TECHNICAL REPORT **Report No.: ESSO-INCOIS-ODICT-TR-02(2026)**

Advancing Oceanic Data Continuity: A Machine Learning Framework for Reconstructing Shortwave Radiation in Observation Networks

by

Pavan Kumar Jonnakuti, M Srujana, TVS Udaya Bhaskar, E Pattabhi Rama Rao

Indian National Centre for Ocean Information Services (INCOIS)

(Earth System Science Organization (ESSO), Ministry of Earth Sciences (MoES)

HYDERABAD, INDIA

www.incois.gov.in

March 09, 2026

DOCUMENT CONTROL SHEET

Earth System Science Organization (ESSO)

Ministry of Earth Sciences (MoES)

Indian National Centre for Ocean Information Services (INCOIS)

ESSO Document Number: ESSO-INCOIS-ODICT-TR-02(2026)

Title of the report:

"Advancing Oceanic Data Continuity: A Machine Learning Framework for Reconstructing Shortwave Radiation in Observation Networks".

Author(s) [Last name, First name]:
Jonnakuti Pavan Kumar, M Srujana, TVS Udaya Bhaskar, & E Pattabhi Rama Rao

Originating unit: Ocean Data Management Group (ODM-ICTD), INCOIS
Type of Document: Technical Report (TR)
Number of pages and figures: 30, **15**
Number of references: 17
Keywords: Shortwave Radiation (SWR), Gap-filling, Moored Buoy, Machine Learning, Satellite Observations.
Security classification: Open
Distribution: Open
Date of publication: March 2026

Abstract (236 words):

Accurate estimation of surface shortwave radiation (SWR) is essential for understanding the ocean–atmosphere energy balance and for improving weather, climate, and ocean circulation models. In-situ observations from moored buoys provide high-frequency measurements of meteorological and oceanographic variables, but these datasets often contain missing values caused by communication failures, sensor malfunctions, harsh environmental conditions, and maintenance interruptions. These data gaps limit the usability of buoy observations for scientific analysis and operational applications. This study presents a machine learning–based framework to reconstruct missing shortwave radiation observations using meteorological parameters obtained from buoy measurements, satellite products from the International Satellite Cloud Climatology Project, and radiation fields from reanalysis datasets. The analysis focuses on the BD10 moored buoy observations between **1 January 2013 and 31 July 2017**, where SWR measurements were recorded at **3-hour intervals**. Multiple machine learning models—including Random Forest, Support Vector Machines, and Artificial Neural Networks—were trained using meteorological variables such as air temperature, humidity, wind speed, and cloud cover. Model performance was evaluated using statistical metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and correlation coefficient. The results demonstrate that the Random Forest model consistently outperforms other approaches, achieving the lowest prediction error and highest correlation with observed SWR values. The reconstructed dataset significantly improves temporal continuity and reliability of the SWR record. The gap-filled dataset can support improved air–sea flux estimation, numerical weather prediction, and climate modelling applications.

Advancing Oceanic Data Continuity: A Machine Learning Framework for Reconstructing Shortwave Radiation in Observation Networks

Pavan Kumar Jonnakuti[1], M Srujana[1], TVS Udaya Bhaskar[1], & E Pattabhi Rama Rao[1]

[1]Indian National Centre for Ocean Information Services

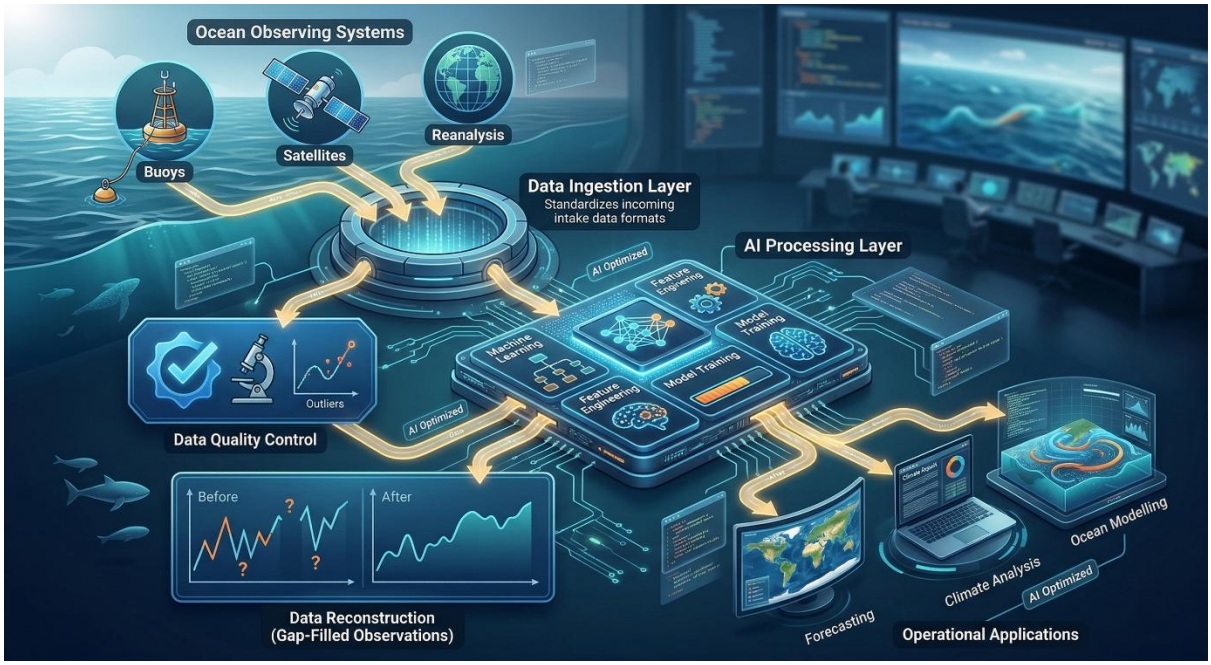
ABSTRACT

Accurate estimation of surface shortwave radiation (SWR) is essential for understanding the ocean–atmosphere energy balance and for improving weather, climate, and ocean circulation models. In-situ observations from moored buoys provide high-frequency measurements of meteorological and oceanographic variables, but these datasets often contain missing values caused by communication failures, sensor malfunctions, harsh environmental conditions, and maintenance interruptions. These data gaps limit the usability of buoy observations for scientific analysis and operational applications.

This study presents a machine learning–based framework to reconstruct missing shortwave radiation observations using meteorological parameters obtained from buoy measurements, satellite products from the International Satellite Cloud Climatology Project, and radiation fields from reanalysis datasets. The analysis focuses on the BD10 moored buoy observations between 1 January 2013 and 31 July 2017, where SWR measurements were recorded at 3-hour intervals. Multiple machine learning models—including Random Forest, Support Vector Machines, and Artificial Neural Networks—were trained using meteorological variables such as air temperature, humidity, wind speed, and cloud cover. Model performance was evaluated using statistical metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and correlation coefficient.

The results demonstrate that the Random Forest model consistently outperforms other approaches, achieving the lowest prediction error and highest correlation with observed SWR values. The reconstructed dataset significantly improves temporal continuity and reliability of the SWR record. The gap-filled dataset can support improved air–sea flux estimation, numerical weather prediction, and climate modelling applications.

Keywords: Shortwave Radiation (SWR), Gap-filling, Moored Buoy, Machine Learning, Satellite Observations.



Visual Abstract

1.0 INTRODUCTION

The ocean, covering approximately 71% of the Earth's surface, plays a fundamental role in regulating the global climate and weather systems. Acting as Earth's largest heat reservoir, it absorbs solar radiation and redistributes this energy across the planet through ocean currents. This thermal exchange between the ocean and atmosphere influences major weather phenomena, including El Niño, La Niña, monsoons, and cyclones. Ocean currents, like the Gulf Stream and thermohaline circulation, transport warm water from the equator toward the poles and return cooler water to the tropics, balancing global temperatures. Sea surface temperature (SST) significantly impacts atmospheric pressure and wind patterns, which in turn affect the information and intensity of tropical storms. Warmer ocean regions tend to intensify cyclones, while cooler areas suppress them. The ocean also plays a critical role as a carbon sink, absorbing nearly a quarter of the CO_2 emitted into the atmosphere, which helps buffer the effects of climate change. However, with increasing greenhouse gas emissions, the ocean is warming **(Gerrit et al., 2014)**, leading to the melting of polar ice, thermal expansion, sea-level rise, and ocean acidification. These physical changes are disrupting marine ecosystems by affecting coral reefs, fish migration, and plankton populations. The consequences are far-reaching, impacting biodiversity and threatening the livelihoods of coastal communities dependent on marine resources. Ocean-atmosphere interactions directly affect weather events such as rainfall, cyclones, and temperature fluctuations **(Showman et al., 2014; Philander et al., 1999)**, especially in coastal areas where small changes in ocean temperature can have pronounced effects. Furthermore, oceans help stabilize daily and

seasonal temperature variations, making coastal climates generally milder than inland areas. Modern weather prediction models rely heavily on ocean data obtained from satellites and buoys, which track sea surface temperature, salinity, currents, and wave patterns. Without these observations, forecasting would be far less accurate. In essence, the ocean is not just a passive body of water but a dynamic force essential to Earth's climate system.

