

Article

A Diffusion Weighted Ensemble Framework for Robust Short-Horizon Global SST Forecasting from Multivariate GODAS Data

Gwangun Yu, GilHan Choi, Moonseung Choi , Sun-hong Min  and Yonggang Kim * 

Department of Software, Kongju National University, Cheonan 31080, Republic of Korea; yugwangun@mail.kongju.ac.kr (G.Y.); choikilhan12@mail.kongju.ac.kr (G.C.); cbhoo3@mail.kongju.ac.kr (M.C.); devminsh02@mail.kongju.ac.kr (S.-h.M.)

* Correspondence: ygkim@kongju.ac.kr

Abstract

Accurate time series forecasting of sea surface temperature (SST) is essential for understanding the ocean climate system and large-scale ocean circulation, yet it remains challenging due to regime-dependent variability and correlated errors across heterogeneous prediction models. This study addresses these challenges by formulating SST ensemble time series forecasting aggregation as a stochastic, sample-adaptive weighting problem. We propose a diffusion-conditioned ensemble framework in which heterogeneous base forecasters generate out-of-sample SST predictions that are combined through a noise-conditioned weighting network. The proposed framework produces convex, sample-specific mixture weights without requiring iterative reverse-time sampling. The approach is evaluated on short-horizon global SST forecasting using the Global Ocean Data Assimilation System (GODAS) reanalysis as a representative multivariate dataset. Under a controlled experimental protocol with fixed input windows and one-step-ahead prediction, the proposed method is compared against individual deep learning forecasters and conventional global pooling strategies, including uniform averaging and validation-optimized convex weighting. The results show that adaptive, diffusion-weighted aggregation yields consistent improvements in error metrics over the best single-model baseline and static pooling rules, with more pronounced gains in several mid- to high-latitude regimes. These findings indicate that stochastic, condition-dependent weighting provides an effective and computationally practical framework for enhancing the robustness of multivariate time series forecasting, with direct applicability to global SST prediction from large-scale geophysical reanalysis data.

Keywords: sea surface temperature (SST) prediction; GODAS reanalysis data; multivariate forecasting; deep learning; diffusion weighted ensemble



Academic Editor: Jonathan Blackledge

MSC: 68T07, 62M10, 60G25, 62G05

Received: 23 January 2026

Revised: 20 February 2026

Accepted: 20 February 2026

Published: 22 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Sea surface temperature (SST) is an important factor to examine in the ocean climate system, as it directly influences air–sea heat exchange, large-scale ocean circulation, and atmosphere–ocean interactions [1]. Fluctuations in SST play a critical role in climate variability, seasonal forecasting, and marine ecosystem dynamics. Consequently, accurate prediction of SST is essential for understanding and monitoring oceanic and climatic processes across multiple temporal scales. Global ocean reanalysis products have been

widely used to analyze and model SST variability [2,3]. Among these, the Global Ocean Data Assimilation System (GODAS), developed by the National Oceanic and Atmospheric Administration (NOAA) and the National Centers for Environmental Prediction (NCEP), provides spatially and temporally consistent near-surface ocean temperature fields by assimilating heterogeneous observational data into numerical ocean circulation models. GODAS reanalysis data offer long-term, global-scale SST records and have become a vital data source for ocean climate studies. However, forecasting future SST from global ocean data remains challenging, as SST variability reflects the combined effects of atmospheric forcing, ocean dynamics, and unresolved sub-grid scale processes [2].

Traditional SST prediction approaches have mainly relied on physics-based numerical models or statistical methods. Physics-based models are physically interpretable but computationally expensive and sensitive to uncertainties in initial conditions and parameterization. Statistical approaches are computationally efficient but often assume linearity or stationarity, which limits their ability to represent nonlinear dynamics and long-range temporal dependencies. These limitations have motivated increasing interest in data-driven methods for SST prediction [4]. Recent advances in deep learning (DL) have significantly improved time series forecasting performance. Transformer-based architectures, such as iTransformer, have further improved multivariate relations and time-series dependency modeling through attention mechanisms, while PatchTST enhances long-horizon forecasting by learning informative representations from segmented temporal patches. In contrast, linear decomposition-based models, such as DLinear, provide efficient modeling of trend and seasonal components with low data complexity. Despite these advances, no single model architecture consistently achieves optimal performance across all temporal scales, regions, and forecasting horizons. SST prediction performance is inherently affected by model uncertainty and data variability. Different models tend to excel under varying temporal regimes, seasonal conditions, and degrees of variability. Relying on a single forecasting model can lead to unstable or suboptimal predictions, particularly under complex oceanic conditions. In addition to model uncertainty, SST predictability is also shaped by horizon-dependent time-series characteristics. At short forecast horizons, predictive skill is often dominated by persistence and strong local autocorrelation, and short-range predictability can degrade abruptly in a seasonally dependent manner [5,6]. As the forecast horizon increases, local persistence becomes less informative, and predictability is increasingly related to seasonal structure and slowly varying, low-frequency components associated with basin-scale modes of SST variability [7]. Moreover, spatiotemporally correlated and nonstationary error behavior has been reported in SST forecast products, implying that model performance can vary across regions even at the same lead time [8]. As a result, ensemble learning has been recognized as an effective strategy to enhance robustness and generalization by integrating the complementary predictive behaviors of multiple models [9,10].

Basic ensemble approaches for SST prediction depend on simple averaging or weighted combinations of deterministic model outputs. While effective to some extent, such approaches do not explicitly model predictive uncertainty or account for the complex, multivariate nature of SST forecasting errors. To address these limitations, probabilistic ensemble frameworks that can represent uncertainty in a principled manner are required [11,12]. Diffusion-weighted generative models provide a robust probabilistic framework for ensemble forecasting by leveraging noise-conditioned representations to estimate optimal ensemble weights without expensive iterative generation. Unlike conventional ensemble methods, the diffusion approach can learn the underlying uncertainty structure of model outputs and generate refined forecasts that accurately represent complex temporal variability [13–15].

In this study, we propose a diffusion-weighted ensemble framework for SST prediction using GODAS reanalysis data. We first generated base forecasts using heterogeneous DL models, including LSTM, iTransformer, PatchTST, and DLinear, each capturing distinct temporal characteristics of SST variability. The proposed diffusion-weighted ensemble model integrates these diverse predictions by modeling correlated error patterns and refining them into a unified forecast. This approach allows the ensemble to adaptively exploit the unique strengths of individual models while reducing prediction variance and improving robustness under region dependent variability within the short-horizon forecasting setting.

- We construct an ensemble composed of diverse model architectures, including LSTM, iTransformer, PatchTST, and DLinear, each capturing complementary temporal characteristics of SST variability, such as long-range dependencies, patchwise temporal patterns, and trend–seasonal components.
- We apply the proposed diffusion-weighted ensemble framework to the multivariate SST GODAS reanalysis data, leveraging consistent global-scale oceanic variables for robust spatiotemporal forecasting.
- We demonstrate that the diffusion-weighted ensemble can effectively refine and combine model predictions by learning their joint uncertainty structure, leading to improved robustness across different forecasting horizons and temporal regimes compared to single models and conventional ensemble approaches.

2. Related Work

2.1. Traditional SST Prediction

Traditional SST prediction has been primarily based on physics-driven numerical ocean models and coupled atmosphere-ocean forecasting systems. Barreto et al. developed and evaluated an operational multigrid ocean forecasting system, demonstrating that physically consistent numerical models can provide reliable SST forecasts over regional domains when properly configured [16]. The main advantage of such systems lies in their physical interpretability and dynamical consistency; however, their performance is sensitive to model resolution and atmospheric forcing errors, leading to persistent regional biases. To mitigate systematic errors in numerical SST forecasts, data-driven bias correction has been explored. Storto et al. proposed a neural network-based surface heat flux correction method embedded in a Nucleus for European Modelling of the Ocean (NEMO) ocean model, showing a significant reduction in SST bias compared to the original configuration [17]. While this hybrid approach improves forecast accuracy without modifying the dynamical core, it remains dependent on the quality and representativeness of the training data and does not explicitly address forecast uncertainty.

In parallel, several studies have concentrated on the refinement and validation of operational forecasting systems. Kong et al. presented validation results for an operational marine forecasting system, highlighting that forecast skill varies substantially across regions and seasons [18]. Although such system-level improvements enhance overall SST performance, the deterministic nature of the forecasts limits their usefulness for probabilistic risk assessment.

The limitations of traditional SST prediction become more evident for extreme events. de Boissésou and Balmaseda investigated the seasonal predictability of marine heatwave occurrence and duration using the European Centre for Medium Range Weather Forecast (ECMWF) seasonal forecast system, showing that while some predictability exists, extreme SST events are often underestimated [19]. Similarly, Koul et al. analyzed the seasonal prediction of marine heatwaves in the Arabian Sea and reported limited reliability for event duration forecasts. These studies indicate that physics-based systems struggle to provide

well-calibrated uncertainty information for extremes, motivating the need for probabilistic post-processing frameworks [20].

2.2. DL-Based SST Prediction

DL approaches have been increasingly applied to SST prediction due to their ability to learn nonlinear spatiotemporal dependencies directly from data. Hao et al. investigated ConvLSTM and ST-ConvLSTM models for SST prediction in the South China Sea, systematically analyzing the impact of traditional methods; nevertheless, performance degraded for longer lead times and under dynamically complex conditions [21]. To enhance temporal dependency modeling, Xu et al. proposed a DL framework for short-term global SST prediction using reanalysis data [22]. The study demonstrated that DL models can outperform conventional statistical baselines for global scale forecasting. Nevertheless, the model focused on deterministic point predictions and did not quantify predictive uncertainty, limiting its applicability to risk-aware forecasting.

Attention mechanisms have been introduced to address the limitations of recurrent architectures. Zrira et al. proposed an attention-based BiLSTM model for SST time series forecasting, demonstrating improved long-range temporal feature extraction and higher accuracy than standard LSTM models [23]. Despite these advancements, the method remains primarily time-series-based and does not fully exploit spatial field structures. To enhance spatial modeling, Shi et al. integrated a deformable attention transformer into a ConvLSTM framework, enabling the model to adaptively capture spatial heterogeneity in SST fields [24]. This approach improves prediction accuracy in regions with significant mesoscale variability but introduces increased model complexity and computational cost.

A growing body of work explores multivariate learning approaches. Fu et al. proposed a hybrid model combining LSTM and Transformer architectures, which integrates sequential learning with global attention, demonstrating improved SST prediction across various coastal regions [25]. Yang et al. further indicated that incorporating various physically related variables into a multifactor DL framework enhances robustness under diverse oceanic conditions [26]. However, most multivariate DL studies still emphasize deterministic accuracy metrics and lack probabilistic evaluation.

