MLAWind: A Monthly Sea Surface Wind Dataset Derived from an Interpretable Machine Learning Approach Integrating In-Situ Observations and Satellite Data

Weihao Guo[1,2,3], Rongwang Zhang[1,2,3], Xin Wang[1,2,3], Dongxiao Wang[4,5]

Affiliations

[1]State Key Laboratory of Tropical Oceanography, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, 510301, China.

[2]Global Ocean and Climate Research Center, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, 510301, China.

[3]Guangdong Key Laboratory of Ocean Remote Sensing and Big Data, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, 510301, China.

[4]School of Marine Sciences, Sun Yat-Sen University, Zhuhai, 519082, China.

[5]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, 519080, China.

Correspondence to: Rongwang Zhang (rwzhang@scsio.ac.cn); Xin Wang (wangxin@scsio.ac.cn)

Abstract. A gridded sea surface wind dataset with long temporal coverage is crucial for understanding atmospheric circulation changes and air-sea interactions at different time scales. This study employs an interpretable machine learning model based on random forest algorithm to generate a 1°×1° monthly sea surface wind dataset (MLAWind) from 1950 to 2023, covering the near-global ocean within 60°S–60°N. The data reconstruction model integrates the Cross-Calibrated Multi-Platform (CCMP) satellite data and the spatially sparse long-term International Comprehensive Ocean-Atmosphere Data Set (ICOADS), exhibiting robust interpretability and generalization capability. Quantitative validations against satellite and in-situ data indicate that MLAWind achieves better consistency with observations than most reanalysis datasets, maintaining robust performance during both training (1993–2022) and testing periods (2023). The performance of MLAWind since 1950 is assessed across multiple time scales. Its characteristics in climatology, annual cycle, and inter-annual variability are comparable to those of reanalysis datasets, even during the non-satellite period prior to 1993. Independent coral records from 1950–1982 are employed as proxy indicators to evaluate long-term sea surface wind trends. The trend derived from MLAWind is corroborated by these proxies, exhibiting significantly higher temporal correlation than those obtained from reanalysis products. MLAWind demonstrates a more reliable reconstruction of long-term sea surface wind variability during historical periods since 1950. The results indicate that MLAWind serves as a robust data resource for global climate change research. The reconstructed MLAWind dataset is publicly accessible at https://doi.org/10.5281/zenodo.17354864 (Guo et al., 2025b).

1. Introduction

Sea surface wind is a critical factor in air-sea interactions, exerting significant impacts on climate change, marine ecosystem and human society (Tokinaga et al., 2012; Wang et al., 2018; Zhou et al., 2022). It governs variations of the hydrological cycle by modulating atmosphere-ocean heat and moisture exchanges (Held and Soden, 2006; Findell et al., 2019). Sea surface wind drives the large-scale redistribution of water masses and energy. The global ocean current systems, including the Kuroshio, Gulf Stream, South China Sea Throughflow, Indonesian Throughflow, and Antarctic Circumpolar Current, are primarily regulated by wind-driven Ekman transport and momentum transfer processes (Deser et al., 1999; Wang et al., 2006, 2023; Zhang et al., 2023; Li et al., 2025). Sea surface wind can deliver energy to the deep ocean, inducing near-inertial waves and near-bottom currents (Zhang et al., 2024). Previous studies have demonstrated that equatorial sea surface wind anomalies are associated with the inter-annual variability of the tropical sea surface temperature (SST) anomalies, which promotes the development of El Niño-Southern Oscillation (ENSO) events (Kug, et al., 2009; Clarke, 2014; Wang, 2019). Sea surface wind is thus utilized as a precursor to improve the predictability of different types of ENSO (Ren et al., 2019; Tseng et al., 2022).

The existing observational sea surface wind datasets exhibit notable limitations despite remarkable advances in observation platforms and sensor technologies. Satellite data are characterized by exceptional spatiotemporal continuity, offering a robust data foundation for retrieving climate system characteristics at the global scale, including those associated with extreme meteorological events (Vinoth and Young, 2011; Yang et al., 2015; Young and Ribal, 2019). However, the relatively short temporal coverage of satellite observations constrains their capacity to analyze climate variability from inter-decadal time scale to long-term trends. The International Comprehensive Ocean-Atmosphere Data Set (ICOADS) is widely acknowledged as the most

extensive and comprehensive global surface ocean dataset. It provides in-situ observations of critical atmospheric and oceanic variables dating back to 1662 (Freeman et al., 2017). In contrast to the extensive spatial coverage of satellite data, ICOADS exhibit an irregular distribution and suffer from significant sampling errors. Due to these deficiencies in existing observational datasets, the Sixth Assessment Report by the Intergovernmental Panel on Climate Change (IPCC AR6) assigns a 'low to medium' confidence level for the historical sea surface wind trends assessment (IPCC, 2023).

A long-term gridded global wind dataset can be constructed by integrating in-situ and satellite observations. Numerous reanalysis products have assimilated the ICOADS data and multi-platform satellite observations through data assimilation techniques, significantly extending the temporal-spatial coverage of the sea surface wind field (e.g. Kalnay et al., 1996; Kobayashi et al.,2015; Hersbach et al., 2019). The inherent limitations of reanalysis datasets primarily stem from two aspects: (1) the assimilation methodologies and models introduce significant biases into the reanalysis products; (2) variations in the types and volumes of in-situ data result in sampling biases. Substantial uncertainties persist in reanalysis datasets with different assimilation methodologies and data sources (Zhang et al., 2023).

In addition to data assimilation techniques, direct reconstruction methods based on ICOADS data are employed in some studies (Berry and Kent, 2011; Tokinaga et al., 2011, 2012). For example, the Wave- and Anemometer-Based Sea Surface Wind (WASWind) achieves high accuracy by integrating anemometer-measured winds and wave height-estimated winds from the ICOADS data (Tokinaga and Xie, 2011). WASWind exhibits a relatively coarse spatial resolution of 4°×4° and limited temporal coverage from 1950 to 2008. The National Oceanography Centre Southampton Flux Dataset v2.0 (NOCS2) is another reconstructed dataset derived from the ICOADS data. It features an enhanced spatial resolution of 1°×1° but covers a shorter time period (1973–

2014) for sea surface wind speed (Berry and Kent, 2011).

Machine learning algorithms demonstrate superior capability in capturing nonlinear relationships between different variables compared to traditional techniques (Reichstein et al., 2019; Jiang et al., 2024; Wang and Li, 2024). These algorithms are applied in the reconstruction of various datasets such as sea surface temperature (Huang et al., 2025), air temperature (Yoo et al., 2018; He et al., 2022), ocean salinity (Tian et al., 2022), ocean heat content (Su et al., 2020; Bagnell and DeVries, 2021), carbon and water fluxes (Leng et al., 2024), and terrestrial water storage (Yin et al., 2023). Traditional machine learning models suffer from interpretability challenges, as their input-output mappings are treated as "black-box" systems lacking physically meaningful explanations. Recently, resolving machine learning interpretability issues has emerged as a critical research priority (Tian et al., 2022; Qin et al., 2024). SHapley Additive exPlanations (SHAP) is a reliable interpretable module based on cooperative game theory, which demonstrates excellent compatibility across diverse machine learning models (Lundberg and Lee, 2017; Kim et al., 2025). By generating both global and local interpretability, SHAP enables a more comprehensive attribution analysis than conventional feature importance metrics. The continuous advancement of interpretable machine learning techniques offers a robust methodology for historical sea surface wind reconstruction.

This study employs an interpretable machine learning algorithm (Tian et al., 2022; Kim et al., 2025) to integrate ICOADS and satellite data, constructing a 1°×1° monthly sea surface wind dataset from 1950–2023 with near-global ocean coverage within 60°S–60°N. The rest of the paper is organized as follows: the datasets, methods and data reconstruction model used in this study are described in Sect. 2; the evaluation processes of the reconstructed dataset are shown in Sect. 3; the performances of the reconstructed dataset at different time scales are examined in Sect. 4; the data availability statement is provided in

Sect. 5; the study is summarized in Sect. 6.

2. Data and methods

2.1 Input data for model establishment

The reconstructed dataset is derived from in-situ observations of the ICOADS dataset. The ICOADS dataset provides an extensive collection of observations from diverse observing systems, containing multiple essential atmospheric and oceanic variables (Freeman et al., 2017). The incorporation of physically relevant feature inputs significantly enhances the reconstruction accuracy of machine learning models (Tian et al., 2022; Zhang et al., 2022). Multiple variables closely associated with sea surface winds are utilized in this study, including sea surface wind speed (WS), wind direction (WD), SST, air temperature (Ta), and sea level pressure (SLP) from 1950 to 2023. Note that the ICOADS observations exhibit pronounced spatiotemporal heterogeneity. During the period 1950–1969, data coverage is relatively limited in the Southern Hemisphere, contrasting with denser distributions in the North Pacific and North Atlantic Oceans. From 1980 to 1999, the data volume gradually increases, followed by significant growth after 2000, which achieves near-global coverage (Fig. 1).

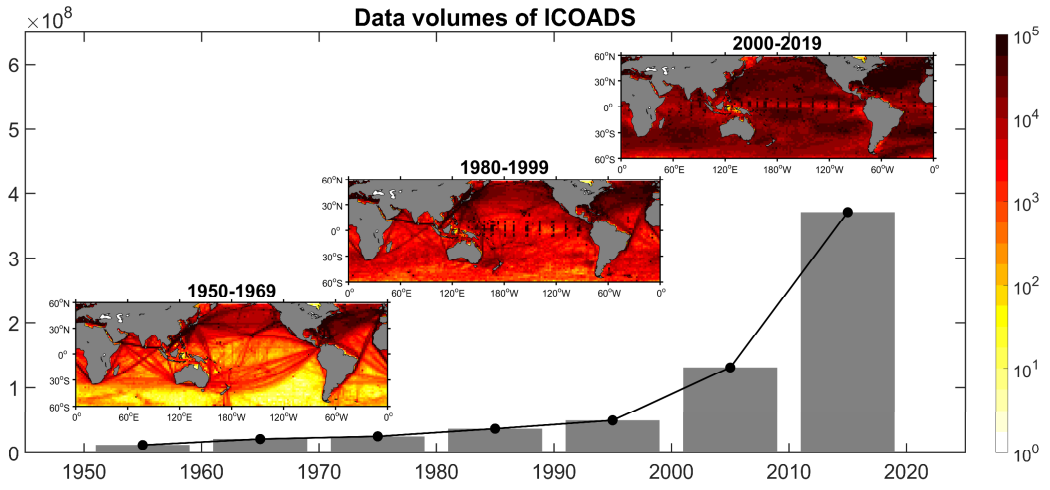


Figure 1. The volumes of ICOADS dataset since 1950. The bar illustrates the decadal cumulative volume of ICOADS dataset. The spatial distributions of the volumes of ICOADS dataset are presented in different periods.

Following previous studies (Berry and Kent, 2011; Tokinaga and Xie, 2011), comprehensive preprocessing is performed on the ICOADS dataset. All variables in the ICOADS dataset are subjected to rigorous quality control through an outlier detection procedure based on a climatological threshold of 3.5 standard deviations. The selection of the threshold aligns with the standardized processing implemented for both the ICOADS Monthly Summary Groups (MSG) product and the WASWind dataset (Tokinaga and Xie, 2011). Additional height adjustments are required for ICOADS sea surface winds, as the altitudes of ships and measurement instruments change over time. The bulk formulae and related parameters established by Smith (1980) are utilized to unify the unadjusted ICOADS sea surface winds to a standard reference height of 10 m. All data are initially processed into monthly means to enable point-to-point learning during model establishment process.