India is located in South Asia, surrounded by three major water bodies: the Arabian Sea to the west, the Bay of Bengal to the east, and the Indian Ocean to the south. These oceanic regions are vital not only for biodiversity and maritime trade but also for regulating India's complex weather systems, especially the monsoon. The southwest monsoon, which delivers the majority of India's annual rainfall, is driven by differential heating between the landmass and the surrounding oceans. Among the three, the Bay of Bengal plays a particularly crucial role due to its consistently warm waters and high cyclone activity. This region significantly affects rainfall distribution, storm formation, and the intensity of weather events along India's eastern coast. To monitor and understand these oceanic and atmospheric processes, India has established an extensive network of observation systems, including moored buoys. One such buoy is BD10, located in the Bay of Bengal, which forms part of the Indian National Data Buoy Program. This buoy collects essential oceanographic and meteorological data, such as sea surface temperature, wind speed and direction, humidity, atmospheric pressure, and solar radiation. The real-time data from BD10 is transmitted via satellite to the Indian National Centre for Ocean Information Services (INCOIS), a government agency responsible for processing, analyzing, and disseminating marine information. This data is crucial for accurate weather forecasts, early cyclone warnings, climate monitoring, and disaster preparedness. It also supports research in ocean-atmosphere interaction and long-term climate modelling. The data supports various sectors, such as shipping, fishing, offshore industries, and environmental monitoring. With ocean warming and climate change becoming more evident, continuous ocean monitoring through buoys like BD10, OMNI, and African-Asian-Australian Monsoon Analysis and Prediction (RAMA) is more important than ever. They help in understanding changing ocean dynamics and their direct influence on India's climate, aiding in developing better predictive models and response strategies.

The National Buoy Program (NBP) is a global initiative aimed at monitoring oceanic and atmospheric conditions through a network of moored and drifting buoys. These buoys collect critical data on sea surface temperature, wind speed, wave height, ocean currents, and radiation fluxes, supporting weather forecasting, climate research, and marine navigation. The program plays a vital role in understanding ocean-atmosphere interactions and improving climate models **(Hamilton et al., 1986).** However, due to sensor failures, transmission issues, and periodic

maintenance, buoys often experience data gaps, hindering accurate scientific analysis and forecasting.



Fig 1. Deployed buoy as part of the National Buoy Program

Buoys are floating platforms used to monitor oceanographic and meteorological conditions in real time. They collect crucial data such as SST, wind speed, humidity, atmospheric pressure, and radiation, aiding climate research and forecasting. Moored buoys, fixed to the seabed, provide continuous observations from specific locations. The Bay of Bengal, a key climatic region in the Indian Ocean, heavily influences monsoons and cyclones **(Venkatesan et al., 2013; Hamilton et al., 1986)**. The OMNI buoy network, along with RAMA and BD10, operates in this area to study air-sea interactions. BD10 measures SST, shortwave radiation (SWR), wind, and pressure. These data help improve weather predictions and ocean models. However, data gaps can arise due to sensor failures or severe weather. Such interruptions may reduce the accuracy of climate analysis and forecasting systems.

To address missing data in ocean observations, datasets from ISCCP, RAMA buoys, and reanalysis products are used for validation and reconstruction. Machine learning techniques are applied to fill these gaps and ensure the continuity and reliability of climate records. In this study, Cloud Amount and SWR values are estimated using advanced machine learning models trained on multi-source datasets. The ISCCP provides satellite-based measurements of cloud cover and solar radiation, while RAMA buoys, including BD10, offer oceanic and atmospheric data. Moored buoys

like BD10 are anchored platforms designed to continuously measure parameters such as SST, wind speed, humidity, atmospheric pressure, ocean currents, and SWR.

These high-resolution, time-series measurements are critical for studying ocean-atmosphere interactions, climate variability, and extreme weather events like cyclones and monsoons. Unlike satellite or reanalysis datasets, buoy data provides direct, high-precision observations with minimal interpolation. In this study, the BD10 buoy is used as a key source of observed SWR data. This SWR data is essential for training machine learning models to analyze their relationship with various meteorological variables and to accurately predict missing values. Additionally, supplementary data from the RAMA network further supports this process. By integrating buoy-based SWR measurements, the study enhances the accuracy of radiation forecasting and contributes to improved oceanographic and climate assessments.

The OMNI and RAMA buoy networks in the Bay of Bengal are essential for climate monitoring, providing real-time data on ocean-atmosphere interactions **(Venkatesan et al., 2013)**. These buoys use surface and subsurface sensors to measure wind speed, atmospheric pressure, SST **(Vishnu et al., 2025)**, salinity, and other parameters. Data is transmitted via telemetry to the processing centre for forecasting and analysis, improving cyclone predictions and weather forecasting. During extreme weather events, such as cyclones, the buoys offer critical information on storm intensity, progression, and environmental changes. However, disruptions like sensor malfunctions or communication failures can lead to data gaps, affecting the accuracy of climate models and forecasts. In such cases, satellite data and reanalysis datasets help fill these gaps, but they lack the precision of in situ measurements. Proper buoy maintenance is necessary to ensure continuous data collection. Combining buoy data with satellite observations and reanalysis products strengthens climate models, improving weather predictions and disaster management strategies in the Bay of Bengal region. Accurate, uninterrupted data flow is crucial for understanding oceanic and atmospheric dynamics and enhancing climate risk management efforts. However, data gaps can occur due to factors like sensor malfunctions, power outages, communication issues, or extreme weather. These gaps pose challenges, compromising the accuracy of ocean-atmosphere studies and climate models. Missing data can significantly impact weather forecasting and disaster management efforts, making continuous, high-quality data collection essential for reliable climate analysis. By combining buoy data with satellite and reanalysis datasets, we can deepen our understanding of ocean-atmosphere interactions. Regular maintenance of buoys is crucial to minimize disruptions, ensure the accuracy of data, and enhance prediction capabilities. This integrated approach improves forecasting and strengthens climate risk

management **(Klaus et al., 2021)**, helping to better prepare for and mitigate the effects of extreme weather events.

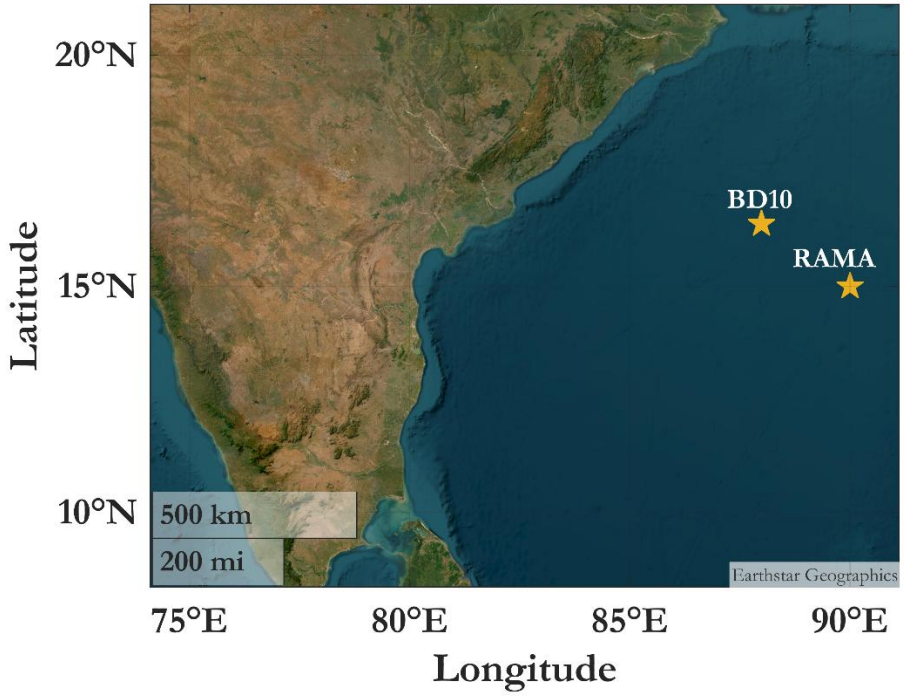


Fig 2: Locations of moored buoy BD10 and RAMA buoy at which the Shortwave Radiation data is analysed