2.3. Ensemble Techniques for SST Prediction

Ensemble techniques have been widely adopted to enhance SST prediction robustness by integrating the complementary strengths of diverse forecasting models. Dai et al. proposed a stacked generalization ensemble that combines multiple DL predictors through a meta-learning stage, demonstrating that aggregating heterogeneous models can consistently outperform individual predictors in SST forecasting [27]. The key advantage of this approach lies in its ability to exploit complementary error characteristics across models; however, it requires training and maintaining multiple base learners, as well as an additional meta-learner, resulting in increased computational and operational complexity. From a modeling perspective, ensemble performance strongly depends on the diversity of base learners. Qian et al. showed that combining predictors that capture different physical and dynamical aspects of the ocean, such as integrating SST with sea surface height anomalies, geostrophic velocities, and wind stress, leads to improved forecast skill in dynamically complex regions [28]. While this multivariate ensemble framework enhances deterministic accuracy, predictive uncertainty is still inferred indirectly from ensemble spread rather than being explicitly modeled.

Ensemble learning has also been applied to SST-related extreme event prediction. Bonino et al. employed machine-learning-based ensemble approaches to predict SST variability and marine heatwave occurrence across multiple Mediterranean subregions,

highlighting that ensemble aggregation improves robustness under anomalous conditions [29]. Nevertheless, their framework focused on deterministic event prediction, and probabilistic uncertainty was not formally quantified or calibrated. A different ensemble perspective leverages diversity across climate model simulations rather than observational predictors. Boschetti et al. trained ensemble machine learning models exclusively on climate model outputs and demonstrated that machine learning can function as an interpolator among ensemble members, enhancing SST predictability assessment [30]. Although effective in utilizing model diversity, this approach remains dependent on the fidelity of the underlying climate simulations and does not directly address forecast uncertainty calibration.

3. Dataset

In this study, we use ocean temperature and related oceanographic variables obtained from the GODAS. GODAS integrates in situ and satellite observations with a numerical ocean circulation model through data assimilation, providing dynamically consistent estimates of the ocean state [31]. Moreover, NCEP GODAS data are provided by the NOAA Physical Sciences Laboratory (PSL), Boulder, CO, USA, on their website at <https://psl.noaa.gov>. Note that GODAS does not provide a variable explicitly labeled as SST, as shown in Figure 1, which represents the temperature value in GODAS. The dataset includes ocean potential temperature fields at different vertical levels. Hereafter, SST refers to the near-surface potential temperature from GODAS unless otherwise stated; specifically, it denotes the ocean temperature at approximately 5–10 m depth. This near-surface temperature is used as a proxy for SST in reanalysis-based ocean climate and prediction studies and is suitable for analyzing large-scale and seasonal SST variability.

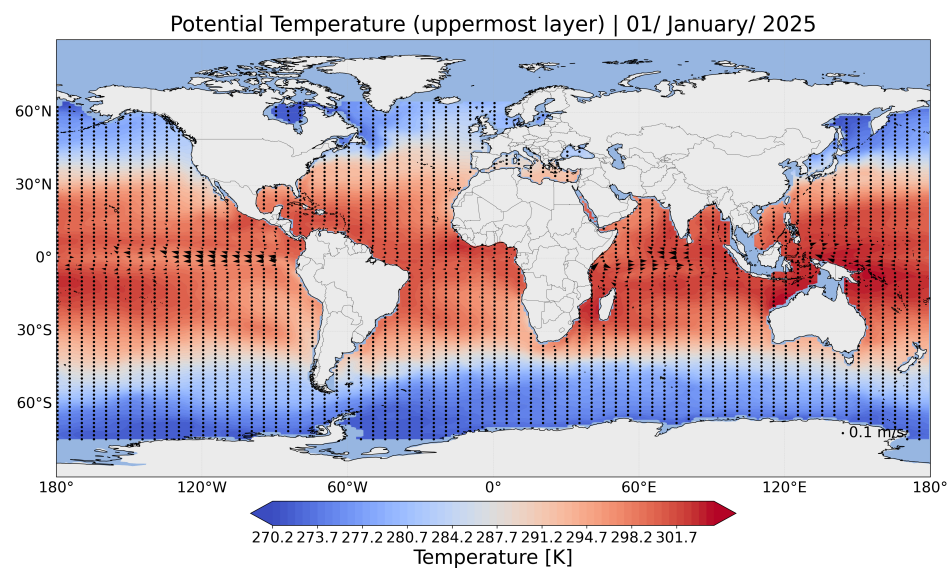


Figure 1. Global map of near-surface potential temperature from GODAS

3.1. Rationale for Selecting GODAS

To contextualize our selection of GODAS, we briefly compare it with two different global ocean reanalysis products, Ocean Reanalysis System 5 (ORAS5) and Global Ocean Physics Reanalysis (GLORYS) 12. ORAS5 provides a global ocean–sea ice ensemble reanalysis and analysis framework, where multiple ensemble members are generated to represent uncertainty in the estimated coupled ocean–ice state [32]. GLORYS12 is a high-resolution global ocean and sea-ice reanalysis designed to represent mesoscale variability with greater detail, particularly during the altimetry era [33]. Both products are, therefore, well suited

for studies that prioritize uncertainty quantification (ORAS5) or eddy-resolving ocean dynamics (GLORYS12). In this study, however, our goal is to benchmark heterogeneous deep forecasters and a diffusion-weighted ensemble under a controlled short-horizon protocol at the global scale, where the primary requirement is a stable, long-record, and globally consistent near-surface thermal field. From this perspective, GODAS provides an appropriate and practical baseline that supports reproducible preprocessing and equitable model comparison.

3.2. Data Structure

The GODAS dataset covers the global ocean domain on a regular latitude–longitude grid. Each grid point represents a fixed geographic location for which various oceanographic variables are available as time series. In this study, land grid points where ocean variables are undefined are removed during preprocessing. The data are organized into a spatio-temporal structure in which each sample corresponds to a specific grid point and time index. For each latitude–longitude location, multivariate time series are constructed by stacking the SST and additional oceanographic variables along the temporal dimension. Table 1 summarizes the oceanographic variables provided by the GODAS dataset. Monthly mean data are used throughout this study.

Table 1. All oceanographic variables are provided by the GODAS reanalysis dataset.

Category	Variable	Description
Temperature	Potential temperature	Ocean potential temperature is provided at multiple vertical levels
Salinity	Salinity	Ocean salinity field provided at multiple vertical levels
Ocean currents	U of current	Zonal (east–west) component of ocean current velocity
	V of current	Meridional (north–south) component of ocean current velocity
Vertical motion	Geometric vertical velocity	Vertical velocity component of ocean flow
Sea level	Sea surface height relative to geoid	Sea surface height referenced to the geoid
Vertical structure	Ocean mixed layer depth below sea surface	Depth of the surface ocean mixed layer
	Ocean isothermal layer depth below sea surface	Depth of the isothermal layer below the sea surface
Surface forcing	Total downward heat flux at surface	Net downward heat flux at the ocean surface
	Zonal momentum flux	Zonal component of surface momentum flux
	Meridional momentum flux	Meridional component of surface momentum flux
	Salt flux	Surface salt flux at the ocean–atmosphere interface

3.3. Data Preprocessing

Several preprocessing steps are applied to the GODAS data prior to model training. Although GODAS provides a wide range of oceanographic variables, not all of them are utilized in this study. Variables are selected to ensure spatial and temporal consistency across the global ocean, which is essential for learning coherent spatio-temporal dependencies in multivariate forecasting models. In particular, variables that exhibit persistent missing values over large contiguous ocean regions or extended temporal intervals are excluded. Such irregular spatial and temporal coverage can introduce structured sparsity, distort local spatio-temporal correlations, and bias the learning of shared representations across grid points and time steps. When incorporated into multivariate time series models, these inconsistencies may lead to unstable training dynamics and degrade generalization by forcing the model to learn from uneven or discontinuous spatio-temporal data. Therefore, we restrict the input set to variables that are spatially continuous and temporally stable over the study period to avoid irregular sampling effects. The final set of input variables used in this study is summarized in Table 2.

Table 2. Oceanographic variables selected from the GODAS dataset and used in this study.

Category	Variable	Description
Temperature	Potential temperature (uppermost level)	Near-surface ocean potential temperature used as a proxy for sea surface temperature
Ocean currents	U of current	Zonal (east-west) component of ocean current velocity
	V of current	Meridional (north-south) component of ocean current velocity
Vertical structure	Ocean mixed layer depth below sea surface	Depth of the surface ocean mixed layer
	Ocean isothermal layer depth below sea surface	Depth of the isothermal layer below the sea surface
Sea level	Sea surface height relative to geoid	Sea surface height referenced to the geoid
Surface forcing	Total downward heat flux at the surface	Net downward heat flux at the ocean surface

To reduce the influence of the mean seasonal cycle and to focus the learning process on interannual and subseasonal variability, anomaly fields are computed for selected variables. For sea surface temperature, surface heat flux, and sea surface height, anomalies $X'(u, v, t)$ are calculated as follows:

$$X'(u, v, t) = X(u, v, t) - X_c^{\text{train}}(u, v). \quad (1)$$

Here, (u, v) index the spatial grid coordinates (latitude and longitude) and t denotes the time index, where $X(u, v, t)$ denotes the raw variable at grid point (u, v) and time t , and $X_c^{\text{train}}(u, v)$ represents the monthly climatological mean for calendar month $c \in \{1, \dots, 12\}$, computed at the same grid point. Other variables, including zonal and meridional ocean currents and upper ocean layer depth variables, are used in their original form.

All input variables are standardized using statistics. Standardization is performed as follows:

$$\tilde{X}(u, v, t) = \frac{X(u, v, t) - \mu_{\text{train}}}{\sigma_{\text{train}}}, \quad (2)$$

where μ_{train} and σ_{train} denote the mean and standard deviation calculated over the training period. This procedure prevents information leakage from the validation and test periods and improves numerical stability during model training.

For variables that contain missing values, such as ocean mixed layer depth and isothermal layer depth, missing values are retained during preprocessing and set to zero after standardization. Binary masks indicating valid observations are maintained to distinguish missing values during model training.

4. Methodology

This section describes the experimental protocol and the proposed ensemble framework for SST prediction. To ensure a fair and controlled comparison across learning paradigms, we fix the input window and forecasting horizon to `seq_len = 12` and `pred_len = 1` for all models. All approaches share the same preprocessing pipeline and the same train/validation/test split described in Section 3. In addition, all normalization statistics (e.g., mean and standard deviation for z-score standardization) are computed using the training split only and then reused for the validation and test splits to prevent any information leakage.

4.1. Base Forecasting Models

We employ a set of heterogeneous forecasters to capture complementary temporal characteristics of SST variability. Let $\mathcal{M}_{\text{DL}} = \{\text{DLinear}, \text{iTransformer}, \text{PatchTST}, \text{LSTM}\}$ denote the deep-learning forecasters. To provide additional reference baselines under the short-horizon setting (`pred_len = 1`), we also include two classical machine-learning regressors: Random Forest (RF) and LinearSVR [34,35]. RF is a strong nonlinear tree-

ensemble baseline that captures feature interactions with minimal modeling assumptions, while LinearSVR offers scalable learning with a linear kernel on high-dimensional inputs. We define the full set of trained models as $\mathcal{M}_{\text{base}} = \mathcal{M}_{\text{DL}} \cup \{\text{LinearSVR}, \text{RF}\}$. For ensemble construction, however, we restrict candidates to the deep-learning set $\mathcal{M}_{\text{ens}} = \mathcal{M}_{\text{DL}}$ to focus subset enumeration and adaptive aggregation on models that explicitly learn temporal structure. This restriction ensures that ensemble aggregation operates on forecasts derived from sequence-aware representations rather than flattened lag features. Such sequence-aware models produce forecasts whose errors reflect regime-dependent temporal dynamics, which is essential for learning meaningful, condition-dependent ensemble weights.