Satellite sea surface wind from the cross-calibrated multiplatform (CCMP) dataset is another important source of the reconstructed dataset. The CCMP dataset covers the period from 1993 to 2023 at a horizontal resolution of 0.25°×0.25° (Mears et al., 2019). It is spatially interpolated onto a 1°×1° grid to facilitate point-to-point learning with ICOADS dataset. The CCMP integrates multiple advanced satellites, including Advanced Scatterometer-A and B (ASCAT-A/B), Special Sensor Microwave/Imager (SSM/I), Special Sensor Microwave Imager/Sounder (SSMIS), Tropical Rainfall Measuring Mission's (TRMM) Microwave Imager (TMI), Global Precipitation Measurement (GPM) Microwave Imager (GMI), Advanced Microwave Scanning Radiometer-EOS (ASMR-E), Advanced Microwave Scanning Radiometer 2 (AMSR2), Quick Scaterometer (QuikScat), and Wind Satellite (WindSat). All satellite data are consistently converted to10-m neutral stability winds, followed by inter-satellite cross-calibration to minimize systematic biases. Moored buoys are used to calibrate the satellite sensors, adjusting the CCMP background surface wind field during the data assimilation phase.

Therefore, the CCMP demonstrates strong agreement with in-situ observations (Mears et al., 2019).

2.2 Data for evaluation

Five reanalysis products, including European Centre for Medium-Range Weather Forecasts (ERA5, Hersbach et al., 2019), the Japanese 55-year Reanalysis (JRA-55, Kobayashi et al., 2015), the National Oceanic and Atmospheric Administration 20th-Century Reanalysis (NOAA-20C, Compo et al., 2011), the National Centers for Environmental Prediction reanalysis 1 (NCEP1, Kalnay et al., 1996) and its upgraded version (NCEP2, Kanamitsu et al., 2002), are compared with the reconstructed dataset. All reanalysis products are spatially interpolated onto a 1°×1° grid. Comprehensive descriptions of these datasets are presented in Table 1.

The Global Tropical Moored Buoy Array (GTMBA) in-situ observation is employed to evaluate the performance of multiple sea surface wind datasets. The GTMBA provides real-time observations in the tropics for weather and climate research. It integrates three buoy arrays, including the Tropical Atmosphere Ocean/Triangle Trans-Ocean Buoy Network (TAO/TRITON) in the Pacific Ocean, the Prediction and Research Moored Array in the Tropical Atlantic (PIRATA), and the Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction (RAMA) in the Indian Ocean (McPhaden et al., 2010). TAO provides a long-term observational record dating back to the early 1990s, while PIRATA and RAMA were deployed in 1997 and 2002, respectively. 56 TAO stations, 19 PIRATA stations, and 28 RAMA stations are utilized in this study.

Three coral Mn/Ca records provided by *Porites* spp. corals are utilized to evaluate the long-term trends of sea surface wind. The coral record derived from Tarawa Atoll (1.33°N, 172.97°E) is resampled at approximately quarterly (4 per year) increments, spanning from 1894 to 1982 (Thompson et al., 2015).

The coral records from Butaritari Atoll (3.07°N, 172.75°E) and Kiritimati Island (1.93°N, 157.49°W) exhibit bimonthly temporal resolution, covering the periods of 1989–2010 and 1994–2011, respectively (Sayani et al., 2021). These islands are located far from continental Mn sources and possess a west-facing lagoon that is sheltered from the easterly trade winds. Their Mn/Ca records demonstrate a strong correlation with westerly wind events in the equatorial Pacific Ocean, serving as a valuable and independent resource for evaluating the zonal wind anomalies (Thompson et al., 2015; Sayani et al., 2021).

Table 1. A list of the datasets used in this study.

Product		Variable	Data period	Resolution	Reference
In-situ observations	ICOADS	WS, WD, SST, Ta, SLP	1950–2023	-	Freeman et al. (2017)
	GTMBA	U, V, WS	1993–2023	-	McPhaden et al. (2010)
Satellite data	CCMP	U, V, WS	1993–2023	0.25°×0.25°	Mears et al. (2019)
Reanalysis datasets	ERA5	U, V, WS	1950–2022	0.25°×0.25°	Hersbach et al. (2019)
	NCEP1	U, V, WS	1950–2022	1.875°×1.90°	Kalnay et al. (1996)
	NCEP2	U, V, WS	1979–2022	1.875°×1.90°	Kanamitsu et al. (2002)
	JRA-55	U, V, WS	1958–2022	1.25°×1.25°	Kobayashi et al. (2015)
	NOAA-20C	U, V, WS	1950–2014	2°×2°	Compo et al. (2011)
Corals	Tarawa	Mn/Ca	1950–1982	-	Thompson et al. (2015)
	Butaritari	Mn/Ca	1989–2010	-	Sayani et al. (2021)
	Kiritimati	Mn/Ca	1994–2011	-	Sayani et al. (2021)

2.3 Data reconstruction model

Figure 2 illustrates the schematic diagram of the construction process of the machine learning-assisted sea surface wind dataset (MLAWind). The random forest model is employed as the core algorithm. It effectively captures nonlinear relationships between feature and label variables through multiple decision trees (Breiman, 2001) and is extensively employed for temporal-

spatial data processing (Yoo et al., 2018; Watt-Meyer et al., 2021; He et al., 2022). Notably, the inherent randomness in random forest model significantly improves its generalization capability, which is critical for robust reconstruction of long-term historical sea surface wind.

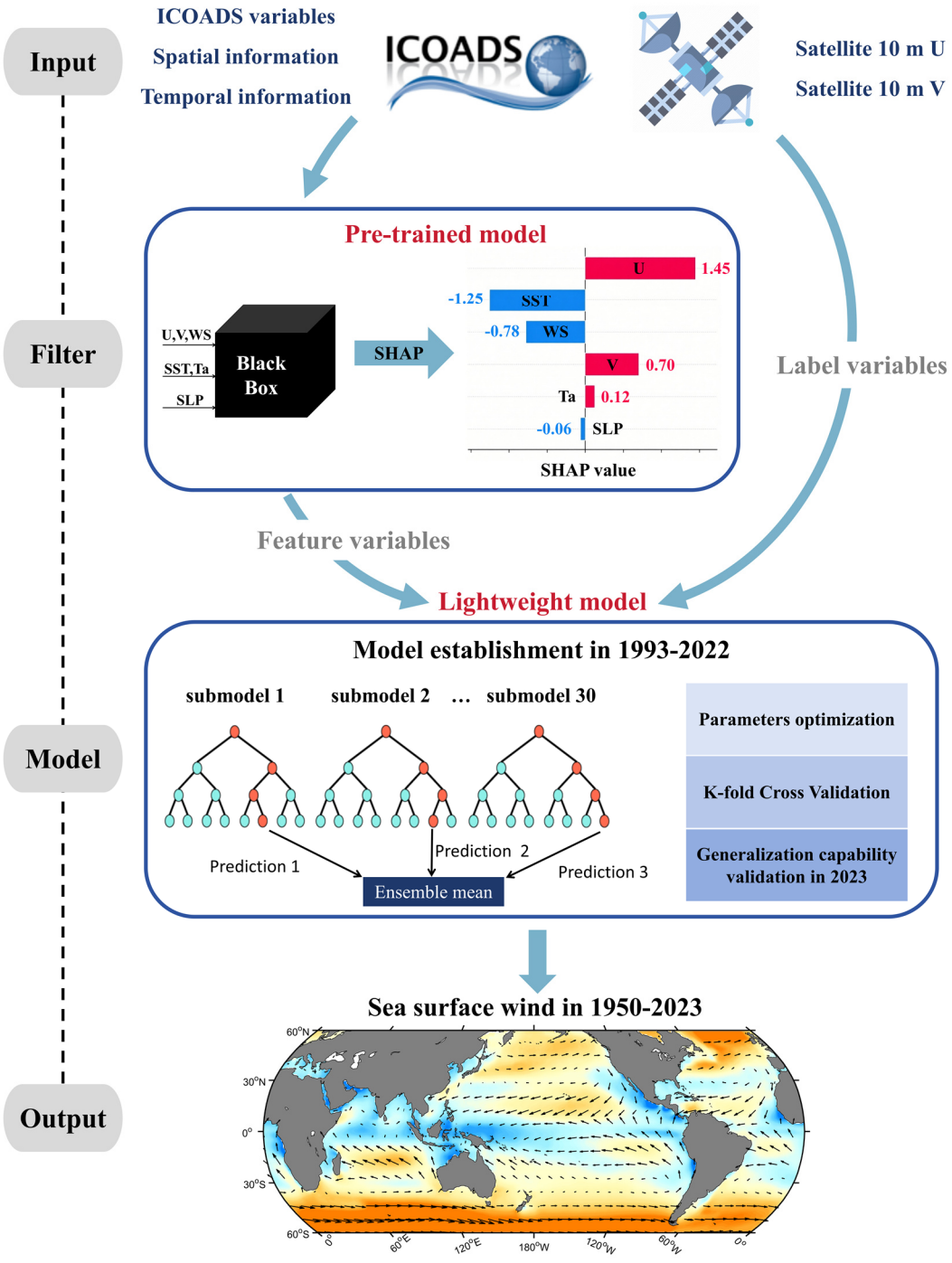


Figure 2. Schematic diagram of the construction process of machine learning-assisted sea surface wind dataset (MLAWind).

The reconstruction framework comprises four essential steps. The first is tokenization and preprocessing of the input data. Given that 10-m zonal wind, 10-m meridional wind and WS are routinely available from widely-used satellite and reanalysis products, these variables are selected as reconstruction targets to enable systematic inter-product comparison. Multiple ICOADS variables (WS, WD, SST, Ta, SLP) from 1993–2022 are used as feature inputs, while the CCMP 10-m wind field from 1993–2022 serves as the label variable for data reconstruction model establishment. The ICOADS data from 1950–1992 are employed in the model inference process to enable reconstruction of sea surface winds during historical periods without satellite data. All feature values for the data reconstruction model are normalized to enhance the efficiency of parameter optimization (Su et al., 2018; Watt-Meyer et al., 2021).

The second step is establishing a pre-trained model integrating the SHAP algorithm to achieve synergistic optimization of model efficiency and interpretability. Specifically, the feature set comprising multiple wind-related variables and their spatiotemporal information is input into a pre-trained model. The SHAP algorithm quantifies the contribution of each feature to the model's output, thereby establishing a feature importance ranking (Lundberg and Lee, 2017). The top-K (e.g., 5–10) highest-contributing features are retained, while redundant variables with relative contribution below 10% are eliminated to reduce model complexity and enhance physical interpretability. Based on the SHAP analysis, the ICOADS 10-m zonal wind, meridional wind, SST, WS, longitude, latitude, and month are identified as the seven key feature variables for sea surface wind reconstruction. They are utilized to establish a lightweight model, while other variables with little contribution (i.e. Ta, SLP and year) are excluded.