Technical difficulties and many kinds of inaccuracies are introduced by buoy data gaps. When points must be approximated using statistical or machine-learning methods **(Tutz, Gerhard et al., 2015)**, data interpolation errors are a significant problem. Inaccuracies and biases can be introduced by improper interpolation, which can result in incorrect results. Furthermore, because missing data points may skew long-term patterns and variability in variables like SST, wind speed, and SWR, gaps can have an impact on trend analysis and anomaly detection. Additionally, incomplete records make it more difficult to validate and calibrate climate models, which lowers their ability to anticipate extreme weather events like monsoons and cyclones.

The loss of temporal continuity, which makes it more difficult to record high-frequency changes in atmospheric and oceanic conditions, is another major obstacle. This is especially troublesome when examining quickly changing phenomena, like tropical cyclones, when tracking variations in wind speed, pressure, and SST requires constant data noted in their paper on big data challenges in ocean observation that the problem is made worse by the fact that large-scale ocean datasets frequently have problems with data inconsistency, incompleteness, and real-time processing. These

difficulties make data assimilation procedures more difficult, and missing data points might result in less accurate models and more uncertain forecasts.

This study utilizes machine learning models like Random Forest (RF), Support Vector Machine (SVM), and Neural Network (NN) to fill in the gaps in buoy data, as do in a study by (**Kolukula and Murty,2025).** By identifying patterns in the existing data, machine learning models assist in predicting these missing values while preserving the consistency needed for precise cyclone and climate modelling. Through the integration of BD10 buoy data, ISCCP-derived parameters, and advanced machine learning techniques, this study successfully enhances SWR prediction and gap-filling. The results contribute to improved ocean-atmosphere interaction analysis and support more reliable climate assessments in the Bay of Bengal region.

2.0 DATA - Multi-Source Data Architecture and Integration

The data used in the study is collected from a variety of sources, each of which contributed a unique set of parameters needed to use machine learning models to fill the Enhanced SWR gap. These data sources include reanalysis data from the RAMA buoy, Buoy Data 10 (BD10), and the International Satellite Cloud Climatology Project (ISCCP). The crucial data from each of these sources is combined to create and improve the prediction models.

The efficacy of machine learning in physical oceanography depends on "data fusion"—the synergistic integration of in-situ buoy observations with satellite-derived products and reanalysis datasets to establish a definitive "ground truth." By aligning disparate data streams, we can capture the complex, multi-scale physics of the ocean-atmosphere interface.

Primary Data Pillars

To reconstruct Shortwave Radiation (SWR), this study integrates four primary pillars of environmental data:

1. **ISCCP (International Satellite Cloud Climatology Project):** Provides three-hourly satellite-based measurements of cloud cover and radiative fluxes, offering the temporal granularity necessary to capture diurnal cycles since 1983.
2. **BD10 Moored Buoy:** A critical node in the Indian National Data Buoy Program located in the Bay of Bengal (87.99°E, 16.36°N), providing high-precision, direct observations of SWR, SST, and atmospheric pressure.
3. **RAMA Network (Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction):** A broad Indian Ocean array that provides the supplemental atmospheric metrics and cross-validation points required to refine predictive correlations.

4. **Reanalysis Products (ERA5/MERRA-2):** These integrate historical observations into numerical weather models, providing a continuous, high-resolution record to bridge extensive temporal voids in buoy data.

2.1 International satellite cloud climatology project (ISSCP)

This study makes use of the ISCCP dataset, which provides annual atmospheric and radiative flux data from January 1, 2013, to July 30, 2017. The three-hourly temporal resolution of the dataset reduces noise and enhances data quality by capturing diurnal fluctuations in cloud cover, radiative fluxes, and surface characteristics. The machine learning model's capacity to forecast SWR and represent variations in solar radiation is improved by this granularity. The NOAA NCEI THREDDS catalogue is the source of the ISCCP dataset **(Tsuang et al.,2008)**, which is renowned for offering long-term, validated satellite-based climate data that is essential for assessing SWR trends and enhancing climate projections in the Bay of Bengal area. Important variables that affect how much solar radiation reaches the Earth's surface include the cosine of the solar zenith angle ($\mu 0$), which was taken from the ISCCP dataset. $\mu 0$ enhances radiation modelling and aids in capturing SWR variability. With its high geographical and temporal resolution readings, the ISCCP dataset has been a significant resource since 1983 and is widely utilised in climate studies. This study advances our knowledge of ocean-atmosphere interactions and strengthens climate forecasts by merging ISCCP data with buoy observations and re-analysis datasets.

2.2 Data from the BD10 moored buoy

Buoy Data 10 (BD10) was placed in the Bay of Bengal as part of the NBP to track oceanic and atmospheric conditions. To research monsoon dynamics and ocean-atmosphere interactions, BD10 offers real-time data on variables such as air temperature, wind speed, humidity, pressure, SST, and SWR. High-resolution information on surface radiation fluxes and climatic variables is provided by the buoy. However, sensor malfunctions, problems with data transmission, or maintenance frequently result in missing data for BD10. To guarantee accurate datasets for climate studies, these gaps must be filled. This study predicts and reconstructs missing SWR data using machine learning approaches such as RF, NN, and SVM. This study advances our understanding of oceanic and atmospheric processes in the Bay of Bengal and improves climate modelling by expanding the dataset of BD10.

2.3 Data from the Rama buoy

In order to investigate ocean-atmosphere interactions, a network of moored buoys has been placed throughout the Indian Ocean as part of the Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction. RAMA was founded in 2004 and offers real-time atmospheric and oceanic data that is essential for climate research and monsoon dynamics. Shortwave radiation, sea surface temperature, wind speed, and air pressure are among the useful metrics provided by the RAMA buoys. These excellent observations improve climate models and close gaps in the SWR dataset. Weather forecasting in under-sampled areas and a better understanding of air-sea interactions depend on RAMA's data. The network contributes significantly to the advancement of climate forecasts and oceanic research.

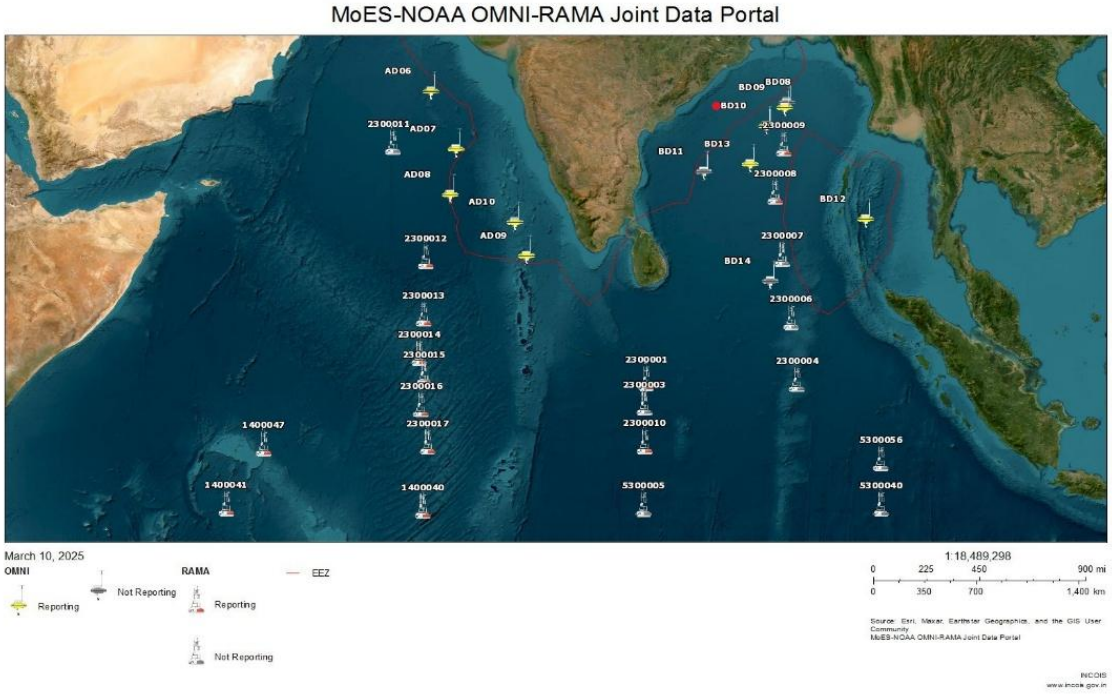


Fig 3. RAMA and OMNI buoy network.