4.2. Input Representation for Classical Baselines

Unlike sequence models, classical regressors require fixed-length vectors. For LinearSVR and RF, we convert each sample's multivariate history into a fixed-dimensional feature vector that matches the temporal context of the deep models. Specifically, the past $\text{seq_len} = 12$ time steps are concatenated into $\mathbf{z} \in \mathbb{R}^{12d}$, where d is the number of input variables. The flattening order across time and variables is kept identical across all experiments to ensure reproducibility. The resulting vectors are standardized using training-set statistics; LinearSVR uses standardized inputs by default, and RF is also trained with the same standardized inputs for consistency.

4.3. Hyperparameter Optimization and Model Selection

For each architecture $m \in \mathcal{M}_{\text{base}}$, we optimize hyperparameters using Optuna [36]. Each Optuna trial is trained on the training split and evaluated on the validation split using a fixed loss metric (e.g., MSE or RMSE). For deep models, we apply early stopping and retain the checkpoint that achieves the lowest validation loss. For LinearSVR and RF, we refit the estimator using the selected hyperparameters on the training split and then obtain validation predictions. After selecting the best configuration for each model, we fix the corresponding forecasters $\{f_m\}_{m \in \mathcal{M}_{\text{base}}}$ and generate out-of-sample predictions on both the validation and test splits. These predictions are used for (i) single-model comparisons and (ii) deep-learning ensemble construction.

4.4. Validation-Based Ensemble Subset Selection

Figure 2 illustrates the overall workflow. Importantly, ensemble fitting and configuration selection are conducted only within the validation split via a meta-train/meta-validation protocol, and the test split is used exactly once for final reporting.

Let $P_{\text{val}} = \{\hat{y}_n^{(m)} \mid m \in \mathcal{M}_{\text{ens}}, n \in \mathcal{D}_{\text{val}}\}$ denote the collection of out-of-sample base predictions on the validation split. Similarly, let $P_{\text{test}} = \{\hat{y}_n^{(m)} \mid m \in \mathcal{M}_{\text{ens}}, n \in \mathcal{D}_{\text{test}}\}$ denote the corresponding predictions on the test split. To exploit complementary error patterns among forecasters, we evaluate all non-empty subset ensembles $S \subseteq \mathcal{M}_{\text{ens}}$. Let $K = |S|$ be the number of models in subset S . Let y_n denote the ground-truth target for sample n . For each sample n , we collect base predictions into a vector

$$\mathbf{x}_{0,n} = [\hat{y}_n^{(m_1)}, \hat{y}_n^{(m_2)}, \dots, \hat{y}_n^{(m_K)}]^\top \in \mathbb{R}^K, \quad \{m_1, \dots, m_K\} = S, \quad (3)$$

where $\hat{y}_n^{(m_i)}$ denotes the prediction of base model $m_i \in S$ for sample n , and y_n denotes the ground-truth target. When $K = 1$, the ensemble reduces to the corresponding single model. When $K \geq 2$, we compare multiple aggregation rules, including uniform averaging, validation-optimized convex weighting, Bayesian model averaging (BMA), quantile regression forests (QRF), and the proposed diffusion-weighted ensemble. To reduce overfitting during ensemble selection and to avoid any leakage from the test split, we adopt a

two-stage protocol on the validation split. Concretely, we first fit any ensemble parameters on a meta-train subset, then select the best subset and aggregation rule using a disjoint meta-validation subset. The selected configuration is subsequently refit using the full validation split, and the test split is used exactly once for final reporting. We denote these two disjoint subsets as $\mathcal{D}_{\text{meta-train}}$ and $\mathcal{D}_{\text{meta-validation}}$ (both drawn from the validation split), where ensemble parameters are fitted on $\mathcal{D}_{\text{meta-train}}$ and configurations are selected based on performance on $\mathcal{D}_{\text{meta-validation}}$.

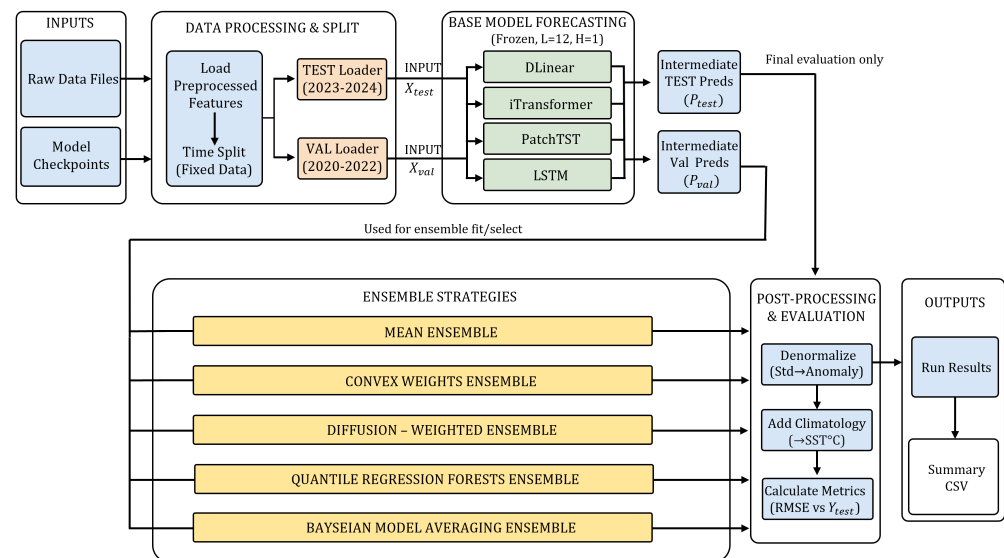


Figure 2. Overview of the deep-learning ensemble pipeline employing noise-conditioned weighting without reverse sampling. Base forecasters (DLinear, iTransformer, PatchTST, LSTM) generate out-of-sample predictions on the validation split (P_{val}). Ensemble parameters and the best subset/aggregation rule are selected using a meta-train/meta-validation protocol within validation only, and the selected ensemble is then applied once to the test predictions (P_{test}) for final evaluation.

4.5. Aggregation Rules for Sample-Adaptive Ensemble Forecasting

Ensemble performance depends not only on which base forecasters are included but also on how their outputs are combined. Because different forecasters can exhibit correlated errors and regime-dependent strengths (e.g., seasonal vs. interannual variability), there is no single universally optimal pooling strategy. Therefore, we compare five aggregation rules with increasing modeling capacity and adaptivity to (i) establish a transparent baseline, (ii) test whether a global, validation-fitted combination improves over naive averaging, (iii) benchmark against evidence-weighted mixtures commonly used in geophysical forecasting, (iv) assess a nonlinear nonparametric combiner that can capture heteroscedastic and asymmetric regimes, and (v) evaluate whether sample-adaptive weighting yields additional gains. Concretely, we consider: (1) a uniform mean as a strong, assumption-light reference; (2) validation-optimized convex weights (linear pool) as a classical forecast-combination method that learns a global mixture while mitigating overfitting via nonnegativity and sum-to-one constraints; (3) Bayesian model averaging (BMA) as an evidence-weighted global mixture baseline; (4) quantile regression forests (QRF) as a nonlinear, distribution-aware combiner; and (5) the proposed diffusion-weighted rule as a sample-adaptive mechanism that can vary mixture weights across different oceanic conditions. This controlled comparison allows us to attribute improvements to the aggregation strategy itself rather than to changes in base models, and it clarifies the value of sample-adaptive weighting beyond both conventional global pooling and recent geophysical ensemble baselines.

Given a subset S of K base forecasters, for each sample n we define the stacked base prediction vector $\mathbf{x}_{0,n} = [\hat{y}_n^{(1)}, \dots, \hat{y}_n^{(K)}]^\top \in \mathbb{R}^K$ and denote by y_n the corresponding ground-

truth target (each sample can be interpreted as a specific spatiotemporal point, e.g., a grid cell at a given time). The uniform mean ensemble is defined as

$$\hat{y}_{\text{mean}} = \frac{1}{K} \sum_{i=1}^K \hat{y}^{(i)}. \quad (4)$$

As a stronger global pooling baseline, we also learn nonnegative convex weights on the meta-train split:

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbb{R}^K} \sum_{n \in \mathcal{D}_{\text{meta-train}}} (\mathbf{w}^\top \mathbf{x}_{0,n} - y_n)^2 \quad \text{s.t. } \mathbf{w} \geq \mathbf{0}, \mathbf{1}^\top \mathbf{w} = 1, \quad (5)$$

and predict by $\hat{y}_{\text{cw},n} = (\mathbf{w}^*)^\top \mathbf{x}_{0,n}$. While convex weighting provides a simple yet competitive global combination rule, it cannot adapt to sample-specific conditions. To strengthen the positioning of our diffusion-weighted aggregation against recent ensemble practices in geophysical forecasting, we additionally include two widely used alternatives: Bayesian model averaging (BMA) and quantile regression forests (QRF). BMA provides an evidence-weighted global mixture by assigning higher weights to base forecasters that better explain the meta-train targets. Concretely, let $\text{mse}_i = \frac{1}{|\mathcal{D}_{\text{meta-train}}|} \sum_{n \in \mathcal{D}_{\text{meta-train}}} (\hat{y}_n^{(i)} - y_n)^2$ denote the meta-train mean squared error of model i . Under a Gaussian-error approximation, we compute AIC-like weights

$$\tilde{w}_i \propto \exp\left(-\frac{1}{2}(|\mathcal{D}_{\text{meta-train}}| \log(\text{mse}_i) + 2)\right), \quad w_i = \frac{\tilde{w}_i}{\sum_{j=1}^K \tilde{w}_j}, \quad (6)$$

and form the BMA prediction as $\hat{y}_{\text{bma},n} = \sum_{i=1}^K w_i \hat{y}_n^{(i)}$. This yields a principled evidence-weighted pooling rule that remains global (sample-invariant) yet goes beyond uniform or convex weighting.