Thirdly, the lightweight model is trained from 1993–2022, facilitating point-to-point learning between ICOADS variables and CCMP sea surface winds. Parameter tuning, K-fold Cross Validation, and generalization

capability validation are three key components for lightweight models: (1) a grid search approach is employed to identify the optimal model parameters. Consequently, the number of regression trees, the minimum number of samples required to split and samples required to be at a leaf node is set to 200, 3, 3, respectively; (2) leave-one-out Cross Validation (LOOCV), a specialized form of K-fold Cross Validation is implemented during the model establishment process (Mao et al., 2014). The 30-year samples from 1993–2022 are segmented into 30 submodels, with each submodel consisting of a 29-year training set and a 1-year validation set. LOOCV exhibits minimal discrepancies among these model, indicating the lightweight model's robust performance; (3) an independent verification is conducted in 2023 to evaluate model's generalization capability. The results will be presented in Sect. 3.

The final step is the model inference process that generates the restructured dataset. The process operates without label variable. Following the training process conducted from 1993–2022, the machine learning model derives a nonlinear relationship between ICOADS variables and CCMP sea surface wind. The established relationship enables the reconstruction of historical sea surface winds during the non-satellite periods by utilizing ICOADS data from 1950–1992. The resultant monthly MLAWind dataset spanning 1950 to 2023, covering near-global oceans between 60°S and 60°N at a horizontal resolution of 1°×1°.

Figure 3 exhibits the reconstruction efficacy in 1950 and 2000. The two years are chosen as examples to represent periods with sparse and abundant original data coverage, respectively (Fig. 1). The original ICOADS from 1950 exhibits significant spatial data gaps, particularly in the tropical Pacific and Southern Ocean. MLAWind shows continuous spatial coverage. Despite limited observations in the Southern Ocean, it demonstrates robust reconstruction of the prevailing westerly wind belt (Garreaud et al., 2013). In 2000, ICOADS exhibits a higher spatial density, while still containing considerable noise.

MLAWind demonstrates smoother sea surface wind characteristics. The CCMP satellite observation is used to evaluate the biases of ICOADS and MLAWind in 2000. ICOADS exhibits significant systematic biases, primarily attributed to uncertainties in ship height measurements and instrumental calibration protocols (Berry and Kent, 2011; Tokinaga and Xie, 2011). MLAWind effectively mitigates these biases and demonstrates high consistency with the CCMP, confirming the validity of the reconstruction.

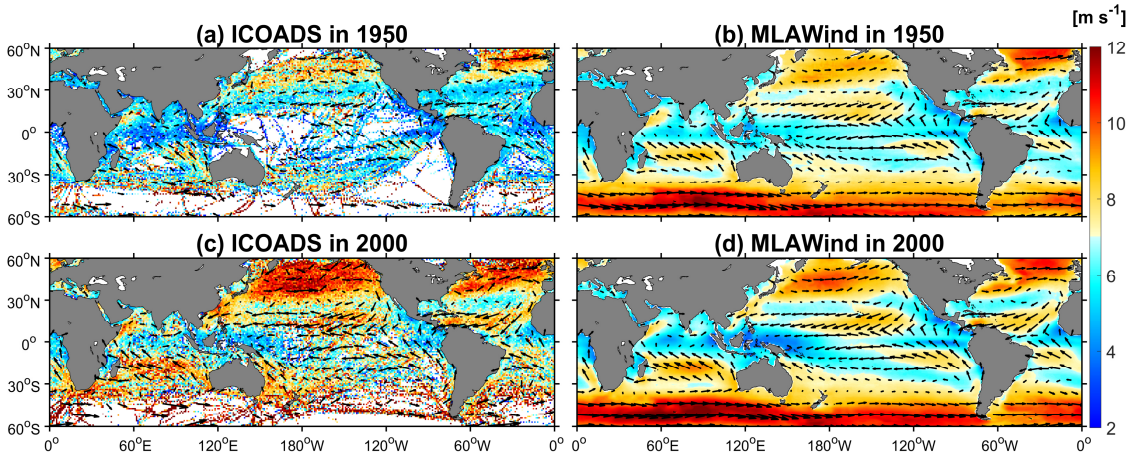


Figure 3. Reconstruction results presentation. 10-m wind velocity (vector, m s^{-1}) and sea surface wind speed (shading, m s^{-1}) of the original ICOADS and MLAWind dataset are compared in 1950 and 2000, respectively. These two years are chosen as examples to represent periods with sparse and abundant original data coverage, respectively.

2.4 Evaluation metrics

Five metrics are employed in the evaluation process, including mean error (Bias), root mean square error (RMSE), Pearson correlation coefficient (CORR), skewness and the coefficient of determination (R-squared). Bias, RMSE and CORR quantify discrepancies between observations and products. Skewness characterizes the asymmetry of a variable's probability distribution. Positive values denote right-skewed distributions with longer tails toward large values and a left-shifted peak. R-square serves as a key metric for assessing model performance. Its values span from 0 to 1, with those closer to 1 indicating

better model performance. These metrics can be described as follows:

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

$$CORR = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}} \quad (3)$$

$$Skewness = \frac{\sqrt{n(n-1)}}{n-2}\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^3}{(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2})^3} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{\hat{y}})^2} \quad (5)$$

where n represents the number of samples, y_i represents the values obtained from different products, $\hat{y}_i$ represents observational values. $\bar{y}$ and $\bar{\hat{y}}$ represent the mean values of y_i and $\hat{y}_i$, respectively.

3. Evaluation of the MLAWind dataset

Based on satellite data, we systematically evaluate the performance of the MLAWind dataset on the training set (1993–2022) and testing set (2023), respectively (Figs. 4-6). During the training period, the MLAWind dataset shows the smallest mean biases for both 10-m zonal and meridional winds (Fig. 4a, e). It achieves the lowest RMSE (0.21 m s^{-1}) for meridional wind, while its RMSE (0.26 m s^{-1}) ranks second only to the ERA5 dataset (0.16 m s^{-1}) for zonal wind. The mean bias and RMSE are quantified across the Pacific, Indian, and Atlantic Oceans. Their performance is comparable to those of near-global coverage for both 10-m zonal (Fig. 4b-d) and meridional wind (Fig. 4f-h). A global evaluation of WD is performed to further confirm the reliability of the MLAWind dataset. The MLAWind dataset demonstrates the lowest mean bias and the second-lowest RMSE across all evaluated products. Despite the absence of direct reconstruction of WD, the MLAWind dataset derives WD with high accuracy.

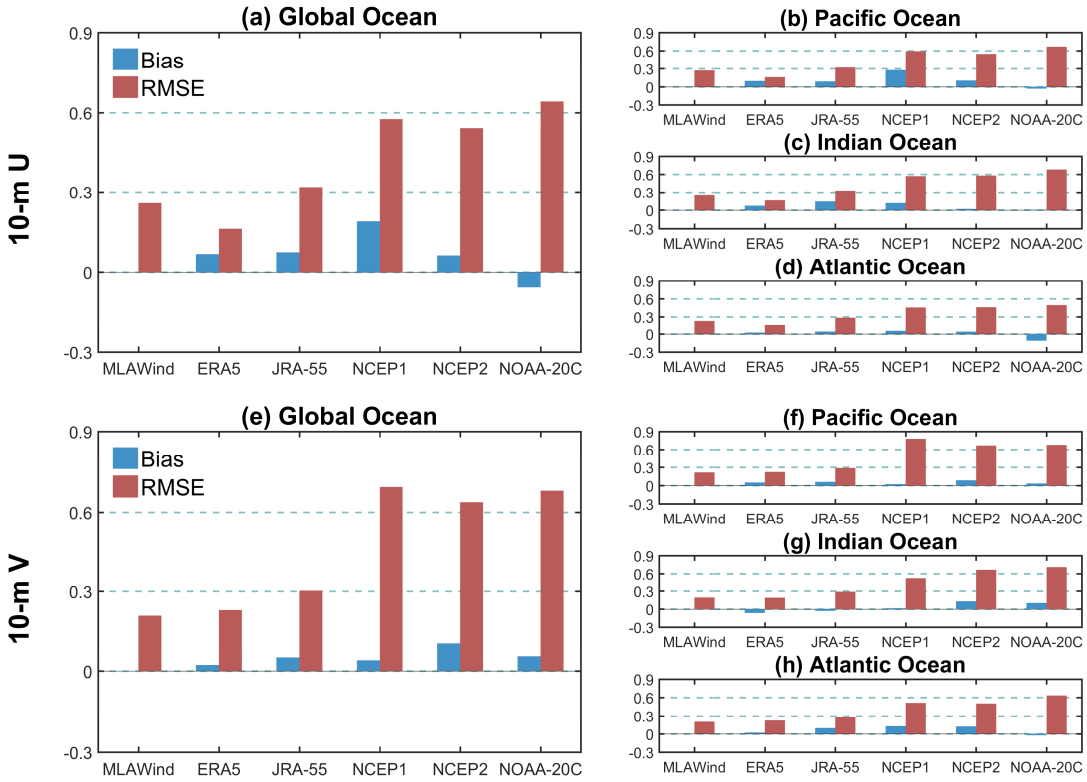


Figure 4. Evaluation based on satellite sea surface wind. (a-d) Bias and RMSE values (m s^{-1}) of multiple datasets evaluated with CCMP 10-m zonal wind over different oceans within 60°S–60°N from 1993–2022. (e-h) Same as (a-d) but for 10-m meridional wind.

The probability density function (PDF) characteristics of different sea surface wind datasets are compared from 1993–2022. The 10-m zonal wind in CCMP, MLAWind, and reanalysis datasets collectively exhibit positive skewness distributions with longer tails toward larger values and left-shifted peaks (Fig. 5a-c). MLAWind demonstrates the closest skewness value (0.19) to CCMP satellite data (0.21) among all datasets. The 10-m meridional wind in MLAWind and CCMP are consistent in their negative skewness distributions, whereas substantial divergences emerge across reanalysis datasets (Fig. 5d-f). In contrast to the CCMP dataset, the 10-m meridional wind in NCEP1, NCEP2, and NOAA-20C exhibit positive skewness in their distributions (Fig. 5f), which may be attributed to their large RMSE with satellite data (Fig. 4).