In this study, BD10 data are supplemented using RAMA buoy data. Because SWR data from RAMA is accessible, it is used to create relationships and correlations that help machine learning techniques recreate missing values in BD10. This study enhances climate modelling efforts and improves the accuracy of SWR forecasts, and advances knowledge of monsoon variability and ocean-atmosphere interactions in the Indian Ocean region by integrating RAMA measurements.

2.4 Data from the Reanalysis

By incorporating historical observational data from weather stations, ships, buoys, and satellites into numerical weather models, reanalysis data is produced. Because they provide a trustworthy long-term record of atmospheric and oceanic conditions, these records are crucial for scientific study, climate studies, and weather forecasting. Accurate SWR prediction depends on continuous, high-resolution climate data, which reanalysis data provides to fill in the gaps in buoy measurements. Important meteorological characteristics, including wind speed and surface radiation fluxes, are provided by sources such as MERRA-2, NCEP/NCAR Reanalysis, and ERA5 **(Hersbach et al., 2020)**. This study improves climate projections and SWR estimations by integrating reanalysis data with buoy observations and machine learning methods.

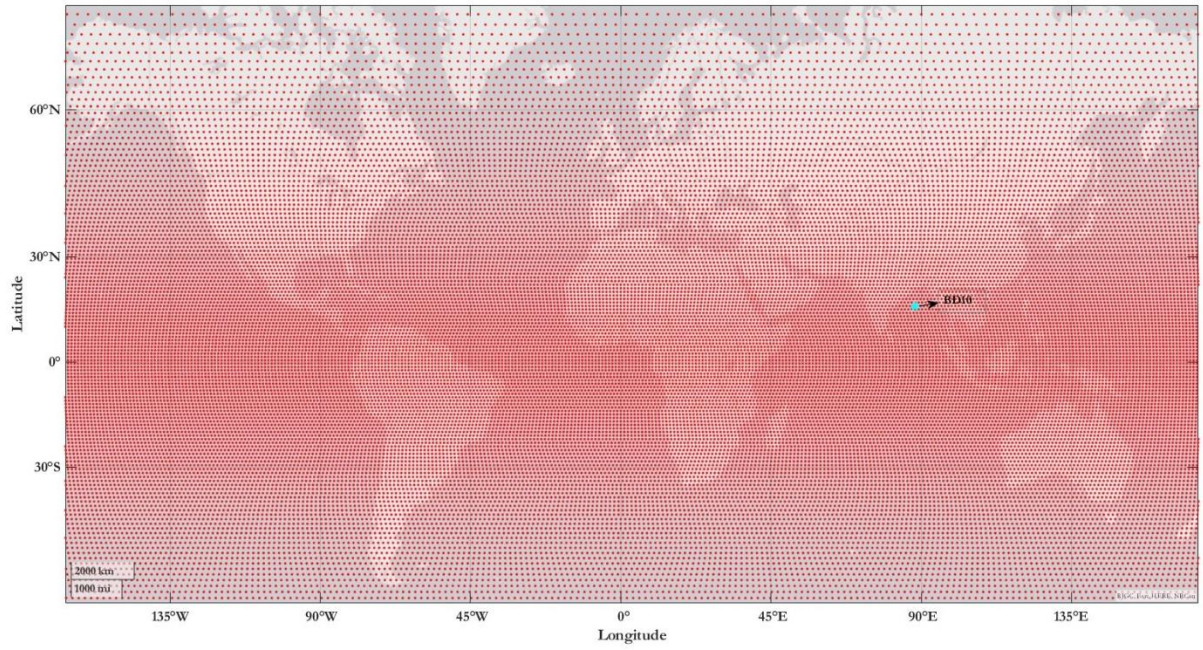


Fig 4. ISSCP CloudAmount distribution with BD10 buoy

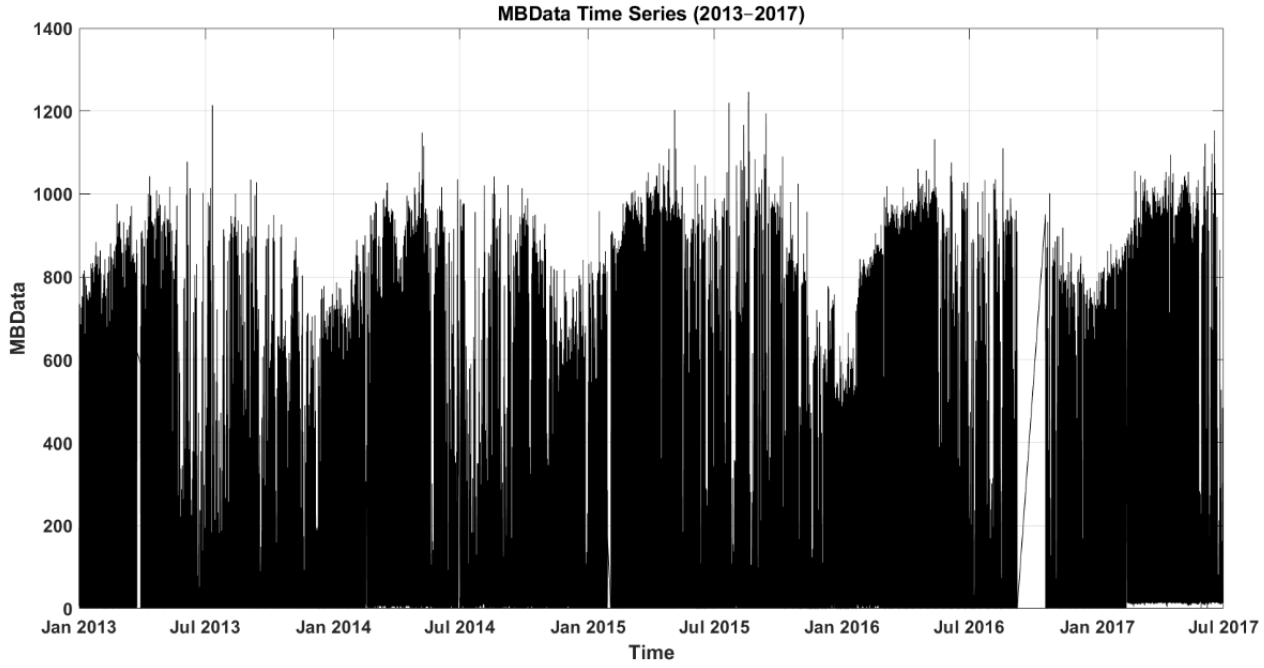


Fig 5. Time series of solar irradiance (W/m²) from 2013 to 2017.

In these recursive predictions, as the future predictions are based on the previous predictions, small errors in each prediction accumulate over time, magnify, and deviate significantly away from the true values, leading to inaccurate predictions. Hence, this approach is suitable to fill small temporal gaps but not suitable to fill long temporal gaps, as shown in Fig. 5.

Figures 5 illustrate the division of meteorological data (MB Data) over the period from 2013 to 2017 for machine learning model development. Notably, the dataset contains visible gaps, which may be due to missing or unavailable measurements during certain time intervals. For model training purposes, the data has been strategically split: the first 80% is used for testing, ensuring robust evaluation across a broad temporal range. The following 10% is allocated for training, where the model learns patterns and relationships from the data. The remaining 10% is reserved for validation, which is critical for fine-tuning the model and preventing overfitting. This temporal partitioning helps in assessing the model's performance and generalization capability on unseen data.