In contrast, QRF serves as a nonlinear, nonparametric combiner that can better accommodate regime-dependent and potentially asymmetric error behavior. Specifically, we train a random-forest regressor g_ϕ on the meta-train split to map the base prediction vector $\mathbf{x}_{0,n}$ to the target y_n . At inference, we collect per-tree predictions $\{g_{\phi,t}(\mathbf{x}_{0,n})\}_{t=1}^T$ and use their empirical quantile as a robust point forecast:

$$\hat{y}_{\text{qrf}} = \text{Quantile}_q\left(\{g_{\phi,t}(\mathbf{x}_{0,n})\}_{b=1}^B\right). \quad (7)$$

Here, b indexes the trees in the random forest, and B is the total number of trees. Unless stated otherwise, we use $q = 0.5$ (the median). Together, BMA and QRF provide complementary and widely recognized baselines—evidence-weighted global pooling and nonlinear, distribution-aware combining, respectively—which help contextualize the gains from sample-adaptive diffusion-based weighting. Building on these global and nonparametric baselines, we propose a diffusion-weighted ensemble, inspired by diffusion probabilistic models [37], that uses only the forward diffusion process to construct noise-conditioned inputs while omitting iterative reverse-time sampling. Specifically, we perturb the base prediction vector via a forward noising process:

$$\mathbf{x}_\tau = \sqrt{\bar{\alpha}_\tau} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_\tau} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K), \quad (8)$$

where $\tau \sim \text{Unif}\{0, \dots, T-1\}$, T is the number of diffusion steps, $\beta_\tau \in (0, 1)$ is a predefined noise schedule, $\alpha_\tau = 1 - \beta_\tau$, and for $\tau \geq 1$ we define $\bar{\alpha}_\tau = \prod_{s=1}^\tau \alpha_s$ (with $\bar{\alpha}_0 = 1$ by convention). In practice, β_τ is set to a monotone increasing schedule over $\tau = 0, \dots, T-1$. To assess sensitivity to this design choice, we consider five schedule families (linear, cosine,

quadratic, sigmoid, and exponential) while keeping $(\beta_{\text{start}}, \beta_{\text{end}}, T)$ fixed. All schedules are mapped to the same range $[\beta_{\text{start}}, \beta_{\text{end}}]$ for a fair comparison. Given the noised input $\mathbf{x}_\tau$, we learn a noise-conditioned weighting network $f_\theta(\cdot)$ that takes the concatenated input $[\mathbf{x}_\tau; \mathbf{e}(\tau)]$ and outputs K logits, which are converted into mixture weights via

$$\mathbf{w}_\tau = \text{softmax}(f_\theta([\mathbf{x}_\tau; \mathbf{e}(\tau)])) \in \mathbb{R}^K, \quad \sum_{i=1}^K w_{\tau,i} = 1, \quad (9)$$

where $\mathbf{e}(\tau) \in \mathbb{R}^{d_e}$ denotes a sinusoidal embedding of the diffusion step τ . The final forecast is computed as a convex combination of the clean base predictions:

$$\hat{y}_{\text{diff}} = \mathbf{w}_\tau^\top \mathbf{x}_0, \quad (10)$$

i.e., the weights are conditioned on the noised input $\mathbf{x}_\tau$ but applied to the clean predictions $\mathbf{x}_0$. We train θ on the meta-train split while freezing all base models by minimizing the expected mean squared error under randomly sampled (τ, ϵ) :

$$\mathcal{L}(\theta) = \mathbb{E}_{(\mathbf{x}_0, y) \in \mathcal{D}_{\text{meta-train}}, \tau, \epsilon} \left[(\mathbf{w}_\tau^\top \mathbf{x}_0 - y)^2 \right], \quad (11)$$

where the expectation is approximated by mini-batch sampling during training. At inference, we perform Monte-Carlo averaging by drawing M pairs (τ, ϵ) , averaging the resulting weights $\bar{\mathbf{w}} = \frac{1}{M} \sum_{j=1}^M \mathbf{w}^{(j)}$, and applying the averaged weights to $\mathbf{x}_0$ to obtain $\hat{y}_{\text{diff}} = \bar{\mathbf{w}}^\top \mathbf{x}_0$. Unless stated otherwise, we use an exponential noise schedule

$$\beta_\tau = \beta_{\text{start}} \left(\frac{\beta_{\text{end}}}{\beta_{\text{start}}} \right)^{\frac{\tau}{T-1}}, \quad \tau = 0, \dots, T-1, \quad (12)$$

with $\beta_{\text{start}} = 10^{-4}$, $\beta_{\text{end}} = 0.02$, and $T = 50$. To address potential sensitivity to the choice of β_τ , we additionally evaluate four alternative schedules (linear, cosine, quadratic, and sigmoid) under the same $(\beta_{\text{start}}, \beta_{\text{end}}, T)$ setting. The overall architecture of the proposed diffusion-weighted ensemble is illustrated in Figure 3. Across all 11 ensemble subsets ($K \in \{2, 3, 4\}$), the differences are small: the exponential schedule yields the lowest average test RMSE, whereas the cosine schedule attains the best single-subset RMSE (Tables A5–A7). For clarity and reproducibility, we adopt the exponential schedule as the default in the main experiments and report schedule sensitivity in the Appendix A. The weighting network f_θ is implemented as a four-hidden-layer MLP with sinusoidal timestep embeddings (embedding dimension $d_e = 32$, hidden dimension 64) and is trained using AdamW (learning rate 10^{-3} , weight decay 10^{-6}) for 20 epochs with gradient clipping at 1.0; at inference, we use $M = 8$ Monte-Carlo draws. We select the MLP depth as 4 hidden layers based on an ablation study over depths $\{2, 4, 8, 16\}$ conducted across all 11 ensemble subsets ($K \in \{2, 3, 4\}$). Appendix A3 summarizes the aggregate behavior of each depth by reporting the mean $\pm$ std and the minimum RMSE over the 11 subsets, which provides a robustness-oriented rationale for the default choice. Appendix A4 complements this summary by listing the full subset-by-depth RMSE matrix (val/test) to document all runs and to show that depth 4 most frequently attains the lowest test RMSE across subsets. Accordingly, depth 4 achieves the lowest average RMSE and yields the overall best RMSE among all tested depths, while overly deep networks (e.g., depth 16) slightly degrade the mean performance.

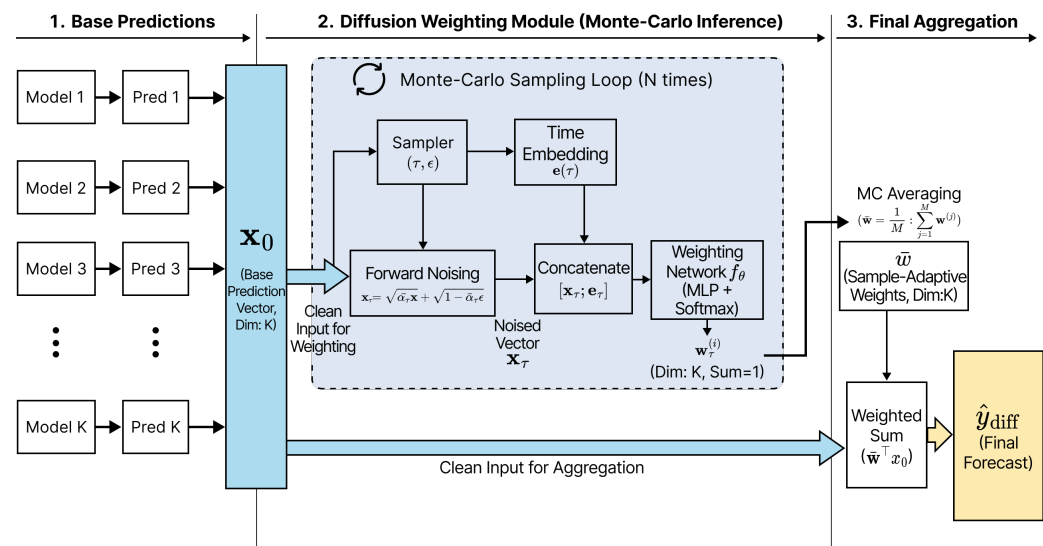


Figure 3. Architecture of the proposed diffusion-weighted ensemble. Unlike standard diffusion models, the network directly predicts mixing weights conditioned on noise levels, bypassing iterative reverse sampling. Given the clean base prediction vector $\mathbf{x}_0$, we apply forward noising to obtain $\mathbf{x}_\tau$ and use a noise-conditioned weighting network $f_\theta([\mathbf{x}_\tau; \mathbf{e}(\tau)])$ to produce convex weights via softmax. At inference, Monte-Carlo averaging over multiple (τ, ϵ) draws yields the final weights $\bar{\mathbf{w}}$, which are applied to $\mathbf{x}_0$ to compute the final forecast $\hat{y}_{\text{diff}} = \bar{\mathbf{w}}^\top \mathbf{x}_0$.

4.6. Relation to Standard Diffusion Models

Standard diffusion generative models learn a reverse-time process to synthesize samples [37]. In contrast, our approach uses only forward noising and noise-conditioning to compute ensemble weights, avoiding iterative reverse sampling. Therefore, advances that accelerate reverse sampling (e.g., EDM, DPM-Solver, Consistency Models, and flow-based reformulations) are orthogonal to our objective [38–41]. Based on the methodology described above, we now evaluate the proposed ensemble strategies on the held-out test set. We compare the performance of individual base forecasters and different ensemble aggregation rules, with particular emphasis on the proposed diffusion-weighted ensemble. All reported results are obtained using the fixed experimental protocol and evaluation metrics described in this section.

5. Results

We evaluate forecasting performance using root-mean-square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). All metrics are reported on the held-out test split unless stated otherwise. For interpretability, we report RMSE both in standardized anomaly units (after z-score normalization) and in physical units ($^{\circ}\text{C}$) by converting predictions back using the training-set statistics. Although the relative improvements in RMSE and MAE may appear numerically modest, they are meaningful in our short-horizon setting. For one-step-ahead SST anomaly forecasting, persistence and strong local autocorrelation already constitute a strong baseline, leaving limited headroom for further error reduction. In such a regime, consistent improvements in aggregate metrics suggest that the proposed method reduces residual error components beyond persistence, potentially including regime-dependent mismatches and correlated model errors. Moreover, because evaluation is conducted over a global domain with a large number of spatial samples, small average error reductions are less likely to arise from isolated cases and instead indicate broadly consistent changes in the overall error distribution. These effects can correspond to more substantial localized gains in dynamically complex regions, which

is consistent with our latitude-band analysis showing more pronounced improvements in several mid- to high-latitude bands.

5.1. Single-Model Performance

Table 3 compares the performance of individual forecasting models, including four DL forecasters (DLinear, iTransformer, PatchTST, LSTM) and two classical machine-learning baselines (LinearSVR and RF). Overall, the best single model achieves the lowest RMSE and MAE and the highest R^2 , indicating stronger agreement with the observed SST anomaly variability. We observe that the classical baselines remain competitive under the short-horizon setting ($\text{pred_len} = 1$), and that flattened lag features can still provide strong predictive signals for near-term SST anomalies.

Table 3. Test performance of individual forecasting models. RMSE is reported in standardized units (std) and in physical units ($^{\circ}\text{C}$). Best values (lowest error/highest R^2) are highlighted in bold.