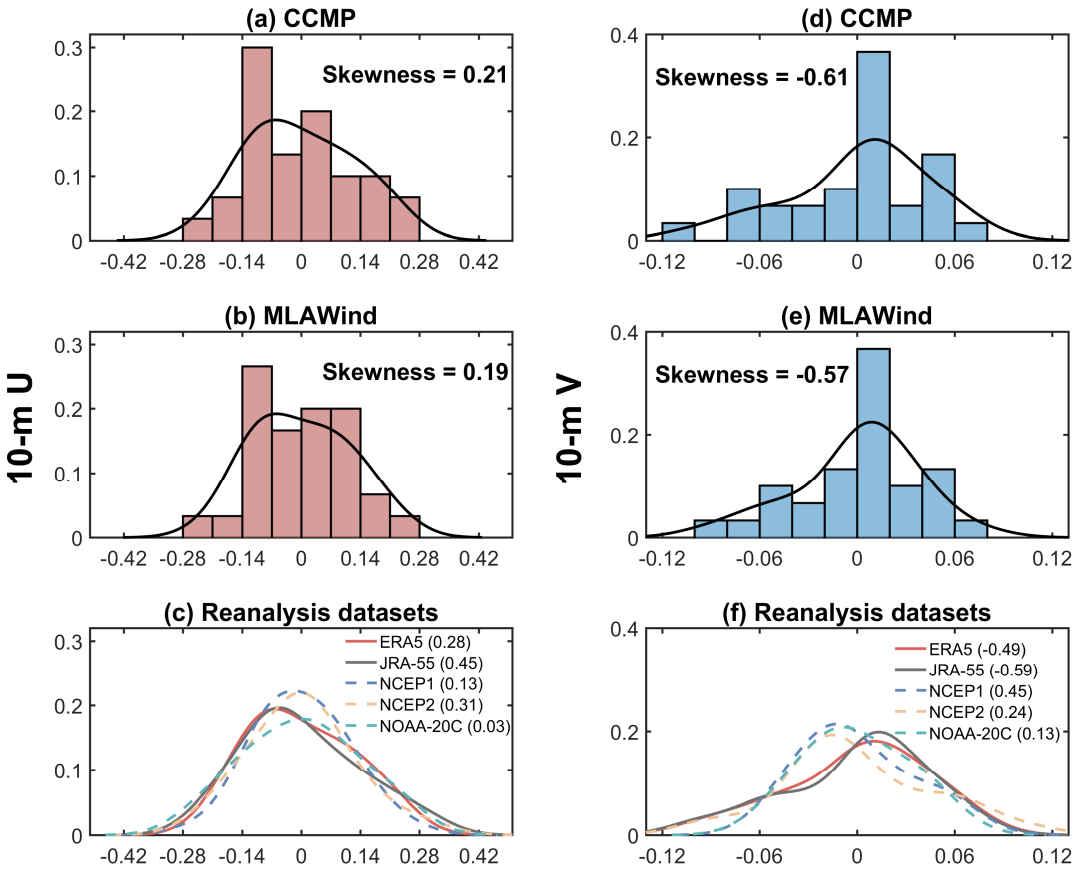


Figure 5. Comparison of probability density functions (PDF) among different datasets. (a-c) PDF and skewness coefficients of 10-m zonal wind in different datasets from 1993–2022. The horizontal axis denotes 10-m zonal wind anomalies. The vertical axis denotes probability densities. The values in parentheses in (c) denote skewness coefficients of reanalysis datasets. (d-f) Same as (a-c) but for 10-m meridional wind.

During the testing phase, the MLAWind dataset maintains strong agreement with satellite observations in 2023 (Fig. 6b). Most values fall near the 1:1 reference line, with relatively larger errors under high wind conditions (≥15 m/s). The mean bias and RMSE between the MLAWind and CCMP datasets are 0.18 m s^{-1} and 0.45 m s^{-1}, respectively. Notably, the MLAWind dataset achieves R-squared values exceeding 0.90 for both the training and testing sets (Table 2). The high consistency exhibits its robust generalization capabilities without signs of overfitting.

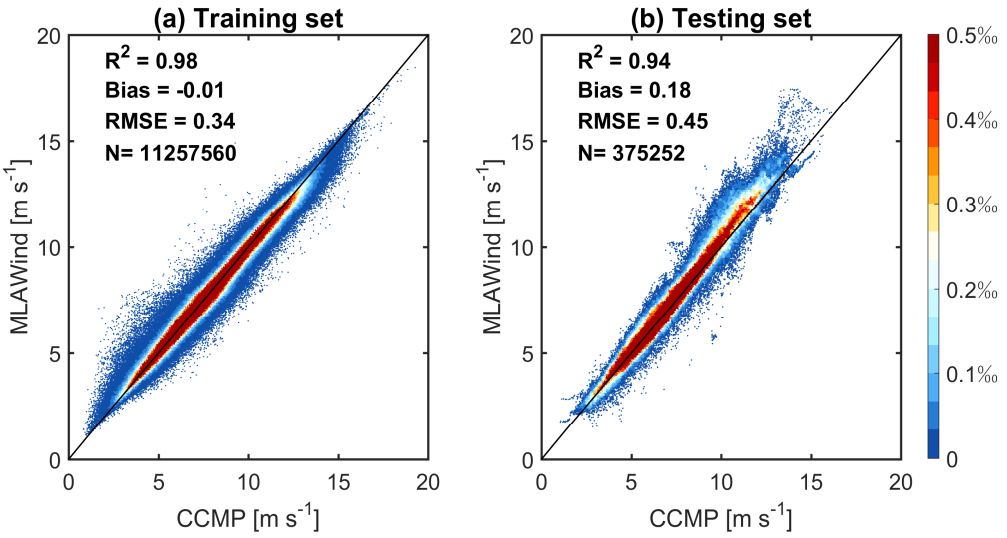


Figure 6. Validation of MLAWind's generalization capability. Density scatter plots compare the MLAWind and CCMP datasets during (a) training phase (1993–2022) and (b) testing phase (2023), respectively. The shading represents data density within each bin. R^2, mean bias, RMSE and sample size (N) are used as quantitative metrics for evaluating MLAWind's performance.

Table 2. Comparison of MLAWind's performance on the training set (1993–2022) and testing set (2023). The CCMP satellite data is used as the evaluation benchmark.

Evaluation metric	Training set	Testing set
R^2	0.98	0.94
Bias (m s^{-1})	-0.01	0.18
RMSE (m s^{-1})	0.34	0.45
N	11257560	375252

In addition to satellite data, the performance of multiple datasets from 1993 to 2023 is compared based on the GTMBA in-situ observations (Fig. 7). MLAWind and ERA5 demonstrate high consistency with the GTMBA, exhibiting smaller errors than other datasets. These two datasets equally achieve low RMSE (~0.40 m/s) and high CORR (0.95), although MLAWind demonstrates a slightly larger bias than ERA5. Note that ERA5 assimilates the GTMBA and thus its evaluation results are not independent. MLAWind achieves reconstruction accuracy comparable to ERA5, without relying on in-situ buoy observations.

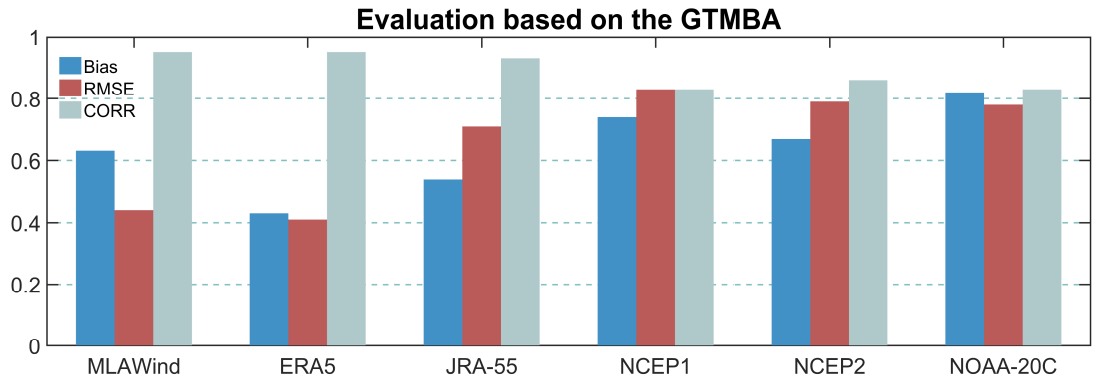


Figure 7. Evaluation of multiple datasets based on the GTMBA observations from 1993–2023.

4. Performance of MLAWind from climatology to long-term trend

The reliability of the MLAWind dataset is validated through satellite and in-situ observations in both training and non-training periods (Figs. 4-7). In this section, the performance of the MLAWind dataset since 1950 are assessed across different time scales, including climatology, seasonality, inter-annual variations, and long-term trends. Its similarities and differences with reanalysis datasets are evaluated, particularly during the period without satellite data.

4.1 Climatology and annual cycle

The climatological spatial pattern of the MLAWind dataset aligns closely with those of multiple reanalysis products, demonstrating robust consistency in capturing the tropical trade wind belt and the mid-latitude westerly belt (Fig. 8a, b). Moreover, it can reproduce the key dynamical features of atmospheric circulation. The divergent wind and velocity potential of the MLAWind dataset exhibit a well-defined convergence–divergence dipole structure of the Walker circulation: intense convergence is concentrated over the Indo-Pacific Warm Pool, while robust divergence centers are distinctly localized in the southeastern Pacific Ocean (Fig. 8c).

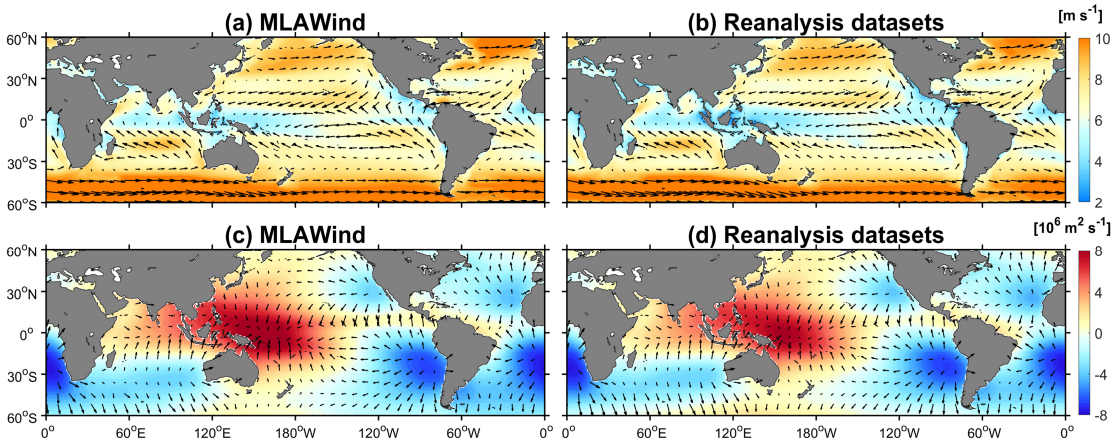


Figure 8. (a-b) Climatology of 10-m wind velocity (vector, m s^{-1}) and sea surface wind speed (shading, m s^{-1}) in the MLAWind dataset and the reanalysis datasets from 1981 to 2010. (c-d) Same as (a-b) but for divergent wind (vector, m s^{-1}) and velocity potential (shading, m^2 s^{-1}). The 30-year period is selected because all datasets are available during this period (Table 1). The climatological distribution of the MLAWind dataset over the 73-year period from 1950 to 2022 are examined and show no significant differences.

The climatological distributions between the MLAWind dataset and the reanalysis datasets are further compared across three distinct periods (i.e. 1950–1969, 1970–1992, and 1993–2022). Similar to Fig. 8, the spatial distributions of MLAWind are highly consistent with those of the reanalysis datasets, even during the pre-satellite periods (1950–1969 and 1970–1992). Both datasets capture prominent westerlies over the Kuroshio, Gulf Stream, and mid-latitude Southern Hemisphere. They reproduce the strong easterlies over the equatorial Pacific and Atlantic, together with the weak annual-mean winds over the equatorial Indian Ocean that arise from monsoon-driven seasonal reversals (Schott et al., 2009). The quantitative evaluation indicates that global-mean climatological sea surface wind speed of the MLAWind dataset exhibits low bias and RMSE relative to the reanalysis datasets, accompanied by a high correlation above 0.97 over the three examined periods (Table 3).

Table 3. Comparison of global-mean climatological sea surface wind speed between the MLAWind dataset and the reanalysis datasets across different periods.