Feature Engineering and Physical Logic

The model architecture utilizes several key parameters to capture the environmental factors influencing surface irradiance:

Feature	Unit	Functional Description
Shortwave Radiation (SWR)	W/m^2	The target variable representing solar irradiance received at the surface.
Cloud Amount	Oktas	A measure of cloud cover (0-8); primarily dictates the attenuation of solar flux.
μ0 (Mu0)	H/m*	The cosine of the solar zenith angle; defines the geometric path of solar radiation.
Zenith Angle	Degrees/Rad	The angle of the sun relative to the local vertical; influences optical path length.
Time/Hour	3-hourly	Captures diurnal and seasonal periodicities in solar exposure.
Lat/Lon	Coordinates	Geographic factors influencing the local solar noon and regional radiation intensity.
Sin/Cos Terms	Unitless	Trigonometric transformations of cyclical features (Hour, Day) to preserve periodicity.

Table 1: Parameters considered in the study and their units

The Fig. 6 displays the inverse relationship between SWR and Total Cloud Cover over the month of June 2017. As evident in the plot, SWR (shown in blue) peaks during periods of low cloud cover, while it significantly drops during times of high cloud presence, indicated by the brown curve approaching values close to 1. This behavior aligns with atmospheric physics, where increased cloud cover obstructs incoming solar radiation, thereby reducing the amount of SWR reaching the Earth's surface. Conversely, on clearer days with minimal cloud cover, solar radiation is largely unimpeded, resulting in higher SWR values. This inverse correlation is a crucial factor in understanding and modeling surface energy balance and radiative transfer in climate and weather studies.

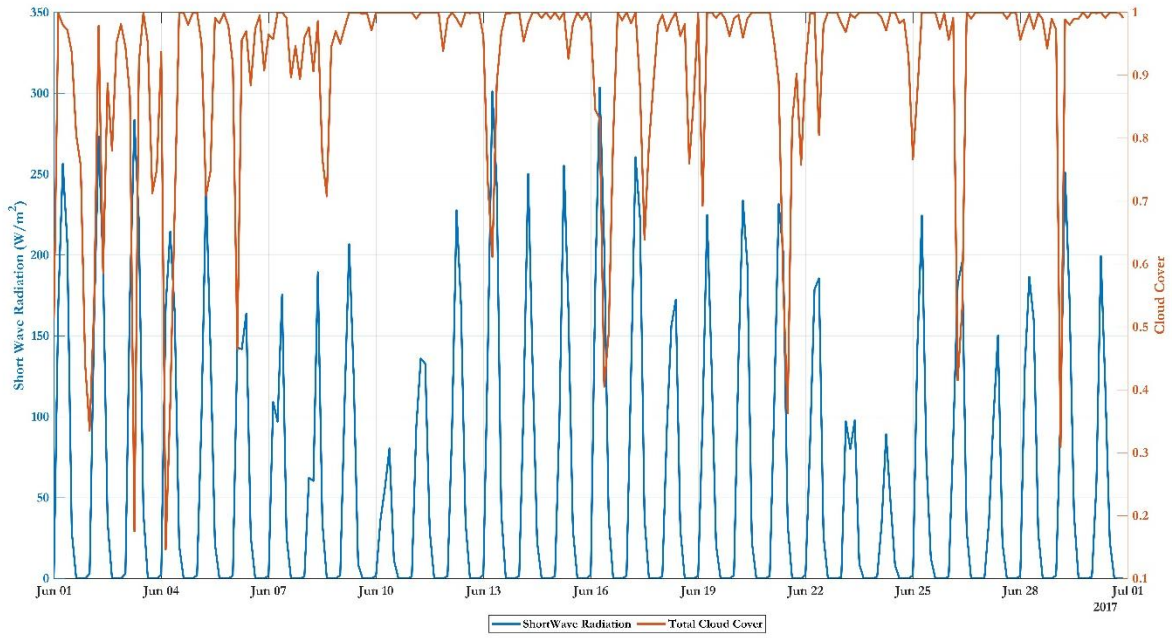


Fig 6. Relationship Between Irradiance and Total Cloud Cover

3.0 METHODOLOGY

In this study, we focus on the time from January 1, 2013, to July 31, 2017, during which the dataset contains several missing values. To address these data gaps, we employ machine learning models such as Random Forest, Support Vector Machine (SVM), and Neural Networks to estimate and fill the missing SWR values **(Kolukula, Murty, 2025)**. Each of these models applies a different methodology: Random Forest operates by constructing an ensemble of decision trees and averaging their outputs, SVM finds an optimal hyperplane in a high-dimensional space for regression, and Neural Networks mimic the human brain by learning complex patterns through interconnected layers of neurons. Among these, Random Forest performs the best in terms of accuracy and robustness, effectively handling non-linear relationships and outliers, making it the preferred model for gap-filling in this analysis.

Focuses on the estimation of SWR, commonly known as irradiance, using machine learning techniques at a specific oceanographic site—OMNI buoy BD10, located at (87.99°E, 16.36°N). This site is part of the larger ocean observation system that monitors meteorological and oceanographic parameters in the Indian Ocean region **(Karnan and Sreedharan, 2024)**. The goal is to accurately estimate irradiance at this location using machine learning models trained on historical data from January 2013 to July 2017. Since solar radiation plays a critical role in the Earth's energy balance and is essential for understanding surface heat fluxes, having a complete and reliable dataset is crucial for both scientific research and operational models. Due to gaps or

missing values in direct observations, your project aims to reconstruct or predict missing irradiance data using model-based estimations. These estimates are not only important for improving climate models but also for applications like solar energy prediction, air-sea interaction studies, and weather forecasting in marine contexts.

The dataset comprises variables collected at three-hour intervals, making it ideal for capturing both diurnal and seasonal patterns in irradiance. Inputs used for model training include date, hour, year, longitude, latitude, mu0 (cosine of the solar zenith angle), and cloud amount. The output is the SWR in W/m². The Mu0 parameter, which is essential for capturing the geometry between the Earth and the Sun, is obtained from the RAMA buoy, located at (90°E, 15°N), which complements the OMNI data. Any missing values (NaN) in mu0 are replaced with -1 to keep the data frame consistent and avoid breaking the model during training. Another vital predictor is cloud amount, which significantly affects how much solar radiation reaches the Earth's surface. CloudAmount is derived using the standard meteorological metric of oktas, calculated using the formula:

$$Cloud\ amount\ (oktas)\ =\ \left(\frac{number\ of\ clouds\ covering\ sky}{Total\ possible\ cloud\ observations}\right)\times\ 8$$

To organize and prepare the data for machine learning, you create two main tables. Table 1 contains the target variable SWR indexed by year, month, day, and hour, while Table 2 contains the corresponding input features, including cloud amount and mu0. These tables are then merged based on common date-time indices to form a consolidated dataset named Table T. This merging step is crucial as it aligns all features and the target variable to the same temporal resolution, ensuring data integrity. After merging, the predictors used include cloudAmount, mu0, year, month, day, and hour, and the target is SWR. This dataset T is then split into training (80%), validation (10%), and testing (10%) subsets, a standard practice in machine learning to avoid overfitting and to ensure the model generalizes well to unseen data. Splitting the data this way allows you to monitor training performance, tune hyperparameters, and evaluate final performance in a structured manner.

3.1 FILLING THE GAP IN DATA

Due to the inherent challenges associated with observational data, such as sensor malfunctions and adverse weather conditions, gaps in the SWR dataset are unavoidable. To address this issue, machine learning models were employed to fill the missing values by learning complex relationships between meteorological parameters and SWR. Several models were evaluated for this purpose, including Random Forest, Support Vector Machine, and Neural Networks.

Random Forest (RF), an ensemble learning method that constructs multiple decision trees and aggregates their predictions, was found to be the most effective model for gap-filling. RF divides the data into subsets and trains individual decision trees on bootstrapped samples, thereby reducing variance and improving prediction accuracy. The model aggregates predictions through averaging (for regression tasks) or majority voting (for classification tasks), resulting in a robust and reliable estimation of missing values.

Support Vector Machine (SVM) is a supervised learning algorithm that identifies an optimal hyperplane to classify or predict data points. SVM is particularly effective in high-dimensional spaces and is capable of handling nonlinear relationships between features. However, due to its computational complexity and sensitivity to hyperparameters, SVM was less effective in handling large datasets with missing values, making it less suitable for this task.

Neural Networks (NN), which mimic the human brain by learning complex patterns through interconnected layers of neurons, were also explored as an alternative approach. While NNs can capture highly nonlinear relationships and provide accurate predictions, they require a large volume of data and computational resources. Given the relatively limited size of the dataset and the need for computational efficiency, RF was chosen as the primary model for gap-filling.