Model	RMSE (std)	RMSE ($^{\circ}\text{C}$)	MAE ($^{\circ}\text{C}$)	R^2
DLinear	0.6478	0.3755	0.2751	0.6622
iTransformer	0.6472	0.3751	0.2731	0.6628
PatchTST	0.6449	0.3738	0.2733	0.6652
LSTM	0.6231	0.3612	0.2627	0.6874
LinearSVR	0.6420	0.3720	0.2683	0.6648
RF	0.6400	0.3710	0.2620	0.6672

Among individual forecasters, LSTM achieves the strongest overall accuracy, yielding the lowest test RMSE (0.3612°C) and MAE (0.2627°C) and the highest R^2 (0.6874) in Table 3. This suggests that, under the short-horizon setting ($\text{pred_len} = 1$), recurrent modeling of local temporal dynamics remains effective for SST anomaly prediction. Notably, the classical baselines (LinearSVR and RF) remain competitive despite using flattened lag features, indicating that near-term SST anomalies contain strong linear and low-order nonlinear signals that can be exploited without explicit sequence modeling.

5.2. Ensemble Performance by Aggregation Rule

We compare three aggregation rules: uniform mean, validation-optimized convex weights, and the proposed diffusion-weighted ensemble. As shown in Table 4, uniform averaging provides a competitive but not consistently superior baseline, whereas validation-optimized convex weighting and the proposed diffusion-weighted ensemble improve upon the best single model.

Table 4. Test performance of ensemble methods. For each aggregation rule, we report the test performance of the best-performing subset selected on the meta-validation split. Best values are highlighted in bold.

Method	RMSE (std)	RMSE ($^{\circ}\text{C}$)	MAE ($^{\circ}\text{C}$)	R^2
Single (LSTM)	0.6231	0.3612	0.2627	0.6874
Ensemble—Mean	0.6273	0.3636	0.2656	0.6833
Ensemble—Convex Weight	0.6209	0.3599	0.2620	0.6896
Ensemble—BMA	0.6231	0.3612	0.2627	0.6874
Ensemble—QRF	0.6321	0.3664	0.2704	0.6770
Ensemble—Diffusion-weighted	0.6188	0.3586	0.2612	0.6918

Relative to the best single model (LSTM), the diffusion-weighted ensemble reduces the test RMSE from 0.3612°C to 0.3586°C , corresponding to a relative improvement of approximately 0.70%. A similar improvement is observed in MAE (from 0.2627°C to 0.2612°C ; $\sim 0.57\%$), while R^2 increases from 0.6874 to 0.6918.

Among global (sample-invariant) pooling rules, uniform mean provides a stable reference but cannot account for correlated forecast errors, and validation-optimized convex weighting improves upon averaging by learning a constrained linear pool. To better contextualize diffusion-based weighting relative to ensemble practices commonly used in geophysical forecasting, we additionally evaluate Bayesian model averaging (BMA) and quantile regression forests (QRF). In our setting, BMA performs comparably to the best single forecaster, suggesting that evidence-weighted yet still global mixing offers limited gains when base-model errors are strongly correlated or when the evidence weights concentrate on a single dominant forecaster. Meanwhile, QRF underperforms, indicating that a generic nonlinear combiner does not automatically translate to improved accuracy under our controlled protocol and fixed feature set.

Overall, the proposed diffusion-weighted aggregation achieves the best performance across all metrics by producing sample-adaptive mixture weights, enabling the ensemble to vary its weighting across different oceanic conditions while consistently outperforming both conventional global pooling and recent geophysical ensemble baselines.

5.3. Latitude-Band Error Analysis

To examine where the proposed ensemble improves the most, we compute test RMSE within 10° latitude bands and report band-wise errors for LSTM and the diffusion-weighted ensemble. This analysis highlights regional differences in predictability and helps interpret whether improvements are concentrated in specific latitude regimes.

Figure 4 and Table 5 demonstrate that the performance of the diffusion-weighted ensemble, configured with an exponential noise schedule, is latitude-dependent. Notably, clearer gains are observed in mid-to-high latitude bands (e.g., 35° – 65°), with the largest improvement at 40° (Imp $\approx 2.61\%$), whereas some low-to-mid latitude bands exhibit marginal or slightly negative changes (e.g., -10° shows $\approx -1.03\%$). A plausible explanation is that mid-to-high latitudes are characterized by stronger variance and regime shifts, which amplify inter-model disagreement and nonstationary error correlations; under such conditions, diffusion-conditioned reweighting can better exploit the complementary strengths of heterogeneous forecasters. In contrast, tropical and low-latitude regimes are often dominated by smoother variability, leading to higher concordance among base predictors and limiting the potential gains from adaptive reweighting. Critically, these results should be interpreted together with absolute error magnitudes: low-latitude regions typically exhibit smaller baseline RMSEs than higher latitudes; so, even small absolute differences can appear as noticeable percentage changes due to the smaller denominators. Thus, slightly negative values in a few low-latitude bands do not necessarily indicate a substantive degradation, but rather reflect the sensitivity of relative metrics in low-variance regimes. Overall, the proposed method is most impactful in complex, high-variance regions, while remaining competitive in smoother regimes.

Table 5. Band-wise test RMSE by latitude. Imp(%) denotes the relative difference in the Diffusion-weighted ensemble over LSTM: $(\text{RMSE}_{\text{LSTM}} - \text{RMSE}_{\text{Diff}}) / \text{RMSE}_{\text{LSTM}} \times 100$. we use half-open intervals [a,b)

Lat. Center	Range	N	RMSE (LSTM)	RMSE (Diff)	Imp (%)
−70	[−75, −65)	3482	0.2362	0.2330	1.3396
−60	[−65, −55)	6252	0.2688	0.2664	0.8983
−50	[−55, −45)	6325	0.2798	0.2788	0.3412
−40	[−45, −35)	5814	0.3711	0.3718	−0.1836
−30	[−35, −25)	5113	0.3599	0.3618	−0.5415
−20	[−25, −15)	4900	0.3173	0.3192	−0.5941
−10	[−15, −5)	4654	0.3322	0.3356	−1.0308

Table 5. Cont.

Lat. Center	Range	N	RMSE (LSTM)	RMSE (Diff)	Imp (%)
0	[−5,5)	4739	0.4028	0.4003	0.6308
10	[5,15)	4752	0.3475	0.3468	0.2088
20	[15,25)	4136	0.3291	0.3292	−0.0230
30	[25,35)	3661	0.4301	0.4256	1.0400
40	[35,45)	3143	0.5953	0.5797	2.6149
50	[45,55)	2594	0.4352	0.4316	0.8316
60	[55,65)	1861	0.4105	0.4025	1.9405

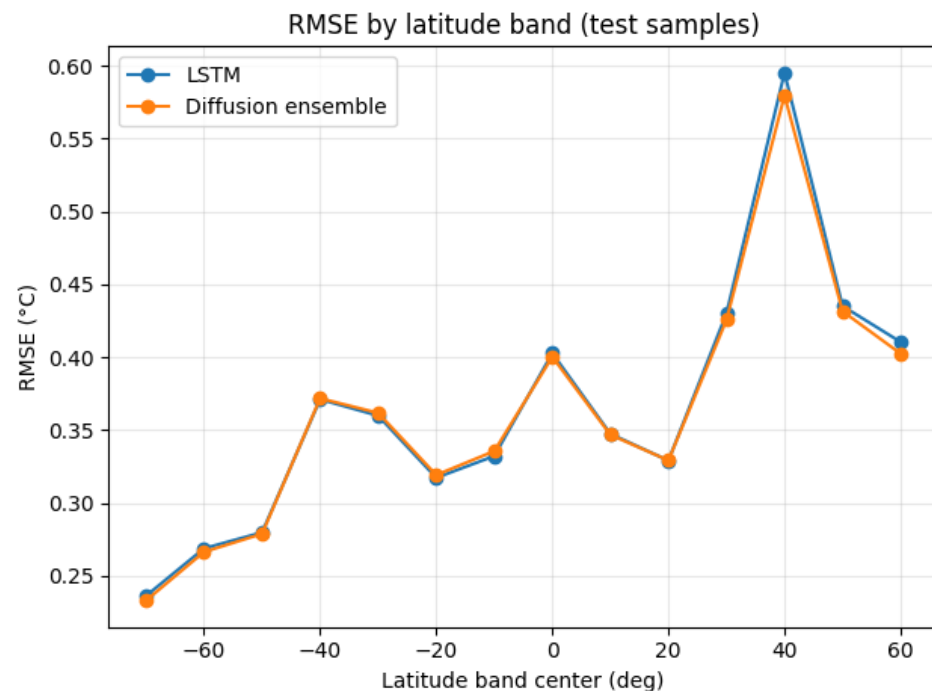


Figure 4. Test RMSE by latitude band (10° bins). The Diffusion-weighted ensemble is compared against the best single model (LSTM) to reveal latitude-dependent performance differences.

5.4. Best Subset Selection: Validation vs. Test

To study whether all four models are necessary, we evaluate all non-empty subsets of $\mathcal{M}_{\text{ens}}$ under each aggregation rule using the meta-train/meta-validation protocol described in Section 4.4. We consider three aggregation rules: uniform mean, validation-optimized convex weights, and the proposed diffusion-weighted ensemble. We report (i) the configuration selected by the lowest validation RMSE (the proper selection criterion) and (ii) the configuration that achieves the lowest test RMSE (reported only as a post hoc reference). Table 6 summarizes both cases.

Table 6. Best ensemble configurations. The left row is selected by the lowest validation RMSE (proper selection). The right row shows the post hoc best test RMSE (for reference) and is not a valid selection method. Here, D = DLinear, I = iTransformer, P = PatchTST, and L = LSTM.

Selection	Subset + Aggregation	Val RMSE (°C)	Test RMSE (°C)	Test MAE (°C)	Test R^2
Val-best	I + P + L; Diffusion	0.3548	0.3597	0.2621	0.6899
Test-best (post hoc)	D + L; Diffusion	0.3559	0.3592	0.2616	0.6909

The Val-best row in Table 6 represents the only configuration that is valid under our experimental protocol because it is selected using validation RMSE without accessing

any test labels. In contrast, the Test-best (post hoc) row is reported solely as a reference point; this test-oracle choice is not a valid selection method and must not be interpreted as a deployable configuration because it implicitly uses test outcomes for model selection. Importantly, the validation-selected diffusion configuration remains highly competitive on the test set (0.3597°C), with only a negligible gap compared to the post hoc best-test configuration (0.3592°C). This small difference suggests that validation-driven subset selection provides a reliable proxy for generalization and that the proposed diffusion-weighted ensemble remains effective even when the ensemble is formed from a reduced subset of base forecasters.

6. Discussion

In this study, we investigated short-horizon sea surface temperature (SST) forecasting from GODAS reanalysis using a diffusion-weighted ensemble framework. We constructed a multivariate global dataset from GODAS and defined SST as the near-surface (uppermost-level) potential temperature, with anomalies computed using training-period monthly climatology to reduce the dominant seasonal cycle. Under a controlled protocol with a fixed input window ($\text{seq_len} = 12$) and one-step forecasting horizon ($\text{pred_len} = 1$), we trained heterogeneous base forecasters (DLinear, iTransformer, PatchTST, and LSTM) and generated out-of-sample predictions for validation and test splits to enable leak-free ensemble construction.