	1950–1969	1970–1992	1993–2022
Bias (m s^{-1})	0.37	0.37	0.28
RMSE (m s^{-1})	0.38	0.32	0.23
CORR	0.97	0.98	0.99

The standard deviations of the 10-m zonal and meridional wind anomalies derived from the MLAWind dataset and the reanalysis datasets are assessed across different periods. The MLAWind dataset exhibits strong spatial consistency with the reanalysis datasets, despite differences in magnitude. Both datasets consistently indicate that high-variability regions are confined to the westerly wind belt and the Indo-Pacific Warm Pool region. These results demonstrate that the MLAWind dataset effectively captures the variability of sea surface winds, even during periods with sparse satellite data and in regions with limited observations.

The annual cycle of mean sea surface winds is evaluated in equatorial regions within 5°S–5°N (Fig. 9). The 10-m zonal winds in the MLAWind dataset and reanalysis datasets show similar seasonal variations over the global equatorial ocean, with their easterlies persisted within a range of 1.55-2.55 m s^{-1} throughout the year. A discrepancy in intensity exists between the two datasets. The 10-m zonal winds in the MLAWind dataset exhibit stronger easterlies during March–May and weaker easterlies during July–September compared to reanalysis datasets (Fig. 9a). The annual cycle characteristics of the MLAWind dataset show a general agreement with those of reanalysis datasets over the three equatorial oceans, albeit with some differences in magnitude (Fig. 9b-d). The 10-m meridional wind in both the MLAWind dataset and reanalysis datasets display a distinct unimodal structure, featuring a southerly wind peak in February and a northerly wind peak in August (Fig. 9e). The MLAWind dataset accurately captures the annual cycle of 10-m

meridional winds across different equatorial oceans, consistent with reanalysis datasets (Fig. 9f-h).

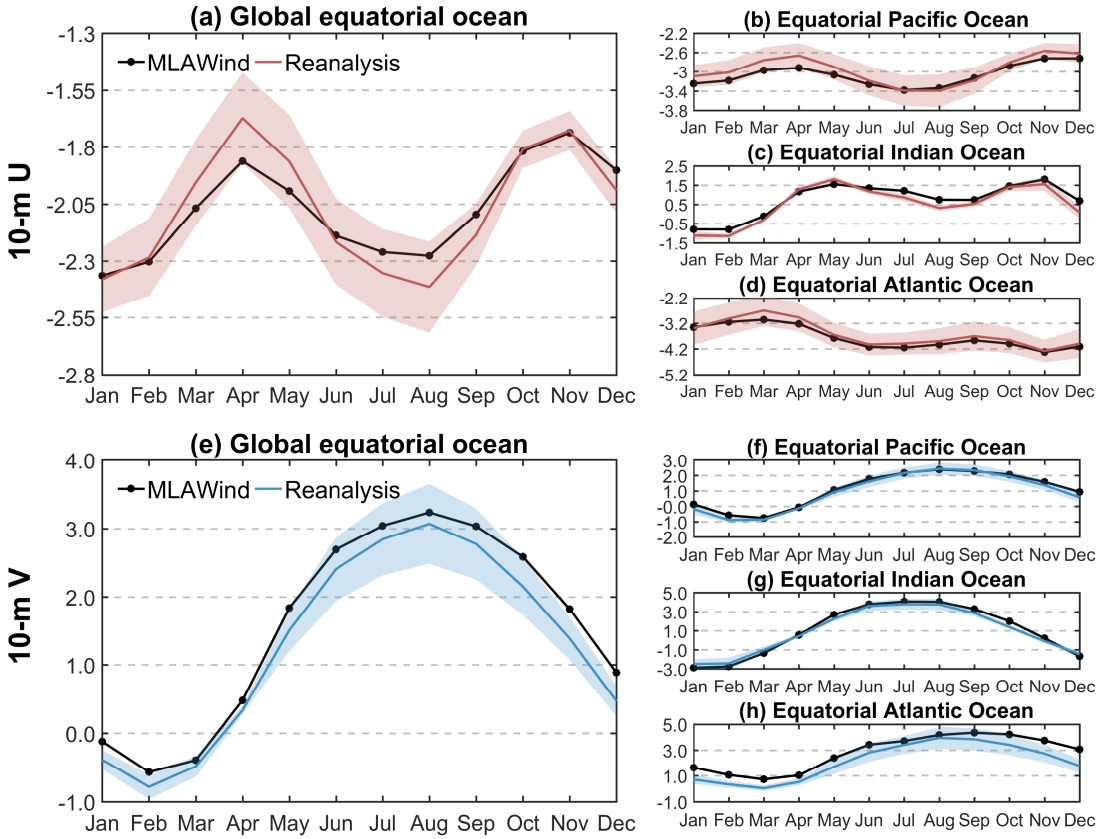


Figure 9. Annual cycle of sea surface wind (m s^{-1}) over equatorial oceans (5°S–5°N) from 1981 to 2010: (a-d) 10-m zonal wind and (e-h) 10-m meridional wind. The red shadings in (a-d) and the blue shading in (e-h) denote one standard deviations of 10-m zonal and meridional winds from multiple reanalysis datasets, respectively. The 30-year period (1981–2010) is selected because all datasets are available during this period (Table 1). The annual cycle of the MLAWind dataset over the 73-year period from 1950 to 2022 are examined and show no significant differences.

4.2 Inter-annual variabilities

ENSO is the dominant inter-annual climate variability in the tropics, exerting significant influences on both regional and global climate systems (e.g. Bjerknes, 1969; Trenberth et al., 1998; McPhaden et al., 2006). Sea surface wind anomalies are greatly associated with the ENSO cycle (Bjerknes, 1969; McPhaden et al., 2006; Kug et al., 2009). Time series of 10-m zonal wind anomalies from multiple datasets are compared over the Niño-3.4 region (Fig.

10). During the satellite observation period (1993–2022), MLAWind, ERA5, and NCEP1 show strong agreement with CCMP, among which MLAWind exhibits the highest correlation coefficient (CORR = 0.98). During the historical period lacking satellite data (1950–1992), MLAWind maintains comparable inter-annual variations to those of both ERA5 and NCEP1, while demonstrating a higher correlation coefficient with the Niño-3.4 index (CORR = -0.86). The 10-m zonal wind anomalies derived from MLAWind show a robust association with ENSO dynamics, exhibiting strong westerly wind anomalies during El Niño events and persistent easterly anomalies during La Niña events, which aligns closely with the canonical Bjerknes feedback mechanism (Bjerknes, 1969; McPhaden et al., 2006; Kug et al., 2009). Moreover, the inter-annual variabilities of MLAWind across other tropical ocean basins have been validated, equally exhibiting high consistency with reanalysis datasets and SST variabilities (Guo et al., 2025a).

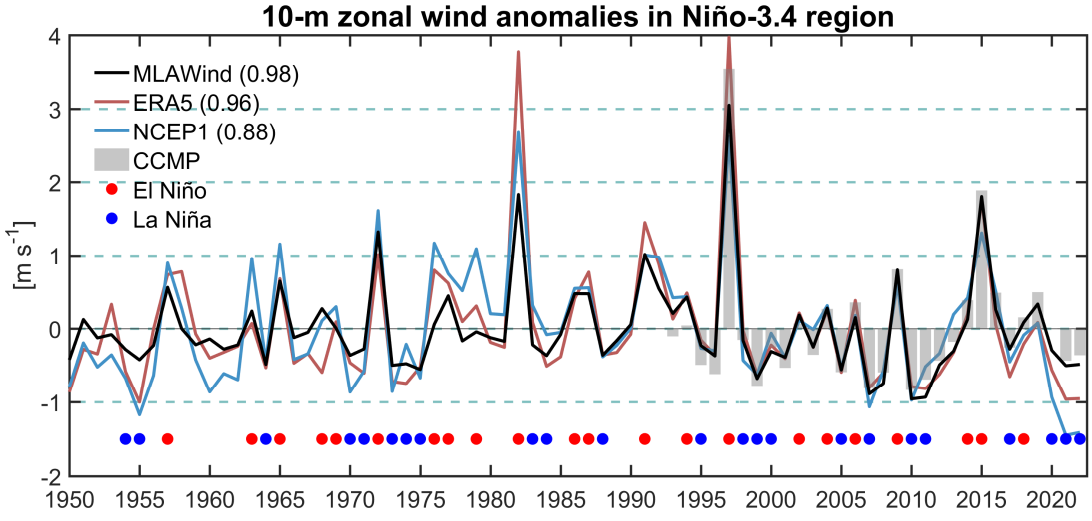


Figure 10. Inter-annual variations of the November–December mean 10-m zonal wind anomalies (m s^{-1}) over the Niño-3.4 region (5°S–5°N, 170°–120°W). Long-term linear trends from 1950–2022 are removed. The grey bars denote 10-m zonal wind anomalies of CCMP from 1993–2022. The values in parentheses denote correlation coefficients with the CCMP data. The red and blue circles denote El Niño and La Niña events, respectively. El Niño and La Niña events are defined when the November–December mean Niño-3.4 index ≥ +0.5°C and ≤ -0.5°C, respectively.

A composite analysis of El Niño events is conducted for the satellite (Fig. 11a) and non-satellite periods (Fig. 11b). The selected El Niño events are marked by red circles in Fig. 10, with a total of 13 events during 1950–1992 and 8 events during 1993–2022. During the satellite period (1993–2022), MLAWind reveal the development of westerly wind anomalies over the equatorial central-western Pacific Ocean (5°S–5°N, 160°E–150°W), peaking around DJF(1). Weak easterly wind anomalies are observed over the equatorial eastern Pacific Ocean (5°S–5°N, 150°–90°W). The pattern is consistent with those of CCMP, with the spatial correlation coefficient exceeding 0.95. ERA5 and NCEP1 exhibit pronounced westerly anomalies in the region west of 150°W, but the easterly anomalies in the equatorial eastern Pacific Ocean are weaker than those of CCMP and MLAWind (Fig. 11a).

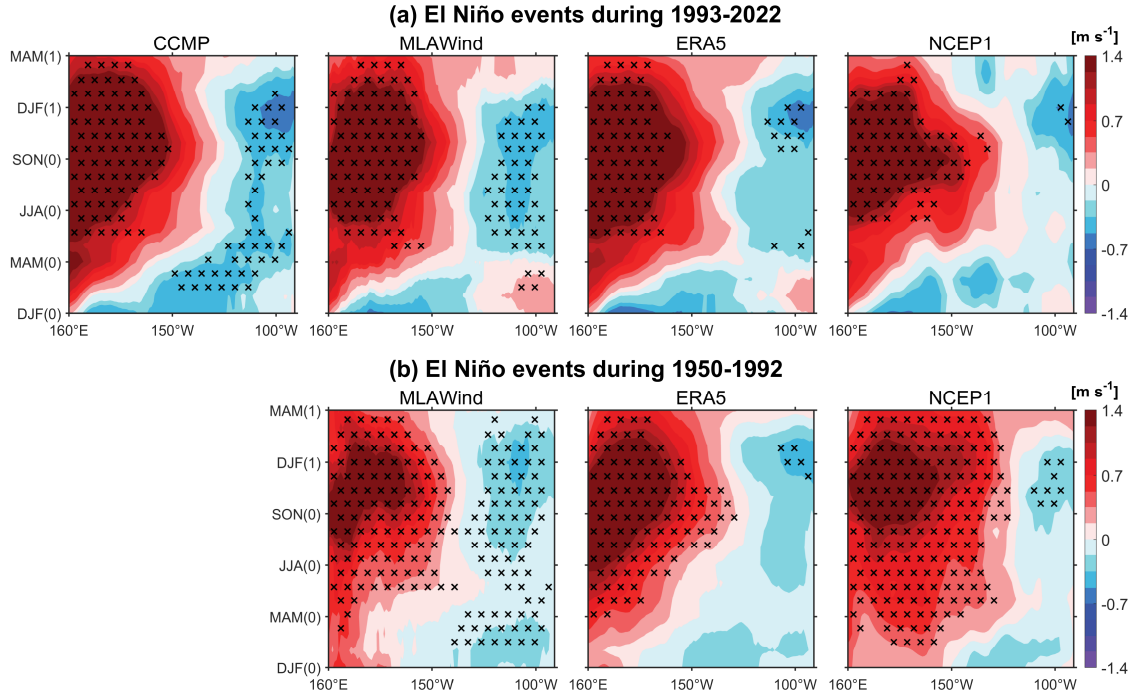


Figure 11. Composites of meridional mean (5°S–5°N) of 10-m zonal wind anomalies. (a) El Niño events during 1993–2022. (b) Same as (a) but for the period lacking satellite data (1950–1992). DJF(0)-SON(0) denote the periods in the developing year while DJF(1)-MAM(1) denote the periods in the decaying year. The black crosses denote the composites exceeding the 99% significance level based on Student's *t*-test.