For the modelling process, we were using three popular machine learning algorithms: RF, SVM, and NN. These models are chosen due to their ability to handle nonlinear relationships, which are common in meteorological data. Among these, Random Forest emerges as the top performer, providing the best accuracy and generalization capability. RF is a bagging ensemble model, which means it builds multiple decision trees on different subsets of the data and averages their outputs to improve prediction and reduce variance. This makes it particularly effective in handling the noisy and irregular nature of environmental datasets. In addition, normalization is applied to the features before model training, especially to benefit distance-based models like SVM and to stabilize neural network training. By scaling the data, each feature contributes proportionally to the model's learning process.

Model performance is evaluated using two key metrics: Root Mean Square Error (RMSE) and R-squared (R^2). RMSE measures the average magnitude of error between the predicted and actual values—lower values indicate better accuracy. R^2 measures how well the model explains the variability in the target variable—a value closer to 1 indicates a better fit. In your case, the Random Forest model achieves the lowest RMSE and highest R^2, confirming its effectiveness. This means that the model is accurately capturing the underlying relationships between cloud cover, solar position, and irradiance. The model can now be applied to estimate SWR values for timestamps

where data is missing or unreliable, thus creating a more complete and consistent irradiance time series for OMNI BD10. This dataset can then support broader research on climate variability, solar energy availability, and ocean-atmosphere interactions in the tropical Indian Ocean.

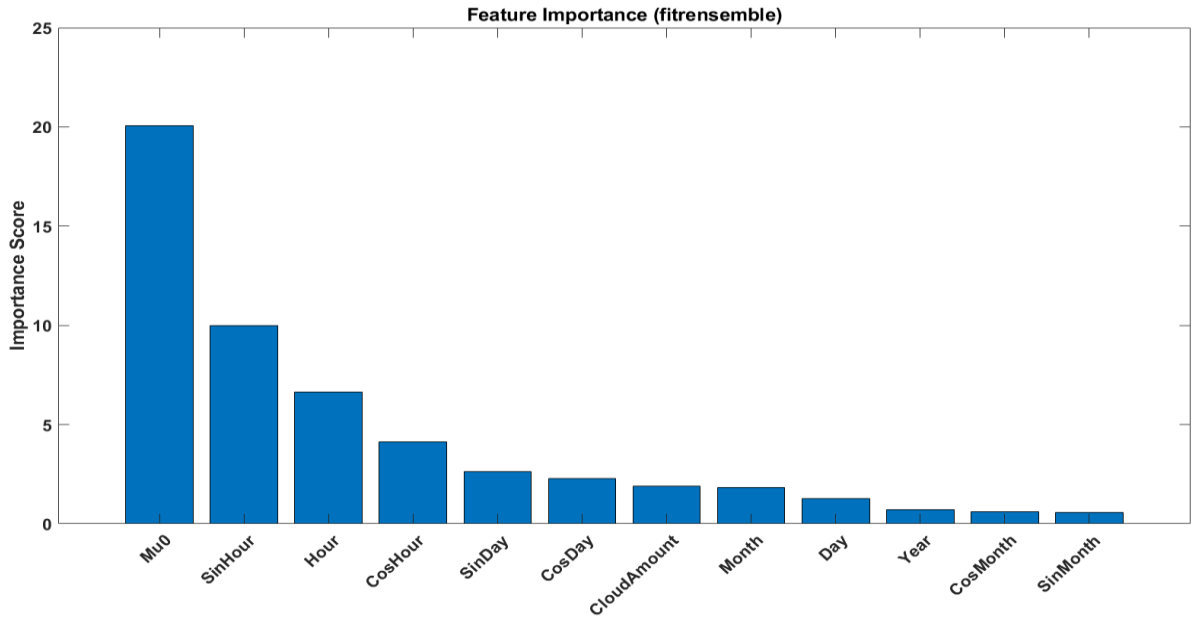


Fig 7: Feature selection

Using an ensemble model (fitrensemble), the bar plot displays the feature relevance for shortwave radiation prediction. The strongest influence of solar geometry is demonstrated by the most significant characteristic, Mu0, which is followed by SinHour and Hour. Trigonometric and temporal transformations such as CosHour, SinDay, and CosDay also make a moderate contribution. Month and CloudAmount are less significant but still significant. Features with little effect on the model's predictions include Year, CosMonth, and SinMonth.

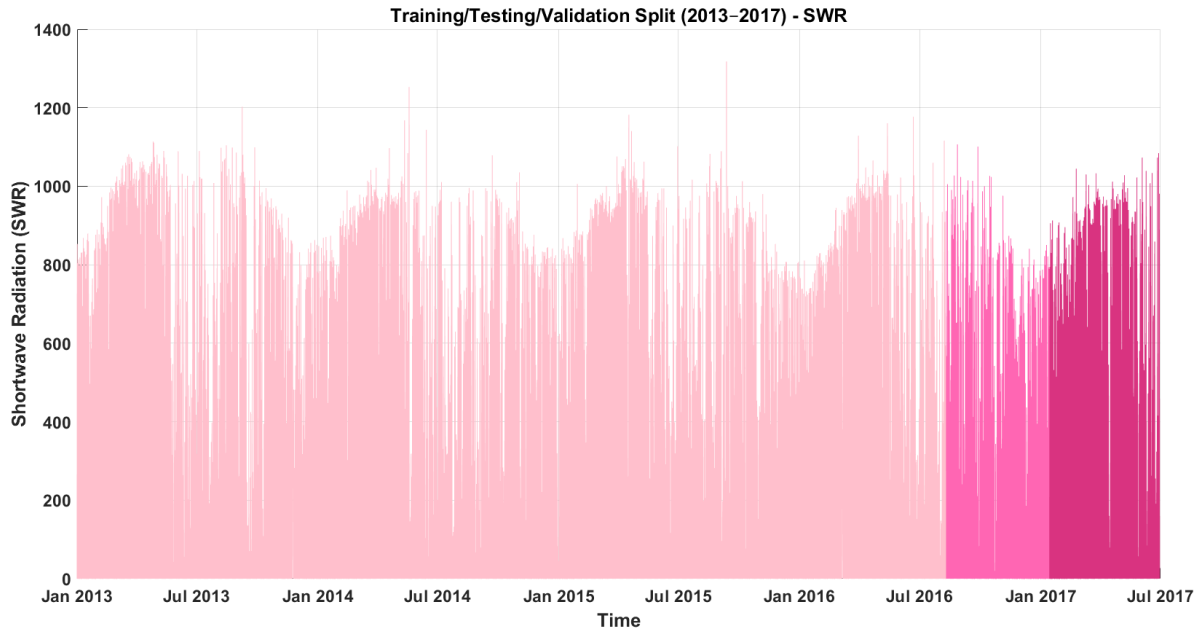


Fig 8. Time Series of SWR Plot of Irradiance from RAMA buoy

In machine learning, the testing dataset plays a crucial role in evaluating the generalization performance of a trained model. As shown in the time series plot from the RAMA buoy data, the pi (80%) represents the data reserved for testing. This dataset is not seen by the model during training or validation; instead, it's used exclusively after the model has been trained and fine-tuned. The purpose is to provide an unbiased estimate of the model's performance on new, unseen data, simulating how the model would behave in real-world scenarios. By analyzing metrics such as RMSE (Root Mean Square Error) and R^2 (Coefficient of Determination) on this dataset, we can assess how well the model predicts shortwave radiation (SWR) across different time periods, ensuring it hasn't simply memorized the training data but has actually learned to generalize underlying patterns. The training phase incorporated a subset of the data (10%) for model learning and another 10% for validation, allowing us to fine-tune hyperparameters and avoid overfitting. During testing (80%), we withheld a large portion of data to evaluate performance, ensuring the model was evaluated on unseen data. With strong metrics (e.g., low RMSE and high R^2), the model demonstrated its capacity to accurately estimate SWR based on inputs like mu0, cloud amount, time, and location.