To combine base forecasts, we compared three aggregation rules with increasing modeling capacity: uniform mean, validation-optimized convex weighting (linear pool), and the proposed diffusion-weighted ensemble that enables one-shot weight estimation without reverse sampling. Ensemble configuration selection was performed strictly within validation using a meta-train/meta-validation split, and the test set was used only once for final reporting. Overall, the results demonstrate that combining heterogeneous deep-learning forecasters improves robustness compared to the best single model, and that the diffusion-weighted ensemble consistently outperforms simpler global pooling strategies.

The latitude band analysis suggests that the benefits of sample-adaptive weighting are not uniform across regions; improvements are more pronounced in several mid-to-high latitude bands, while some low-latitude bands show marginal or negative changes. Finally, subset enumeration indicates that competitive performance can be achieved with reduced ensembles (e.g., three-model subsets), implying that strong performance does not necessarily require all candidate models when selection is conducted properly on validation.

Despite these promising results, several limitations remain. First, our experiments focus on a single short-horizon setting ($\text{pred_len} = 1$) and monthly GODAS data; extending the framework to longer lead times and higher temporal resolution (e.g., weekly or daily products) is required to fully characterize the advantages of diffusion-based adaptive aggregation under increased uncertainty. Second, GODAS is a reanalysis product and therefore reflects model and assimilation biases; evaluating the proposed ensemble on independent observational datasets and alternative reanalyses would strengthen the generality of the conclusions. Third, while we analyzed latitude-dependent errors, additional diagnostics (e.g., seasonality, regional ocean basins, and extreme-event regimes such as marine heatwaves) would provide deeper insight into when and why sample-adaptive weighting is most beneficial.

Our current experiments are limited to one-step forecasting. Nonetheless, the proposed diffusion-weighted aggregation is, in principle, compatible with multi-step forecasting settings. (i) In a recursive rollout, one-step base forecasts are iteratively propagated; the ensemble can recompute diffusion-conditioned weights at each step using the available

base prediction vector for that step, while acknowledging the well-known risk of error accumulation and compounding errors. (ii) In a direct multi-step setting, where an H -step output is predicted in a single forward pass, the weighting network could be conditioned on the lead time (e.g., via a horizon embedding) or could output an $H \times K$ set of weights, allowing the mixture to vary across lead times if error correlations differ. A systematic assessment of these extensions (including training base forecasters for multi-horizon targets and re-tuning diffusion/noising-related hyperparameters) is beyond the scope of this study and is left for future work.

Future work will extend the proposed framework in three directions. (i) We will evaluate multi-step forecasting horizons and probabilistic calibration, including uncertainty-aware metrics, to quantify whether diffusion-weighted aggregation improves reliability as lead time increases. Moreover, the increased sample size at daily or sub-seasonal resolution may support training higher-capacity forecasters; however, any performance gain must be verified empirically under the same leak-free selection protocol. (ii) We will incorporate additional physically relevant predictors and alternative anomaly definitions to assess the sensitivity of model performance to preprocessing choices. (iii) We will explore lightweight and computationally efficient variants of the weighting network to enable deployment-oriented inference while preserving the benefits of sample-adaptive mixture weights.

7. Conclusions

Collectively, this work demonstrates that diffusion-weighted, sample-adaptive aggregation provides a practical and effective pathway to improving the robustness of SST forecasting from multivariate global reanalysis data.

Author Contributions: Conceptualization, Y.K.; methodology, G.Y.; investigation, G.Y.; formal analysis, G.Y. and G.C.; validation, M.C. and S.-h.M.; writing—original draft preparation, G.Y. and G.C.; writing—review and editing, Y.K.; and supervision, Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Regional Innovation System & Education(RISE) program through the Chungnam RISE center, funded by the Ministry of Education(MOE) and the Chungcheongnam-do, Republic of Korea (2025-RISE-12-003).

Data Availability Statement: The data used in this study are derived from public domain resources. The Global Ocean Data Assimilation System (GODAS) reanalysis data are publicly available from the NOAA National Centers for Environmental Information (NCEI) at <https://psl.noaa.gov>.

Acknowledgments: This work was supported by the Regional Innovation System & Education(RISE) program through the Chungnam RISE center, funded by the Ministry of Education(MOE) and the Chungcheongnam-do, Republic of Korea (2025-RISE-12-003).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Complete Results

This appendix provides full validation/test results to ensure transparency and reproducibility. All ensemble configurations are selected using validation only (meta-train/meta-validation protocol; Section 4.4), and the test split is used once for final reporting. We additionally report a test-oracle configuration solely as a post hoc reference to illustrate the upper bound.

Appendix A.1. Single-Model Baselines

Table 3 compares all single models on validation and test splits using RMSE, MAE, and R^2 . For the deep-learning models, we report both standardized metrics (z-score space) and

physical-unit errors in °C (after inverse standardization). Classical baselines (LinearSVR, RF) are evaluated in standardized space. Note that RMSE/MAE in °C are obtained by inverse-transforming standardized predictions using training-set statistics, whereas z-space metrics directly reflect errors in normalized anomaly units.

Table A1. Single-model performance on the validation and test splits.

Model	Val RMSE (z)	Test RMSE (z)	Test MAE (z)	Test R^2	Test RMSE (°C)	Test MAE (°C)
DLinear	0.6331	0.6478	0.4746	0.6622	0.375	0.275
iTransformer	0.6303	0.6472	0.4757	0.6628	0.375	0.276
PatchTST	0.6254	0.6449	0.4715	0.6652	0.374	0.273
LSTM	0.6156	0.6231	0.4533	0.6874	0.361	0.263
LinearSVR	0.6248	0.6421	0.4684	0.6648	0.372	0.271
RF	0.6188	0.6398	0.4616	0.6672	0.370	0.267

For deep-learning models, we additionally report errors in physical units (SST in °C) after inverse standardization. For classical baselines (LinearSVR, RF), only standardized metrics are available.

For brevity, we use the following model abbreviations: D (DLinear), I (iTransformer), P (PatchTST), and L (LSTM). For ensemble rules, single denotes the base model, mean the uniform mean ensemble, cw the validation-optimized convex weighting, and diff the diffusion-weighted ensemble.

Appendix A.2. Ablation on the Depth of the Diffusion Weighting Network

To justify the architecture choice for the diffusion weighting network f_θ , we ablate the number of hidden layers (MLP depth) in $\{2, 4, 8, 16\}$ while keeping all other hyperparameters and the experimental protocol fixed. For each depth, we report the configuration selected by the lowest meta-validation RMSE and its corresponding test performance. As shown in Table A2, depth = 4 achieves the best meta-validation RMSE and the strongest (or tied-best) test accuracy; thus, we use depth = 4 as the default setting throughout this paper. Table A3 further reports aggregate statistics (mean $\pm$ std and minimum RMSE over the 11 subsets) to quantify how consistently each depth performs on average, whereas Table A4 provides the complete subset-by-depth RMSE matrix (val/test) as a full experimental record, enabling verification of which depth wins for each subset and demonstrating that the ablation was exhaustively evaluated rather than selectively reported.

Table A2. Effect of MLP depth (number of hidden layers) in the diffusion weighting network f_θ . For each depth, we report the best configuration selected on meta-validation RMSE and its test performance. Best values are highlighted in bold.

MLP Depth	Val-Best Subset	Val RMSE (°C)	Test RMSE (°C)	Test MAE (°C)	Test R^2
2	I + P + L	0.3548	0.3597	0.2621	0.6899
4	I + P + L	0.3546	0.3590	0.2613	0.6911
8	D + I + P + L	0.3547	0.3592	0.2614	0.6909
16	I + P + L	0.3548	0.3599	0.2621	0.6896

Table A3. Aggregate performance across all 11 ensemble subsets ($K \in \{2, 3, 4\}$). We report mean $\pm$ std and minimum RMSE (°C) over the 11 subsets for each MLP depth. Lower is better.

MLP Depth	Meta-Val RMSE		Test RMSE	
	Mean $\pm$ Std	Min	Mean $\pm$ Std	Min
2	0.3571 $\pm$ 0.0029	0.3548	0.3634 $\pm$ 0.0054	0.3592
4	0.3571 $\pm$ 0.0029	0.3546	0.3633 $\pm$ 0.0055	0.3590
8	0.3571 $\pm$ 0.0028	0.3547	0.3635 $\pm$ 0.0054	0.3592
16	0.3572 $\pm$ 0.0029	0.3548	0.3637 $\pm$ 0.0053	0.3596

Table A4. Full results of the diffusion-weighting ablation across all 11 ensemble subsets ($K \in \{2, 3, 4\}$) and all MLP depths. Each cell reports val/test RMSE ($^{\circ}\text{C}$), and the best test RMSE per row is bolded.

K	Ensemble Subset	Depth = 2	Depth = 4	Depth = 8	Depth = 16
2	D + L	0.3559/0.3592	0.3559/0.3593	0.3559/0.3593	0.3559/0.3596
2	D + P	0.3609/0.3717	0.3609/0.3717	0.3607/0.3714	0.3611/0.3713
2	D + I	0.3625/0.3703	0.3624/0.3703	0.3625/0.3706	0.3626/0.3713
2	P + L	0.3551/0.3599	0.3550/0.3598	0.3551/0.3601	0.3551/0.3603
2	I + L	0.3552/0.3596	0.3551/0.3595	0.3552/0.3595	0.3554/0.3600
2	I + P	0.3595/0.3694	0.3595/0.3695	0.3594/0.3697	0.3597/0.3696
3	D + P + L	0.3551/0.3598	0.3550/0.3593	0.3551/0.3599	0.3551/0.3600
3	D + I + L	0.3551/0.3592	0.3551/0.3592	0.3552/0.3594	0.3553/0.3597
3	D + I + P	0.3595/0.3691	0.3596/0.3694	0.3593/0.3694	0.3597/0.3693
3	I + P + L	0.3548/0.3597	0.3546/0.3590	0.3548/0.3598	0.3548/0.3599
4	D + I + P + L	0.3548/0.3595	0.3546/0.3591	0.3547/0.3592	0.3548/0.3600

Appendix A.3. Sensitivity to the Noise Schedule

In practice, β_{τ} is set to a monotone increasing schedule over $\tau = 0, \dots, T - 1$. To assess sensitivity to this design choice, we consider five schedule families (linear, cosine, quadratic, sigmoid, and exponential) while keeping $(\beta_{\text{start}}, \beta_{\text{end}}, T)$ fixed. Following our implementation, all schedules are mapped to the same range $[\beta_{\text{start}}, \beta_{\text{end}}]$ for a fair comparison. Table A5 summarizes the aggregate performance across all 11 ensemble subsets ($K \in \{2, 3, 4\}$), while Table A6 reports the best-case subset under each schedule. For completeness, Table A7 provides the full subset-by-schedule RMSE matrix. In addition, the mathematical definitions of the different scheduling functions are provided as follows. Let $\tau \in \{0, \dots, T - 1\}$ and define the normalized diffusion time

$$u = \frac{\tau}{T - 1}. \quad (\text{A1})$$

We consider five monotone schedule families for β_{τ} . For a fair comparison, all schedules share the same endpoints $(\beta_{\text{start}}, \beta_{\text{end}})$ and the same number of steps T .

