediction. *IEEE Geoscience and Remote Sensing Letters*, 15(2), 207–211. <https://doi.org/10.1109/LGRS.2017.2780843>
- Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159–175. [https://doi.org/10.1016/S0925-2312\(01\)00702-0](https://doi.org/10.1016/S0925-2312(01)00702-0)
- Zhang, Z., Ma, X., & Yang, Y. (2003). Bounds on the number of hidden neurons in three-layer binary neural networks. *Neural Networks*, 16(7), 995–1002. [https://doi.org/10.1016/S0893-6080\(03\)00006-6](https://doi.org/10.1016/S0893-6080(03)00006-6)
- Zhao, Y., Yang, D., He, J., Zhu, K., & Deng, X. (2024). Hierarchical stacked spatiotemporal self-attention network for sea surface temperature forecasting. *Ocean Modelling*, 191, 102427. <https://doi.org/10.1016/j.ocemod.2024.102427>