The following steps were followed in this document to address the data gaps:

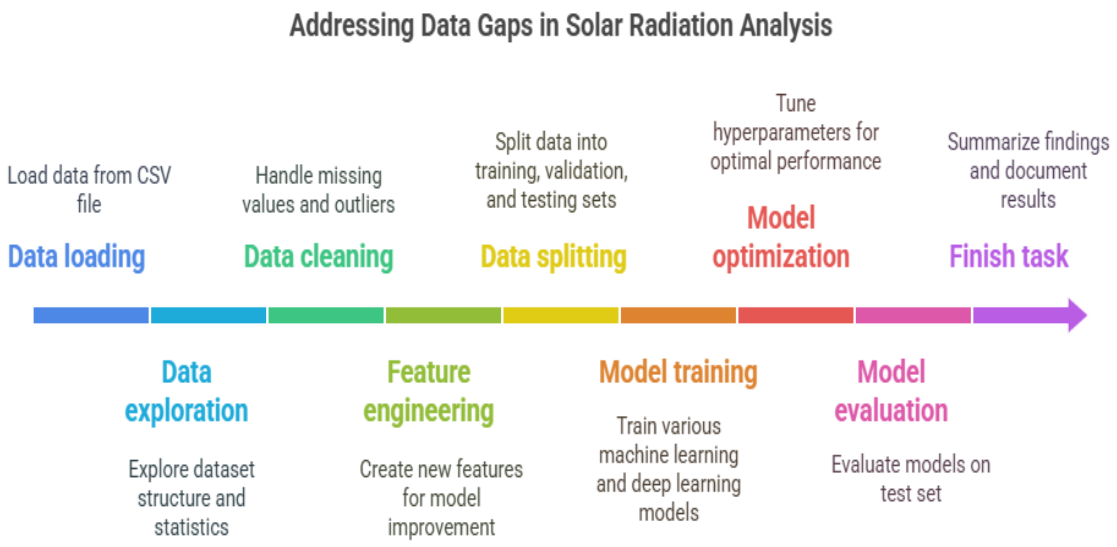


Fig 9. SWR prediction stepwise flowchart

3.2 SHORTWAVE RADIATION

Short-wave radiation is the solar radiation received at the ground. It is defined as the sum of the direct and the diffuse solar radiation:

$$\mathbf{R_{sw} = (R_{dir} + R_{dif})\ (1 - a)}$$

where the factor *a* is the albedo and describes the amount of reflected radiation. The albedo is set to 0.25

The direct solar radiation arrives from the direction of the solar disk and includes a small component scattered forward. Its amount on a horizontal surface is obtained from the ERA5 database as the total sky direct solar radiation at the surface (it is measured in $\left[\frac{MJ}{m^2 h}\right]$. Since the terrain conformation is not flat, the radiation has been rescaled as the amount of direct radiation received by an inclined surface, with the slope and aspect of the local terrain (for more details about the slope and aspect, please visit the supplementary materials). The flux of radiation that reaches an inclined surface can be computed as:

$$\boldsymbol{R_{dir} = \frac{S_{Dir}}{\cos\theta}\zeta[\cos\theta\cos\beta + \sin\theta\sin\beta\cos(\emptyset - \alpha)]}$$

where S_{Dir} is the the total sky direct solar radiation at surface obtained from the ERA5 database, ζ is defined as a binary mask which indicates if the surface is shaded or not, θ is the sun zenith, ϕ is

the sun azimuth, β is the surface slope and α is the surface aspect (for more details about all these variables please visit the supplementary materials).

The diffuse radiation describes all scattered radiation received from the blue sky and from clouds, either by reflection or transmission. The diffuse radiation on a horizontal surface is computed as the surface solar radiation downwards (the amount of solar radiation that reaches a horizontal plane at the surface of the Earth) minus the direct solar radiation computed as described above (it is measured in $\left[\frac{MJ}{m^2h}\right]$). The diffuse radiation on an inclined surface can be computed as proposed by dividing it in an isotropic component R_I, an anisotropic component R_A and a component reflected by the surroundings R_R:

$$\mathbf{R_{dif} = R_I + R_A + R_R}$$

The isotropic component R_I can be computed as:

$$\mathbf{R_I = \frac{1}{2} D\ (1 + \cos\ (\propto))\ (1 - k) s_{vf}}$$

where *D* is the diffuse radiation on a horizontal surface obtained from, α is the slope of the surface, *k* is the anisotropic index defined as $\mathbf{k = \frac{S_{dir}}{\cos(\theta)\ R_\circ}}$, where $R_\circ = 4.87\ \left[\frac{MJ}{m^2h}\right]$ and s_{vf} is the sky view factor (for more details about all these variables, please visit the supplementary materials).

The anisotropic component R_A can be computed as:

$$\boldsymbol{R_A = \frac{R_{dir} D}{R_0}}$$

where R_{dir} is the direct short-wave radiation computed in the section above. R_A describes the circumsolar radiation.

The component reflected by the surroundings R_R can be computed as:

$$\boldsymbol{R_{A\ } = \frac{R_{dir} D}{R_0}}$$

where R_{dir} is the direct short-wave radiation computed in the section above. R_A describes the circumsolar radiation.

The component reflected by the surroundings R_R can be computed as:

$$\boldsymbol{R_{R = \frac{1}{2} D a_r (1 - \cos(\alpha + h_a))}}$$

where a_r is the mean albedo of the surroundings (it is set to 0.25 as pointed out) and h_a is the mean horizon angle (for more details about all these variables, please visit the supplementary

Table 2: Description of input features used for shortwave radiation prediction

Feature	Description
date, hour, month, year	Time features capture seasonal and diurnal variations
Mu0	Cosine of the solar zenith angle – a key factor in solar radiation intensity
cloud amount	Affects how much solar radiation reaches the surface
lat, lon	Geographic location impacts sun angle and radiation levels
sin() and cos() terms	Transform cyclical features like hour, month, year to preserve periodicity

Several supervised machine learning techniques are used in this study to forecast surface SWR **(Sarafanov, Mikhail et al., 2020; Smola, Alex et al., 2004)** using temporal and meteorological variables. Neural Networks (NN) **(Krasnopolsky, Vladimir et al., 2016; Gurney, Kevin et al., 2018)**, Support Vector Regression (SVR), and ensemble learning techniques (such as Random Forest, Bagging, Boosting, and Least Squares Boosting (LSBoost) are the models used. A dataset comprising inputs like solar zenith cosine (μ_0), cloud quantity, latitude, longitude, and cyclical encodings of time variables (month, hour, year) is used to train each model.

3.3 SUPPORT VECTOR MACHINE (SVM)

Support Vector Regression is a kernel-based algorithm adapted from Support Vector Machines (SVMs) for regression tasks. SVR fits a function:

$$\mathbf{f(x)= \langle w,x \rangle + b}$$

by minimizing the regularized loss:

$$\textbf{Minimize } \frac{1}{2}|\boldsymbol{w}|^2 + \mathbf{C} \sum_{i=1}^{N} (\boldsymbol{\varepsilon_i} + \boldsymbol{\varepsilon_i^*})$$

subject to:

$$yi - \langle w,xi \rangle - b \leq \epsilon + \varepsilon_i$$

$$\langle w,xi \rangle + b - yi \leq \epsilon + \varepsilon_i^*$$

$$\varepsilon_i , \varepsilon_i^* \geq 0$$

where ϵ\epsilonϵ is the error tolerance margin, CCC is a regularization parameter, and εi,εi∗ are slack variables. SVR is particularly useful for capturing non-linear relationships by employing kernel functions like RBF or polynomial kernels. Hyperparameters such as CCC, ϵ\epsilonϵ, and the kernel type are optimized for best performance.

3.4 NEURAL NETWORK (NN)

Neural Networks (NNs) are layered architectures inspired by biological neural systems. In this work, feedforward neural networks (FNNs) are used for regression, consisting of:

- Input Layer: Takes features like cloud amount, μ_0, and transformed time variables.
- Hidden Layers: One or more layers with neurons that apply a weighted sum followed by an activation function.
- Output Layer: Produces a continuous value representing SWR.

Each neuron performs:

$$Z = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

Each neuron takes all its inputs (**X**), multiplies each by a weight (**W**), sums the result, adds a bias (**b**), and then passes this total through an activation function (**f**) to produce an output (**Z**). This output becomes the input for the next layer in the network.