$$\textbf{Linear: } \beta_{\tau} = \beta_{\text{start}} + (\beta_{\text{end}} - \beta_{\text{start}}) u, \quad (\text{A2})$$

$$\textbf{Quadratic: } \beta_{\tau} = \beta_{\text{start}} + (\beta_{\text{end}} - \beta_{\text{start}}) u^2, \quad (\text{A3})$$

$$\textbf{Exponential: } \beta_{\tau} = \beta_{\text{start}} \left(\frac{\beta_{\text{end}}}{\beta_{\text{start}}} \right)^u, \quad (\text{A4})$$

$$\textbf{Sigmoid: } g(u) = \text{sigmoid}(\kappa(u - \tfrac{1}{2})), \quad \beta_{\tau} = \beta_{\text{start}} + (\beta_{\text{end}} - \beta_{\text{start}}) \frac{g(u) - g(0)}{g(1) - g(0)}, \quad (\text{A5})$$

$$\textbf{Cosine: } \bar{\alpha}(u) = \cos^2\left(\frac{\pi}{2} \cdot \frac{u + \epsilon}{1 + \epsilon}\right), \quad \bar{\alpha}_{\tau} = \bar{\alpha}\left(\frac{\tau}{T - 1}\right). \quad (\text{A6})$$

For the cosine family, we first derive an unnormalized $\tilde{\beta}_{\tau}$ from $\bar{\alpha}_{\tau}$ using

$$\tilde{\beta}_{\tau} = 1 - \frac{\bar{\alpha}_{\tau}}{\bar{\alpha}_{\tau-1}} \quad (\tau \geq 1), \quad \tilde{\beta}_0 = \tilde{\beta}_1. \quad (\text{A7})$$

We then linearly rescale $\tilde{\beta}_{\tau}$ to match the common endpoint range:

$$\beta_{\tau} = \beta_{\text{start}} + (\beta_{\text{end}} - \beta_{\text{start}}) \cdot \frac{\tilde{\beta}_{\tau} - \min_{\tau'} \tilde{\beta}_{\tau'}}{\max_{\tau'} \tilde{\beta}_{\tau'} - \min_{\tau'} \tilde{\beta}_{\tau'} + \delta}, \quad (\text{A8})$$

where δ is a small constant for numerical stability. Unless stated otherwise, we use $T = 50$, $\beta_{\text{start}} = 10^{-4}$, and $\beta_{\text{end}} = 0.02$.

Table A5. Schedule sensitivity across all 11 ensemble subsets ($K \in \{2, 3, 4\}$) with MLP depth = 4. We report test RMSE ($^{\circ}\text{C}$) mean $\pm$ std and the min/max over subsets. Lower is better.

Schedule	Mean $\pm$ Std	Min	Max
Exponential	0.362844 ± 0.005577	0.358443	0.371630
Cosine	0.362955 ± 0.005654	0.358298	0.371515
Quadratic	0.363018 ± 0.005517	0.358791	0.371663
Sigmoid	0.363220 ± 0.005538	0.358764	0.371615
Linear	0.363314 ± 0.005362	0.359125	0.371676

Table A6. Best-case subset under each schedule (MLP depth = 4), selected by the lowest test RMSE ($^{\circ}\text{C}$).

Schedule	Best Subset (K)	Test RMSE ($^{\circ}\text{C}$)	Test MAE ($^{\circ}\text{C}$)
Cosine	D + I + P + L (4)	0.358298	0.261126
Exponential	D + I + L (3)	0.358443	0.261240
Sigmoid	D + I + L (3)	0.358764	0.261607
Quadratic	I + P + L (3)	0.358791	0.261433
Linear	D + I + L (3)	0.359125	0.261746

Table A7. Full RMSE ($^{\circ}\text{C}$) matrix for schedule sensitivity (MLP depth = 4). Each column highlights the best subset for that schedule.

Subset	Linear	Cosine	Quadratic	Sigmoid	Exponential
D + L	0.359326	0.359068	0.358982	0.359233	0.359073
D + P	0.371676	0.371515	0.371663	0.371615	0.371630
D + I	0.370289	0.370074	0.369975	0.370347	0.369631
P + L	0.359841	0.359919	0.359803	0.359762	0.359864
I + L	0.359651	0.358999	0.359089	0.359364	0.358720
I + P	0.370618	0.370190	0.370290	0.370626	0.370122
D + P + L	0.359254	0.359168	0.358947	0.359123	0.358608
D + I + L	0.359125	0.358791	0.358963	0.358764	0.358443
D + I + P	0.370006	0.369484	0.369484	0.369652	0.369384
I + P + L	0.359472	0.359043	0.358791	0.359053	0.358725
D + I + P + L	0.359227	0.358298	0.358830	0.358781	0.358653

Appendix A.4. Best Ensemble Configurations

We report (i) the best ensemble configuration selected using validation RMSE (leak-free selection) and its corresponding test performance (Table A8), and (ii) the oracle best configuration by test RMSE for reference (Table A9). All configurations are selected from the candidate set that includes the following aggregation rules: uniform mean (mean), validation-optimized convex weights (cw), Bayesian model averaging (bma), quantile regression forests (qrf), and the proposed diffusion-weighted ensemble (diff). The “validation-selected” configuration is the only deployable choice under a proper generalization protocol. The “oracle” configuration is included for comparison only and must not be interpreted as a valid selection rule.

Table A8. Best ensemble configuration selected by validation RMSE (standardized).

Subset	K	Rule	Val RMSE (z)	Test RMSE (z)	Test R^2	Test RMSE ($^{\circ}\text{C}$)
D + I + P + L	4	diff	0.6115	0.6188	0.6918	0.3586

This follows the leak-free protocol: we select the subset+aggregation rule using validation only, then report the corresponding test performance once.

Table A9. Oracle best ensemble configuration by test RMSE (standardized).

Subset	K	Rule	Val RMSE (z)	Test RMSE (z)	Test R^2	Test RMSE (°C)
D + I + L	3	diff	0.6126	0.6185	0.6921	0.3584

Shown for reference only. Selecting a configuration by test performance is not permitted for proper generalization evaluation.

Appendix A.5. All Subset and Aggregation Combinations

Tables A10 and A11 enumerate all non-empty subsets of $\{D, I, P, L\}$ and all aggregation rules (uniform mean/convex weights/Bayesian model averaging/quantile regression forests/diffusion-weighted). This complete listing enables direct comparison between subset choice and aggregation strategy under the same evaluation pipeline.

Table A10. Complete results for all subset and aggregation-rule combinations (deep-learning ensembles only) (Part 1/2).

Subset	K	Rule	Val RMSE (z)	Val MAE (z)	Val R^2	Test RMSE (z)	Test MAE (z)	Test R^2	Test RMSE (°C)
D	1	single	0.6331	0.4510	0.6469	0.6478	0.4746	0.6622	0.375
L	1	single	0.6156	0.4406	0.6662	0.6231	0.4533	0.6874	0.361
P	1	single	0.6254	0.4475	0.6554	0.6449	0.4715	0.6652	0.374
I	1	single	0.6303	0.4505	0.6500	0.6472	0.4757	0.6628	0.375
D + L	2	cw	0.6141	0.4382	0.6678	0.6205	0.4518	0.6900	0.360
D + L	2	diff	0.6144	0.4386	0.6674	0.6198	0.4515	0.6907	0.359
D + L	2	mean	0.6166	0.4391	0.6650	0.6238	0.4553	0.6867	0.362
D + L	2	bma	0.6151	0.4391	0.6667	0.6224	0.4533	0.6883	0.361
D + L	2	qrf	0.6165	0.4395	0.6651	0.6246	0.4556	0.6859	0.362
D + P	2	cw	0.6230	0.4448	0.6581	0.6407	0.4687	0.6696	0.371
D + P	2	diff	0.6237	0.4454	0.6573	0.6428	0.4700	0.6672	0.372
D + P	2	mean	0.6236	0.4449	0.6574	0.6404	0.4687	0.6699	0.371
D + P	2	bma	0.6250	0.4460	0.6558	0.6446	0.4715	0.6651	0.374
D + P	2	qrf	0.6242	0.4453	0.6567	0.6422	0.4699	0.6678	0.372
D + I	2	cw	0.6255	0.4459	0.6553	0.6407	0.4701	0.6695	0.371
D + I	2	diff	0.6261	0.4464	0.6546	0.6417	0.4706	0.6683	0.372
D + I	2	mean	0.6254	0.4455	0.6555	0.6404	0.4696	0.6699	0.371
D + I	2	bma	0.6300	0.4493	0.6502	0.6471	0.4757	0.6629	0.375
D + I	2	qrf	0.6267	0.4469	0.6539	0.6429	0.4716	0.6670	0.373
P + L	2	cw	0.6128	0.4374	0.6693	0.6215	0.4522	0.6890	0.360
P + L	2	diff	0.6131	0.4376	0.6689	0.6212	0.4523	0.6894	0.360
P + L	2	mean	0.6136	0.4379	0.6683	0.6244	0.4549	0.6862	0.362
P + L	2	bma	0.6169	0.4402	0.6647	0.6238	0.4533	0.6869	0.362
P + L	2	qrf	0.6152	0.4389	0.6666	0.6250	0.4561	0.6855	0.362
I + L	2	cw	0.6130	0.4366	0.6690	0.6211	0.4523	0.6895	0.360
I + L	2	diff	0.6136	0.4371	0.6683	0.6220	0.4525	0.6884	0.360
I + L	2	mean	0.6145	0.4375	0.6673	0.6240	0.4553	0.6866	0.362
I + L	2	bma	0.6157	0.4406	0.6662	0.6230	0.4533	0.6875	0.361
I + L	2	qrf	0.6152	0.4381	0.6667	0.6241	0.4561	0.6865	0.362
I + P	2	cw	0.6206	0.4428	0.6607	0.6377	0.4668	0.6727	0.370
I + P	2	diff	0.6206	0.4428	0.6607	0.6377	0.4669	0.6727	0.370
I + P	2	mean	0.6207	0.4428	0.6607	0.6377	0.4668	0.6726	0.370
I + P	2	bma	0.6252	0.4460	0.6556	0.6447	0.4715	0.6649	0.374
I + P	2	qrf	0.6226	0.4440	0.6586	0.6407	0.4683	0.6696	0.371

Table A11. Complete results for all subset and aggregation-rule combinations (deep-learning ensembles only) (Part 2/2).