During the historical period of 1950–1992, both MLAWind and reanalysis

datasets show comparable sea surface wind patterns to those observed during the satellite period. They collectively demonstrate significant westerly wind anomalies that are strongly linked to the development of El Niño events (Fig. 11b). It is noted that there are several discrepancies in the three datasets. MLAWind shows relatively stronger easterly anomalies in the eastern equatorial Pacific Ocean. NCEP1 exhibits more pronounced westerly anomalies during the DJF(0)-MAM(0) period, which extends further eastward compared to those of MLAWind and ERA5 (Fig. 11b). Although the absence of robust benchmarks hinders the assessment of their relative accuracy, the consistency with satellite data and the Niño-3.4 index (Fig. 10) suggests that MLAWind may provide more reliable sea surface wind variations.

In addition to the tropical Pacific Ocean, the inter-annual variability of the MLAWind dataset is systematically assessed on a global scale. Figure 12 presents the CORR and RMSE between MLAWind and ERA5. During 1993–2023, the 10-m zonal winds from MLAWind exhibit the strongest positive correlations with those from ERA5, which can be attributed to the incorporation of CCMP satellite data in the generation of both datasets (Hersbach et al., 2019). During 1950–1992, MLAWind maintains significant positive correlations with ERA5 over most of the globe. It indicates that MLAWind can effectively reproduce the inter-annual variability characteristics of reanalysis datasets, even in the absence of satellite data (Fig. 12).

The 10-m zonal wind of MLAWind exhibits minimal RMSE values during the satellite period. During 1950–1992, its RMSE remains modest over most of the global ocean, with relatively large values concentrated in the Southern Ocean and the eastern tropical Indian Ocean (Fig. 12). The pronounced discrepancies over the Southern Ocean can be attributed to the sparse observational coverage within the 30°S–60°S band, which introduces substantial uncertainties across all products. In addition, previous study has revealed considerable uncertainties among different datasets in the tropical Indian Ocean, where MLAWind are found to be more reliable than the existing

reanalysis datasets (Guo et al., 2025a).

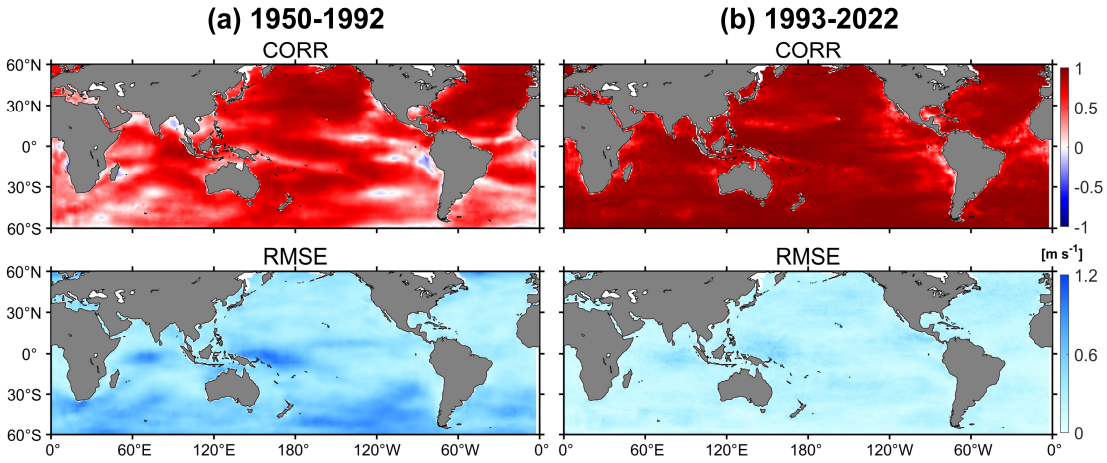
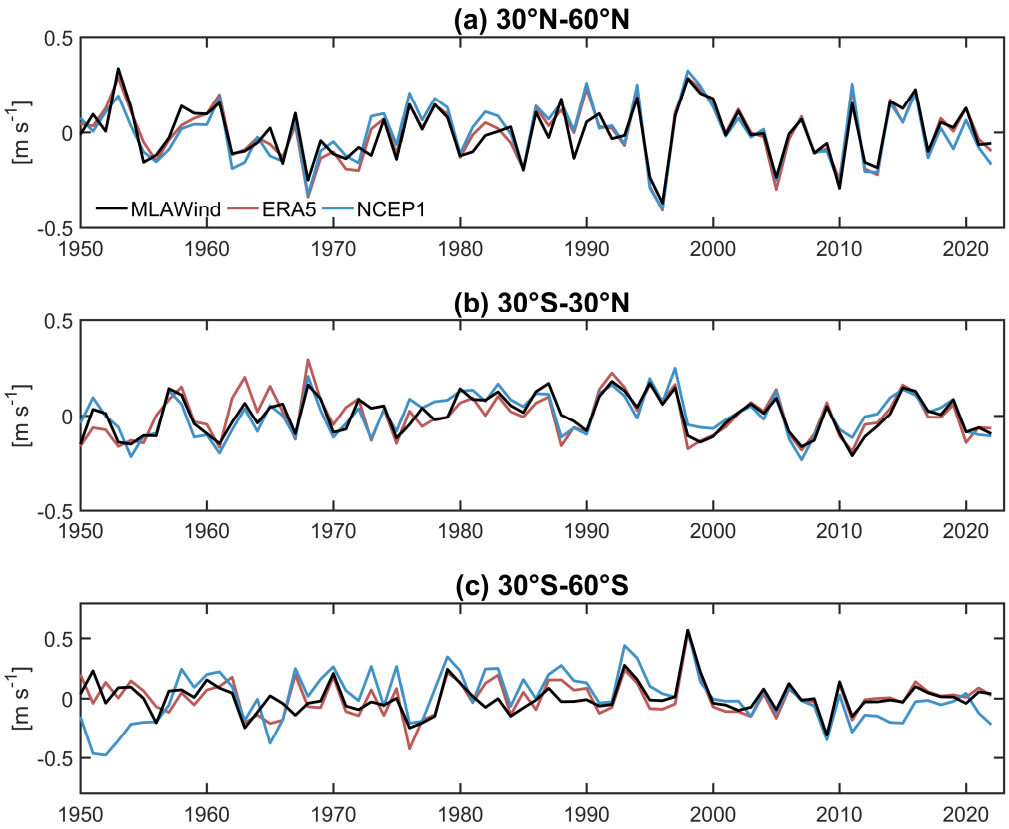


Figure 12. (a) Correlation coefficient (CORR) and Root Mean Square Error (RMSE) between MLAWind and ERA5 for 10-m zonal winds during 1950–1992. (b) Same as (a) but for 1993–2022.

To further compare the performance of the MLAWind dataset and the reanalysis datasets in different regions, their inter-annual variations are assessed across three broad latitude bands: 30°S–60°S, 30°S–30°N, and 30°N–60°N (Fig. 13). The 10-m zonal wind anomalies from MLAWind exhibit strong temporal consistency with those from reanalysis datasets across different regions, even during periods without satellite observations. Despite the relatively large RMSE between MLAWind and ERA5 over the westerly wind belt in the Southern Hemisphere (Fig. 12), MLAWind is effective in capturing the inter-annual variability of the 10-m zonal wind anomalies within the 30°S–60°S region (Fig. 13c). The result demonstrates that MLAWind maintain robust applicability during non-satellite periods and over the Southern Ocean. It can serve as a reliable data source for global wind field change studies.



552

553 **Figure 13.** Inter-annual variations of the 10-m zonal wind anomalies (m s^{-1}) across different latitude
554 bands. Long-term linear trends from 1950–2022 are removed.

555 **4.3 Long-term trends**

556 The long-term trend can filter the interference from short-term
557 fluctuations, and thus reveal the linear change over a long time span (usually
558 decades or longer periods). The long-term trends of the MLAWind dataset are
559 evaluated in both the satellite (after 1993) and non-satellite (before 1993)
560 periods to analyze its reliability across the two periods. The analysis during
561 1993–2022 reveals that the MLAWind dataset attains the most robust spatial
562 concordance with observations among all evaluated datasets (CORR = 0.95).
563 This strong agreement is due to its low bias and RMSE with satellite data (Fig.
564 4).

565 The long-term trends among different products exhibit significant
566 divergence during historical periods without satellite data (Zhang et al., 2023;
567 Guo et al., 2025a). Coral records are used to assess the performance of sea

surface wind during these periods. The skeletal Mn/Ca ratio in corals from atolls with west-facing lagoons serves as an effective proxy for identifying westerly wind anomalies (Shen et al., 1992), thereby reflecting the strength and variability of zonal winds at both inter-annual and multi-decadal time scales (Thompson et al., 2015). The employed coral Mn/Ca records are collected from three distinct sites in the tropical Pacific Ocean, including Tarawa Atoll (1.33°N, 172.97°E), Butaritari Atoll (3.07°N, 172.75°E) and Kiritimati Island (1.93°N, 157.49°W).

The coral Mn/Ca records from Tarawa Atoll (4°S–6°N, 166°–176°E) provide a valuable proxy for the period of 1950–1982, when satellite observations are unavailable. The pronounced increasing trend in Mn/Ca ratios from 1950 to 1982 suggests an intensification of the westerlies, yet it does not afford a quantitative estimate of the magnitude of sea surface wind variability. Significant uncertainties remain in the long-term trends across multiple datasets (Table 4), primarily attributed to discrepancies in their data assimilation models and technical approaches. Consistent with Mn/Ca records, the MLAWind and ERA5 datasets reveal a significant strengthening of westerly winds ($p < 0.05$), with a more pronounced trend observed in the MLAWind dataset. The trends derived from the NCEP1 and NOAA-20C datasets are weaker and statistically insignificant, demonstrating their relatively lower reliability for the period lacking satellite data. Moreover, the positive temporal correlation between MLAWind and Mn/Ca surrounding Tarawa Atoll is much stronger than those observed in ERA5, NCEP1, and NOAA-20C, indicating that MLAWind provides a more reliable and physically consistent reconstruction of long-term sea surface wind variability during non-satellite periods (Table 4). Two supplementary Mn/Ca records from Butaritari Atoll (1989–2010) and Kiritimati Island (1994–2011) are also examined. The MLAWind dataset likewise exhibits strong agreement with the long-term trends of these records.