Subset	K	Rule	Val RMSE (z)	Val MAE (z)	Val R ²	Test RMSE (z)	Test MAE (z)	Test R ²	Test RMSE (°C)
D + P + L	3	cw	0.6127	0.4372	0.6693	0.6212	0.4520	0.6894	0.360
D + P + L	3	diff	0.6124	0.4370	0.6695	0.6206	0.4521	0.6902	0.360
D + P + L	3	mean	0.6157	0.4386	0.6660	0.6265	0.4573	0.6840	0.363
D + P + L	3	bma	0.6158	0.4408	0.6661	0.6236	0.4533	0.6871	0.362
D + P + L	3	qrf	0.6150	0.4387	0.6669	0.6233	0.4548	0.6875	0.361
D + I + L	3	cw	0.6130	0.4366	0.6690	0.6210	0.4523	0.6895	0.360
D + I + L	3	diff	0.6111	0.4357	0.6711	0.6184	0.4513	0.6921	0.358
D + I + L	3	mean	0.6164	0.4384	0.6654	0.6261	0.4575	0.6845	0.363
D + I + L	3	bma	0.6157	0.4406	0.6662	0.6230	0.4533	0.6875	0.361
D + I + L	3	qrf	0.6151	0.4385	0.6667	0.6239	0.4556	0.6866	0.362
D + I + P	3	cw	0.6205	0.4424	0.6609	0.6372	0.4664	0.6732	0.369
D + I + P	3	diff	0.6206	0.4425	0.6607	0.6373	0.4664	0.6731	0.369
D + I + P	3	mean	0.6211	0.4426	0.6602	0.6371	0.4665	0.6733	0.369
D + I + P	3	bma	0.6299	0.4492	0.6503	0.6470	0.4757	0.6631	0.375
D + I + P	3	qrf	0.6227	0.4441	0.6584	0.6405	0.4682	0.6699	0.371
I + P + L	3	cw	0.6122	0.4360	0.6698	0.6209	0.4521	0.6896	0.360
I + P + L	3	diff	0.6118	0.4358	0.6702	0.6194	0.4517	0.6913	0.359
I + P + L	3	mean	0.6138	0.4370	0.6682	0.6254	0.4564	0.6851	0.362
I + P + L	3	bma	0.6157	0.4406	0.6662	0.6230	0.4533	0.6875	0.361
I + P + L	3	qrf	0.6151	0.4384	0.6667	0.6239	0.4556	0.6866	0.362
D + I + P + L	4	cw	0.6122	0.4360	0.6698	0.6209	0.4521	0.6896	0.360
D + I + P + L	4	diff	0.6112	0.4357	0.6710	0.6206	0.4519	0.6899	0.360
D + I + P + L	4	mean	0.6155	0.4380	0.6662	0.6273	0.4583	0.6833	0.364
D + I + P + L	4	bma	0.6157	0.4406	0.6662	0.6230	0.4533	0.6875	0.361
D + I + P + L	4	qrf	0.6141	0.4371	0.6679	0.6250	0.4540	0.6854	0.362

References

- Deser, C.; Alexander, M.A.; Xie, S.P.; Phillips, A.S. Sea surface temperature variability: Patterns and mechanisms. *Annu. Rev. Mar. Sci.* **2010**, *2*, 115–143.
- Storto, A.; Alvera-Azcárate, A.; Balmaseda, M.A.; Barth, A.; Chevallier, M.; Counillon, F.; Domingues, C.M.; Drevillon, M.; Drillet, Y.; Forget, G.; et al. Ocean reanalyses: recent advances and unsolved challenges. *Front. Mar. Sci.* **2019**, *6*, 418.
- Sun, D.; Li, F.; Jing, Z.; Hu, S.; Zhang, B. Frequent marine heatwaves hidden below the surface of the global ocean. *Nat. Geosci.* **2023**, *16*, 1099–1104.
- Ham, Y.G.; Kim, J.H.; Luo, J.J. Deep learning for multi-year ENSO forecasts. *Nature* **2019**, *573*, 568–572.
- Tian, F.; Ren, H.L.; Liu, M.; Su, B.; Wang, R. Intensity and timing of persistence barriers of global sea surface temperature anomalies. *Geosci. Lett.* **2023**, *10*, 16.
- Cluett, A.; Jacox, M.; Amaya, D.; Alexander, M.; Scott, J. Atmospheric precursors of skillful SST prediction in the Northeast Pacific. *J. Clim.* **2024**, *37*, 5337–5353.
- Sengupta, A.; Waliser, D.E.; DeFlorio, M.J.; Guan, B.; Delle Monache, L.; Ralph, F.M. Role of evolving sea surface temperature modes of variability in improving seasonal precipitation forecasts. *Commun. Earth Environ.* **2025**, *6*, 256.
- Yang, Z.; Liu, J.; Yang, C.Y.; Hu, Y. Correcting Nonstationary Sea Surface Temperature Bias in NCEP CFSv2 Using Ensemble-Based Neural Networks. *J. Atmos. Ocean. Technol.* **2023**, *40*, 885–896.
- Montero-Manso, P.; Hyndman, R.J. Principles and algorithms for forecasting groups of time series: Locality and globality. *Int. J. Forecast.* **2021**, *37*, 1632–1653.
- Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; Sun, L. Transformers in time series: A survey. *arXiv* **2022**, arXiv:2202.07125.
- De León Pérez, D.; Salazar-Galán, S.; Francés, F. Beyond Deterministic Forecasts: A Scoping Review of Probabilistic Uncertainty Quantification in Short-to-Seasonal Hydrological Prediction. *Water* **2025**, *17*, 2932.
- Ehmimed, N.; Chkouri, M.Y.; Touhafi, A. Reliable and Adaptive Probabilistic Forecasting for Event-Driven Water-Quality Time Series Using a Gated Hybrid-Mixture Density Network. *Sensors* **2025**, *25*, 7560.
- Li, L.; Carver, R.; Lopez-Gomez, I.; Sha, F.; Anderson, J. Generative emulation of weather forecast ensembles with diffusion models. *Sci. Adv.* **2024**, *10*, eadk4489.

14. Chen, F.; Gao, L. Learning Residual Distributions with Diffusion Models for Probabilistic Wind Power Forecasting. *Energies* **2025**, *18*, 4226.
15. Su, C.; Cai, Z.; Tian, Y.; Chang, Z.; Zheng, Z.; Song, Y. Diffusion models for time series forecasting: A survey. *arXiv* **2025**, arXiv:2507.14507.
16. Barreto, F.T.; Curbani, F.E.; Zielinsky, G.M.; da Silva, M.B.; Lacerda, K.C.; Rodrigues, D.F. Development of a multigrid operational forecast system for the oceanic region off Rio de Janeiro State. *Ocean. Model.* **2023**, *184*, 102206.
17. Storto, A.; Frolov, S.; Slivinski, L.; Yang, C. Correction of sea surface biases in the NEMO ocean general circulation model using neural networks. *Geosci. Model Dev.* **2025**, *18*, 4789–4804.
18. Kong, W.; Lam, C.c.; Lau, D.s.; Chow, C.k.; Chong, S.n.; Chan, P.w.; Leung, N.c. Model validation and applications of wave and current forecasts from the Hong Kong Observatory's Operational Marine Forecasting System. *Ocean. Model.* **2024**, *190*, 102393.
19. de Boissésou, E.; Balmaseda, M.A. Predictability of marine heatwaves: assessment based on the ECMWF seasonal forecast system. *Ocean. Sci.* **2024**, *20*, 265–278.
20. Koul, V.; Brune, S.; Akimova, A.; Düsterhus, A.; Pieper, P.; Hövel, L.; Parekh, A.; Schrum, C.; Baehr, J. Seasonal prediction of Arabian Sea marine heatwaves. *Geophys. Res. Lett.* **2023**, *50*, e2023GL103975.
21. Hao, P.; Li, S.; Song, J.; Gao, Y. Prediction of sea surface temperature in the South China Sea based on deep learning. *Remote. Sens.* **2023**, *15*, 1656.
22. Xu, T.; Zhou, Z.; Li, Y.; Wang, C.; Liu, Y.; Rong, T. Short-term prediction of global sea surface temperature using deep learning networks. *J. Mar. Sci. Eng.* **2023**, *11*, 1352.
23. Zrira, N.; Kamal-Idrissi, A.; Farssi, R.; Khan, H.A. Time series prediction of sea surface temperature based on BiLSTM model with attention mechanism. *J. Sea Res.* **2024**, *198*, 102472.
24. Shi, B.; Ge, C.; Lin, H.; Xu, Y.; Tan, Q.; Peng, Y.; He, H. Sea surface temperature prediction using convlstm-based model with deformable attention. *Remote. Sens.* **2024**, *16*, 4126.
25. Fu, Y.; Song, J.; Guo, J.; Fu, Y.; Cai, Y. Prediction and analysis of sea surface temperature based on LSTM-transformer model. *Reg. Stud. Mar. Sci.* **2024**, *78*, 103726.
26. Yang, Y.; Lam, K.M.; Dong, J.; Ju, Y. Multi-Factor Deep Learning Model for Sea Surface Temperature Forecasting. *Remote Sens.* **2025**, *17*, 752.
27. Dai, H.; Lei, F.; Wei, G.; Zhang, X.; Lin, R.; Zhang, W.; Shang, S. Sea surface temperature prediction by stacked generalization ensemble of deep learning. *Deep. Sea Res. Part I Oceanogr. Res. Pap.* **2024**, *209*, 104343.
28. Qian, J.; Wang, Q.; Liang, P.; Peng, S.; Wang, H.; Wu, Y. Deep learning-based ensemble forecast and predictability analysis of the Kuroshio intrusion into the south China sea. *J. Phys. Oceanogr.* **2024**, *54*, 1503–1517.
29. Bonino, G.; Galimberti, G.; Masina, S.; McAdam, R.; Clementi, E. Machine learning methods to predict sea surface temperature and marine heatwave occurrence: A case study of the Mediterranean Sea. *Ocean. Sci.* **2024**, *20*, 417–432.
30. Boschetti, F.; Feng, M.; Hartog, J.R.; Hobday, A.J.; Zhang, X. Sea surface temperature predictability assessment with an ensemble machine learning method using climate model simulations. *Deep. Sea Res. Part II Top. Stud. Oceanogr.* **2023**, *210*, 105308.
31. Behringer, D.W.; Ji, M.; Leetmaa, A. An improved coupled model for ENSO prediction and implications for ocean initialization. Part I: The ocean data assimilation system. *Mon. Weather. Rev.* **1998**, *126*, 1013–1021.
32. Zuo, H.; Balmaseda, M.A.; Tietsche, S.; Mogensen, K.; Mayer, M. The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: A description of the system and assessment. *Ocean. Sci.* **2019**, *15*, 779–808.
33. Jean-Michel, L.; Eric, G.; Romain, B.B.; Gilles, G.; Angélique, M.; Marie, D.; Clément, B.; Mathieu, H.; Olivier, L.G.; Charly, R.; et al. The Copernicus global 1/12 oceanic and sea ice GLORYS12 reanalysis. *Front. Earth Sci.* **2021**, *9*, 698876.
34. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. <https://doi.org/10.1023/A:1010933404324>.
35. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>.
36. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19), Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. <https://doi.org/10.1145/3292500.3330701>.
37. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv* **2020**, arXiv:2006.11239.
38. Karras, T.; Aittala, M.; Aila, T.; Laine, S. Elucidating the Design Space of Diffusion-Based Generative Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LO, USA, 28 November– 9 December 2022.
39. Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; Zhu, J. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LO, USA, 28 November– 9 December 2022.

40. Song, Y.; Dhariwal, P.; Chen, M.; Sutskever, I. Consistency Models. In Proceedings of the International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023.
41. Liu, X.; Gong, C.; Liu, Q. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *arXiv* **2022**, arXiv:2209.03003.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.