Table 4. Long-term trends of coral Mn/Ca and 10-m zonal wind surrounding Tarawa Atoll (4°S–6°N, 166°–176°E) from 1950–1982. Positive values of coral Mn/Ca indicate an intensification of westerly winds. The temporal correlation coefficients between Mn/Ca and multiple datasets are assessed. All datasets are smoothed using a 11-year running mean filter to remove the impact of high-frequency signals.

Variable		Long-term trend	*P* value	Temporal correlation coefficient
Coral record (nmol mol^{-1}/decade)	**Mn/Ca**	**1.89**	***P*<0.05**	**-**
10-m U ($m\ s^{-1}$/decade)	**MLAWind**	**0.49**	***P*<0.05**	**0.48**
	ERA5	0.28	*P*<0.05	0.02
	NCEP1	0.14	*P*>0.05	-0.64
	NOAA-20C	0.14	*P*>0.05	-0.22

A comprehensive analysis of sea surface wind trends in the tropical Indian Ocean (TIO) reveals significant discrepancies among multiple datasets (Guo et al., 2025a). Different from reanalysis results, the MLAWind dataset exhibits a significant weakening of sea surface wind speed over the TIO during 1950–2019. The trend is corroborated by the asymmetrical west-east variations in both thermocline depth and sea surface height since 1950 (Guo et al., 2025a), demonstrating the reliability of the MLAWind dataset in capturing sea surface wind variations over long time periods.

5. Data availability

The MLAWind dataset reconstructed in this study is available at https://doi.org/10.5281/zenodo.17354864 (Guo et al., 2025b). Here, we provide a monthly gridded sea surface wind product at 1°×1° horizontal resolution from

1950 to 2023, covering the near-global ocean between 60°S and 60°N.

6. Conclusions

This study employs a random forest algorithm integrated with the SHAP interpretable module to generate a monthly sea surface wind dataset (MLAWind). The MLAWind dataset effectively combines the respective strengths of in-situ and remote sensing observations, covering the near-global ocean between 60°S and 60°N from 1950 to 2023, with a spatial resolution of 1°×1°. The evaluations based on satellite data reveal that MLAWind exhibits better agreement with observations during the training period (1993–2022) compared to most reanalysis datasets. It shows strong generalization capability, maintaining robust performance during the testing period (2023). The performance of MLAWind since 1950 is evaluated across different time scales, including climatology, annual cycle, inter-annual variation, and long-term trends. The assessments based on reanalysis datasets and coral records indicate that MLAWind demonstrates better skill in capturing long-term variations of sea surface winds compared to other datasets. MLAWind is expected to become an important resource for global climate change research.

The MLAWind dataset will continue to be optimized and updated in future work. The current version of the MLAWind dataset is limited within 60°S–60°N, owing to the insufficient data coverage of the ICOADS and CCMP datasets in high-latitude regions. More reliable high-latitude datasets will be integrated into the machine learning model in the future, enabling the development of a global sea surface wind dataset. In addition, the MLAWind dataset, currently available at a horizontal resolution of 1°×1°, demonstrates robust capability in capturing large-scale circulation signals. Further enhancement of its spatial resolution to 0.25°×0.25° will significantly advance mesoscale marine and climate research. Several potential approaches can be used to construct a higher-resolution version of the MLAWind dataset, including the integration of higher-resolution input variables (Tian et al., 2022),

the application of machine learning-based interpolation (He et al., 2022) and super-resolution techniques (Wang et al., 2021; Doury et al., 2023). An assessment of diverse approaches will be conducted in future work to ensure the upcoming high-resolution MLAWind dataset retains the existing version's accuracy while capturing mesoscale atmospheric processes.

Author contributions

WG developed the data reconstruction model, generated the MLAWind dataset and drafted the initial manuscript. RZ participated in data analysis and revised the manuscript. XW conceptualized this study and contributed to the interpretation of the results. DW provided guidance on improving the manuscript. All authors provided valuable feedback and helped shape this study.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors would like to express their gratitude for the publicly available data resources that are utilized in this study. ICOADS is obtained from https://icoads.noaa.gov/products.html. GTMBA is obtained from https://www.pmel.noaa.gov/gtmba/. CCMP is obtained from https://data.remss.com/ccmp/. ERA5 is obtained from https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels-monthly-means?tab=overview. NCEP1 is obtained from https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html. NCEP2 is obtained from https://psl.noaa.gov/data/gridded/data.ncep.reanalysis2.html. JRA-55 is obtained from https://climatedataguide.ucar.edu/climate-data/jra-55. NOAA-20C is obtained from https://psl.noaa.gov/data/gridded/data.20thC_ReanV2c.html. Coral records is obtained from https://www.ncei.noaa.gov/products/paleoclimatology/coral-sclerosponge.

Financial support

This work is supported by the National Natural Science Foundation of China (Grant No. 42330404), the National Key R&D Program of China (2022YFF0801400), the National Natural Science Foundation of China (42376027, 42406197), and the Special Fund of South China Sea Institute of Oceanology of the Chinese Academy of Sciences (SCSIO2023QY01).

References

Bagnell, A. and DeVries, T.: 20(th) century cooling of the deep ocean contributed to delayed acceleration of Earth's energy imbalance, Nat. Commun., 12, 4604, https://doi.org/10.1038/s41467-021-24472-3, 2021.

Berry, D. I. and Kent, E. C.: Air–Sea fluxes from ICOADS: the construction of a new gridded dataset with uncertainty estimates, Int. J. Clim., 31, 987–1001, https://doi.org/10.1002/joc.2059, 2011.

Bjerknes, J.: Atmospheric teleconnections from the equatorial Pacific, Mon. Weather. Rev., 97, 163–172, https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2, 1969.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Clarke, A. J.: El Niño physics and El Niño predictability, Ann. Rev. Mar. Sci., 6, 79–99, https://doi.org/10.1146/annurev-marine-010213-135026, 2014.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J.: The Twentieth Century Reanalysis Project, Q. J. R. Meteorol. Soc., 137, 1–28, https://doi.org/10.1002/qj.776, 2011.

Deser, C., Alexander, M. A., and Timlin, M. S.: Evidence for a wind-driven intensification of the Kuroshio Current extension from the 1970s to the 1980s, J. Clim., 12, 1697–1706, https://doi.org/10.1175/1520-0442(1999)012<1697:EFAWDI>2.0.CO;2, 1999.

Doury, A., Somot, S., Gadat, S., Ribes, A., and Corre, L.: Regional climate model emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling approach, Clim. Dyn., 60, 1751–1779, https://doi.org/10.1007/s00382-022-06343-9, 2023.

Findell, K. L., Keys, P. W., Ent R. J. V. D., Lintner, B. R., Berg, A., and Krasting, J. P.: Rising temperatures increase importance of oceanic evaporation as a source for continental precipitation, J. Clim., 32, 7713–7726, https://doi.org/10.1175/JCLI-D-19-0145.1, 2019.

Freeman, E., Woodruff, S. D., Worley, S. J., Lubker, S. J., Kent, E. C., Lubker, W. E., Berry, D. I., Brohan, P., Eastman, R., Gates, L., Gloeden, W., Ji, Z., Lawrimore, J., Rayner, N. A., Rosenhagen, G., and Smith, S. R.: ICOADS Release 3.0: a major update to the historical marine climate record, Int. J. Clim., 37, 2211–2232, https://doi.org/10.1002/joc.4775, 2017.

Garreaud, R., Lopez, P., Minvielle, M., and Rojas, M.: Large-scale cont rol on the Patagonian climate, J. Clim., 26, 215–230, https://doi.or g/10.1175/JCLI-D-12-00001.1, 2013.

Guo, W., Zhang, R., Wang, X., Wang, C., Li, X., Han, W., and Zhang, L.: Unveiling the drivers of tropical Indian Ocean warming through machine learning-assisted surface wind, J. Clim., 38, 6763–6779, https://doi.org/10.1175/JCLI-D-25-0003.1, 2025a.

Guo, W., Zhang, R., and Wang, X.: Machine learning-assisted sea surface wind dataset (MLAWind), Zenodo [data set], https://doi.org/10.5281/zenodo.17354864, 2025b.

He, Q., Wang, M., Liu, K., Li, K., and Jiang, Z.: GPRChinaTemp1km: a high-resolution monthly air temperature dataset for China (1951–2020) based on machine learning, Earth Syst. Sci. Data., 14, 3273–3292, https://doi.org/10.5194/essd-2021-442, 2022.

Held, I. M. and Soden, B. J.: Robust responses of the hydrological cycle to global warming, J. Clim., 19, 5686–5699, https://doi.org/10.1175/JCLI3990.1, 2006.

Hersbach, H., Bell, B., Berrisford, P., Horányi, A., Muñoz Sabater, J., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C., and Dee, D.: Global reanalysis: goodbye ERA-Interim, hello ERA5, Meteorology section of

ECMWF Newsletter, 159, 17–24, https://doi.org/10.21957/vf291hehd7, 2019.

Huang, B., Yin, X., Boyer, T., Liu, C., Menne, M., Rao, Y. D., Smith, T., Vose, R., and Zhang, H.-M.: Extended Reconstructed Sea Surface Temperature Version 6 (ERSSTv6): Part I. An Artificial Neural Network Approach, J. Clim., 38, 1105–1121, https://doi.org/10.1175/JCLI-D-23-0707.1, 2025.

IPCC: Climate Change 2021: The Physical Science Basis, in: Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., Cambridge University Press, 2391 pp, https://doi.org/10.1017/9781009157896, 2023.

Jiang, S., Sweet, L.-b., Blougouras, G., Brenning, A., Li, W., Reichstein, M., Denzler, J., Shangguan, W. Yu, G., Huang, F., and Zscheischler, J.: How interpretable machine learning can benefit process understanding in the geosciences, Earth's Future, 12, e2024EF004540, https://doi.org/10.1029/2024EF004540, 2024.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D.: The NCEP/NCAR 40-Year Reanalysis Project, Bull. Am. Meteorol. Soc., 77, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, 1996.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J. J., Fiorino, M., and Potter, G. L.: NCEP–DOE AMIP-II Reanalysis (R-2), Bull. Amer.

761 Meteor. Soc., 83, 1631–1643, https://doi.org/10.1175/BAMS-83-11-1631,
762 2002.
763 Kim, S., Alizamir, M., Heddam, S., Chang, S. W., Chung, I. M., Kisi, O., and
764 Kulls, C.: Development of the machine learning and deep learning models
765 with SHAP strategy for predicting groundwater levels in South Korea, Sci.
766 Rep., 15, 35523, https://doi.org/ 10.1038/s41598-025-19545-y, 2025.
767 Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi,
768 K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi,
769 K.: The JRA-55 Reanalysis: General Specifications and Basic
770 Characteristics, J. Meteorol. Soc. Jpn. Ser. II, 93, 5–48.
771 https://doi.org/10.2151/jmsj.2015-001, 2015.
772 Kug, J.-S., Jin, F.-F., and An, S.-I.: Two types of El Niño events: Cold tongue
773 El Niño and warm pool El Niño, J. Clim., 22, 1499–1515,
774 https://doi.org/10.1175/2008JCLI2624.1, 2009.
775 Leng, J., Chen, J. M., Li, W., Luo, X., Xu, M., Liu, J., Wang, R. Rogers, C. Li,
776 B., and Yan, Y.: Global datasets of hourly carbon and water fluxes
777 simulated using a satellite-based process model with dynamic
778 parameterizations, Earth Syst. Sci. Data, 16, 1283–1300,
779 https://doi.org/10.5194/essd-16-1283-2024, 2024.
780 Li, R., Li, Y., Lyu, Y., Sprintall, J., and Wang, F.: Role of the Indian Ocean
781 Wind-Driven Dynamics in the Indonesian Throughflow Variability, J.
782 Geophys. Res.: Oceans, 130, e2025JC022503,
783 https://doi.org/10.1029/2025JC022503, 2025.
784 Lundberg, S. and Lee, S.-I.: A unified approach to interpreting model
785 predictions, Adv. Neural Inf. Process. Syst., 30, 1–10,
786 https://doi.org/10.48550/arXiv.1705.07874, 2017.
787 Mao, W., Mu, X., Zheng, Y., and Yan, G.: Leave-one-out cross-validation-based
788 model selection for multi-input multi-output support vector machine,

Neural Comput. App., 24, 441–451, https://doi.org/10.1007/s00521-012-1234-5, 2014.

McPhaden, M. J., Zebiak, S. E., and Glantz, M. H.: ENSO as an integrating concept in Earth science, Science, 314, 1740–1745, https://doi.org/10.1126/science.1132588, 2006.

McPhaden, M. J., Ando, K., Bourle`s, B., Freitag, H. P., Lumpkin, R., Masumoto, Y., Murty, V., Nobre, P., Ravichandran, M., Vialard, J., Vousden, D., and Yu, W.: The global tropical moored buoy array. In: Hall, J., Harrison, D. E., Stammer, D. (eds) Proceedings of the ‘‘OceanObs’09: Sustained Ocean Observations and Information for Society’’ conference, vol 2, Venice, Italy, 21–25 Sep 2009, ESA Publication WPP-306, https://doi.org/10.5270/OceanObs09.cwp.61, 2010.

Mears, C. A., Scott, J., Wentz, F. J., Ricciardulli, L., Leidner, S. M., Hoffman, R., and Atlas, R.: A near-real-time version of the cross-calibrated multiplatform (CCMP) ocean surface wind velocity data set, J. Geophys. Res.: Oceans, 124, 6997–7010, https://doi.org/10.1029/2019JC015367, 2019.

Qin, L., Zhu, L., Liu, B., Li, Z., Tian, Y., Mitchell, G., Shen, S., Xu, W., and Chen, J.: Global expansion of tropical cyclone precipitation footprint, Nat. Commun., 15, https://doi.org/10.1038/s41467-024-49115-1, 2024.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Ren, H. L., Zuo, J., and Deng, Y.: Statistical predictability of Niño indices for two types of ENSO, Clim. Dyn., 52, 5361–5382. https://doi.org/10.1007/s00382-018-4453-3, 2019.

Sayani, H. R., Thompson, D. M., Carilli, J. E., Marchitto, T. M., Chapman, A. U., and Cobb, K. M.: Reproducibility of coral Mn/Ca-based wind

reconstructions at Kiritimati Island and Butaritari Atoll, Geochem. Geophy. Geosy., 22, e2020GC009398, https://doi.org/10.1029/2020GC009398, 2021.

Schott, F. A., Xie, S.-P., and McCreary, J. P.: Indian ocean circulation and climate variability, Rev. Geophys., 47, RG1002, https://doi.org/10.1029/2007RG000245, 2009.

Shen, G. T., Linn, L. J., Campbell, T. M., Cole, J. E., and Fairbanks, R. G.: A chemical indicator of trade wind reversal in corals from the western tropical Pacific, J. Geophys. Res.: Oceans., 97, 12689–12697, https://doi.org/10.1029/92JC00951, 1992.

Smith, S. D.: Wind stress and heat flux over the ocean in gale force winds, J. Phys. Oceanogr., 10, 709–726, https://doi.org/10.1175/1520-0485(1980)010<0709:WSAHFO>2.0.CO;2, 1980.

Su, H., Li, W. and Yan, X.-H.: Retrieving temperature anomaly in the global subsurface and deeper ocean from satellite observations, J. Geophys. Res.: Oceans., 123, 399–410, https://doi.org/10.1002/2017JC013631, 2018.

Su, H., Zhang, H., Geng, X., Qin, T., Lu, W., and Yan, X. H.: OPEN: A new estimation of global ocean heat content for upper 2000 meters from remote sensing data, Remote Sens., 12, 2294, https://doi.org/10.3390/rs12142294, 2020.

Thompson, D. M., Cole, J. E., Shen, G. T., Tudhope, A. W., and Meehl, G. A.: Early twentieth-century warming linked to tropical Pacific wind strength, Nat. Geosci., 8, 117–121, https://doi.org/10.1038/ngeo2321, 2015.

Tian, T. Cheng, L., Wang, G., Abraham, J., Ren, S., Zhu, J., Song, J., and Leng, H.: Reconstructing ocean subsurface salinity at high resolution using a machine learning approach, Earth Syst. Sci. Data, 14, 5037–5060, https://doi.org/10.5194/essd-2022-236, 2022.

Tokinaga, H. and Xie, S.-P.: Wave- and Anemometer-Based Sea Surface Wind (WASWind) for Climate Change Analysis*, J. Clim., 24, 267–285, https://doi.org/10.1175/2010JCLI3789.1, 2011.

Tokinaga, H., Xie, S.-P., Timmermann, A., McGregor, S., Ogata, T., Kubota, H., and Okumura, Y. M.: Regional Patterns of Tropical Indo-Pacific Climate Change: Evidence of the Walker Circulation Weakening*, J. Clim., 25, 1689–1710, https://doi.org/10.1175/JCLI-D-11-00263.1, 2012.

Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N. C., and Ropelewski, C.: Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures, J. Geophys. Res.: Oceans., 103, 14291–14324, https://doi.org/10.1029/97jc01444, 1998.

Tseng, Y. H., Huang, J. H., and Chen, H. C.: Improving the Predictability of Two Types of ENSO by the Characteristics of Extratropical Precursors, Geophys. Res. Lett., 49, e2021GL097190, https://doi.org/10.1029/2021GL097190, 2022.

Vinoth, J. and Young, I. R.: Global Estimates of Extreme Wind Speed and Wave Height, J. Clim., 24, 1647–1665, https://doi.org/10.1175/2010JCLI3680.1, 2011.

Wang, C.: Three-ocean interactions and climate variability: A review and perspective, Clim. Dyn., 53, 18–5136, https://doi.org/10.1007/s00382-019-04930-x, 2019.

Wang, D., Liu, Q., Huang, R. X., Du, Y., and Qu T.: Interannual variability of the South China Sea throughflow inferred from wind data and an ocean data assimilation product, Geophys. Res. Lett., 33, L14605, https://doi.org/10.1029/2006GL026316, 2006.

Wang, H. and Li, X.: DeepBlue: Advanced convolutional neural network applications for ocean remote sensing, IEEE Geoscience and Remote

874 Sensing Magazine, 12, 138–161,
875 https://doi.org/10.1109/MGRS.2023.3343623, 2024.
876 Wang, J., Liu, Z., Foster, I., Chang, W., Kettimuthu, R., and Kotamarthi, V. R.:
877 Fast and accurate learned multiresolution dynamical downscaling for
878 precipitation, Geosci. Model. Dev., 14, 6355–6372,
879 https://doi.org/10.5194/gmd-14-6355-2021, 2021.
880 Wang, Q., Zeng, L., Chen, J., He, Y., Liu, Q., Sui, D., and Wang, D.: Phase shift
881 of the winter South China Sea western boundary current over the past two
882 decades and its drivers, Geophys. Res. Lett., 50, e2023GL103145,
883 https://doi.org/10.1029/2023GL103145, 2023.
884 Wang, X., Dickinson, R. E., Su, L., Zhou, C., and Wang, K.: PM2.5 pollution
885 in China and how it has been exacerbated by terrain and meteorological
886 conditions, Bull. Am. Meteorol. Soc., 99, 105–119,
887 https://doi.org/10.1175/bams-d-16-0301.1, 2018.
888 Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., and Bretherton, C.
889 S.: Correcting weather and climate models by machine learning nudged
890 historical simulations, Geophys. Res. Lett., 48, e2021GL092555,
891 https://doi.org/10.1029/2021GL092555, 2021.
892 Yang, L., Wang, D., Huang, J., Wang, X., Zeng, L., Shi, R., He, Y., Xie, Q.,
893 Wang, S., Chen, R., Yuan, J., Wang, Q., Chen, J., Zu, T., Li, J., Sui, D.,
894 and Peng, S.: Toward a Mesoscale Hydrological and Marine
895 Meteorological Observation Network in the South China Sea, Bull. Amer.
896 Meteor. Soc., 96, 1117–1135, https://doi.org/10.1175/BAMS-D-14-
897 00159.1, 2015.
898 Yin, J., Slater, L. J., Khouakhi, A., Yu, L., Liu, P., Li, F., Pokhrel, Y., and
899 Gentine, P.: GTWS-MLrec: global terrestrial water storage reconstruction
900 by machine learning from 1940 to present, Earth Syst. Sci. Data, 15, 5597–
901 5615, https://doi.org/10.5194/essd-15-5597-2023, 2023.

Yoo, C., Im, J., Park, S., and Quackenbush, L. J.: Estimation of daily maximum and minimum air temperatures in urban landscapes using MODIS time series satellite data, ISPRS J. Photogramm. Remote Sens, 137, 149–162, https://doi.org/10.1016/j.isprsjprs.2018.01.018, 2018.

Young, I. and Ribal, A.: Multiplatform evaluation of global trends in wind speed and wave height, Science, 364, 548, https://doi.org/10.1126/science.aav9527, 2019.

Zhang, H., Chen, D., Liu, T., Tian, D., He, M., Li, Q., Wei, G., and Liu, J.: MASCS 1.0: synchronous atmospheric and oceanic data from a cross-shaped moored array in the northern South China Sea during 2014–2015, Earth Syst. Sci. Data, 16, 5665–5679, https://doi.org/10.5194/essd-16-5665-2024, 2024.

Zhang, R., Guo, W., and Wang, X.: Deep Learning to Near-Surface Humidity Retrieval from Multi-Sensor Remote Sensing Data over the China Seas, Remote Sens., 14, 4353, https://doi.org/10.3390/rs14174353, 2022.

Zhang, R., Guo, W., Wang, X., and Wang, C.: Ambiguous Variations in Tropical Latent Heat Flux since the Years around 1998, J. Clim., 36, 3403–3415, https://doi.org/10.1175/JCLI-D-22-0381.1, 2023.

Zhang, X., Nikurashin, M., Pena-Molino, B., Rintoul, S. R., and Doddridge, E.: A Theory of Standing Meanders of the Antarctic Circumpolar Current and Their Response to Wind, J Phys. Oceanogr., 53, 235–251, https://doi.org/10.1175/JPO-D-22-0086.1, 2023.

Zhou, C., Azorin-Molina, C., Engstrm, E., Minola, L., Wern, L., Hellstrm, S., Lnn, J., and Chen, D.: HomogWS-se: a century-long homogenized dataset of near-surface wind speed observations since 1925 rescued in Sweden, Earth Syst. Sci. Data, 14, 2167–2177, https://doi.org/10.5194/essd-14-2167-2022, 2022.