The loss function used is Mean Squared Error (MSE):

$$MSE = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Weights are optimized via backpropagation using gradient descent:

$$w \leftarrow w\ \eta \frac{\partial MSE}{\partial w}$$

3.5 ENSEMBLE LEARNING: RANDOM FOREST, BAGGING, BOOSTING, AND LSBOOST

Ensemble learning combines multiple base models to improve overall prediction accuracy and generalization. This study uses the following ensemble techniques:

- **Random Forest**: A collection of decision trees trained using bootstrapped subsets of the data and random feature selection. The final prediction is the average of predictions from all trees. Random Forest helps reduce overfitting and increases stability.
- **Bagging (Bootstrap Aggregating)**: Builds multiple models (typically decision trees) on different bootstrapped samples of the training data. Each model votes, and the final output is averaged (regression) or majority-voted (classification). It reduces variance and is particularly effective in high-variance models.
- **Boosting:** A sequential ensemble method where each model tries to correct the errors of its predecessor. Boosting reduces both bias and variance, leading to high performance on complex datasets.
- **LSBoost (Least Squares Boosting):** A specific form of boosting for regression tasks. It fits models sequentially by minimizing the residual errors using least squares loss. LSBoost builds additive regression models:

$$\boldsymbol{F_m(x) = F_{m-1}(x) + \eta h_m(x)}$$

where $h_m(x)$ is the weak learner (typically a shallow tree), and η is the learning rate. LSBoost is highly effective for capturing complex, non-linear relationships in regression tasks.

Hyperparameters such as the number of trees, tree depth, learning rate (for boosting), and subsampling ratio are optimized using grid search or cross-validation.

3.6 PERFORMANCE ANALYSIS & MODEL EVALUATION

In climate science, Root Mean Square Error (RMSE) and R-squared (R^2) serve as the definitive benchmarks for model accuracy, supplemented by Mean Absolute Error (MAE) to understand error distribution. The performance of the candidate architectures is detailed below. Per the methodology in the source context, the dataset was partitioned into a non-standard but robust split: **10% for training, 10% for validation, and 80% for testing** to ensure maximum generalization across a broad temporal range.

Table 3: Machine learning model training and testing performance comparison.

MODEL	Training			Testing		
	RMSE(W/m²)	R-square	MAE	RMSE(W/m²)	R-square	MAE
Neural Network	162.5149	0.8304	96.5099	162.3753	0.8273	95.7471
Support Vector Machine	185.1430	0.7987	102.1217	184.6536	0.7915	101.5281
Least-squares Boosting	122.8348	0.9077	67.5942	130.6343	0.8923	71.9679
Regression Tree	183.6462	0.7769	126.4442	183.5211	0.7727	125.4130

4.0 RESULT AND DISCUSSION

Evaluation and Feature Selection

The **Bagging/Random Forest** model emerged as the superior solution, achieving an RMSE of **71.05** and an R^2 of **0.9703**. This high degree of accuracy is critical for "So What?" considerations: it enables the reliable calculation of net heat flux, the primary driver of oceanic heat content studies.

Feature importance analysis indicates that solar geometry (**mu_0**) and temporal variables (**Hour** and **SinHour**) are the dominant predictors. The use of trigonometric transformations (Sin/Cos) is vital as it preserves the cyclical nature of time, preventing the model from treating the transition from 23:00 to 00:00 as a linear jump. The statistical validity of these predictions was confirmed via Taylor Diagrams, showing high correlation and minimal standard deviation variance compared to observed data.

The Bagging model's superior performance can be attributed to its ability to reduce variance and improve prediction accuracy by combining the predictions of multiple models. This is particularly useful when dealing with complex datasets where individual models may struggle to capture the underlying patterns.

Bagging, or Bootstrap Aggregating, is a machine learning ensemble method that combines the predictions of multiple models to improve prediction accuracy and reduce variance. By training multiple models on different subsets of the data and combining their predictions, Bagging can reduce overfitting and improve the model's robustness. It is the concept of bootstrap sampling, where multiple subsets of the data are created by sampling with replacement. Each subset is then used to train a separate model, and the predictions of each model are combined to produce the final prediction.

In contrast, the other models, such as the Neural Network, Support Vector Machine, Least-squares Boosting, and Regression Tree, performed relatively higher with RMSE values and lower R-squared values. Based on these results, we will proceed with the Bagging model for further analysis and prediction.

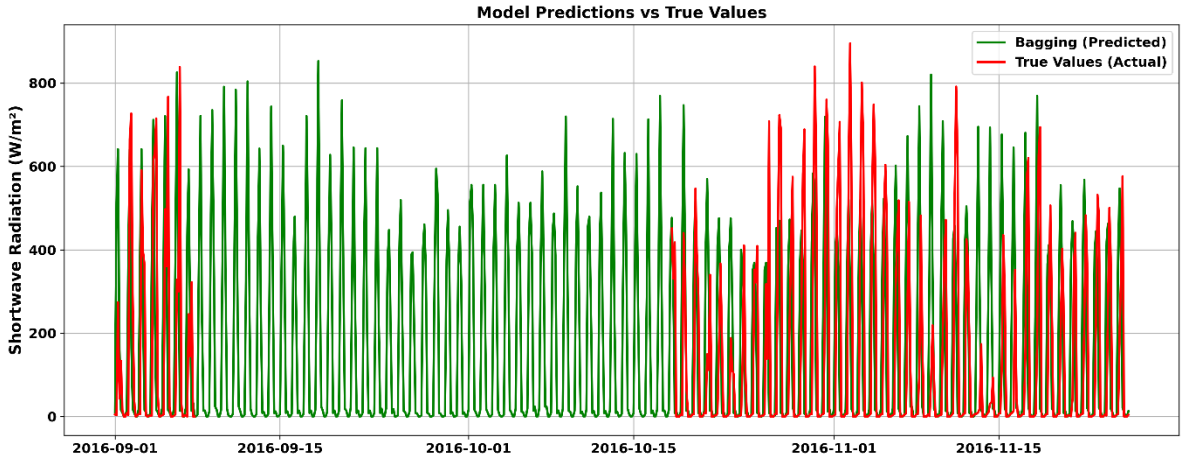


Fig 10. Bagging Model vs Observed

The Bagging model's projections for shortwave radiation are plotted against actual observed values. Red bars show the actual radiation levels, and green bars show the model's forecasts. The model occasionally understates peak values but does a good job of capturing the overall trend and periodicity. In general, Bagging exhibits good agreement, particularly when there is intense radiation and clearer conditions.

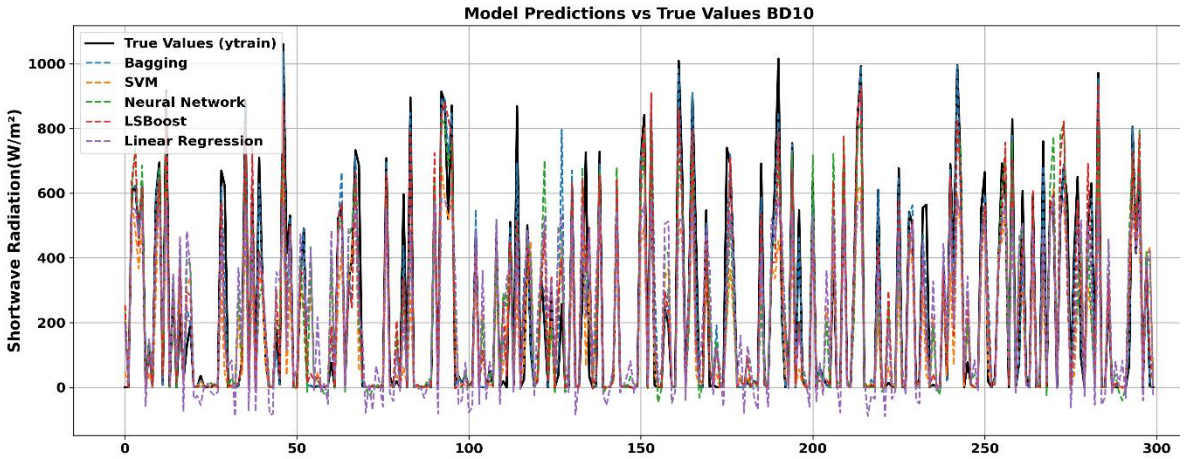


Fig 11. Comparison of ML Models

Five machine learning algorithms' predictions of shortwave radiation are contrasted in this graphic. True values are displayed by the black line, while predictions from regression, LSBoost, SVM,

bagging, and neural networks are represented by the other lines. High accuracy is indicated by the tight adherence of LSBoost and bagging to the true values. Larger variances are displayed in linear regression results, particularly at higher radiation peaks.

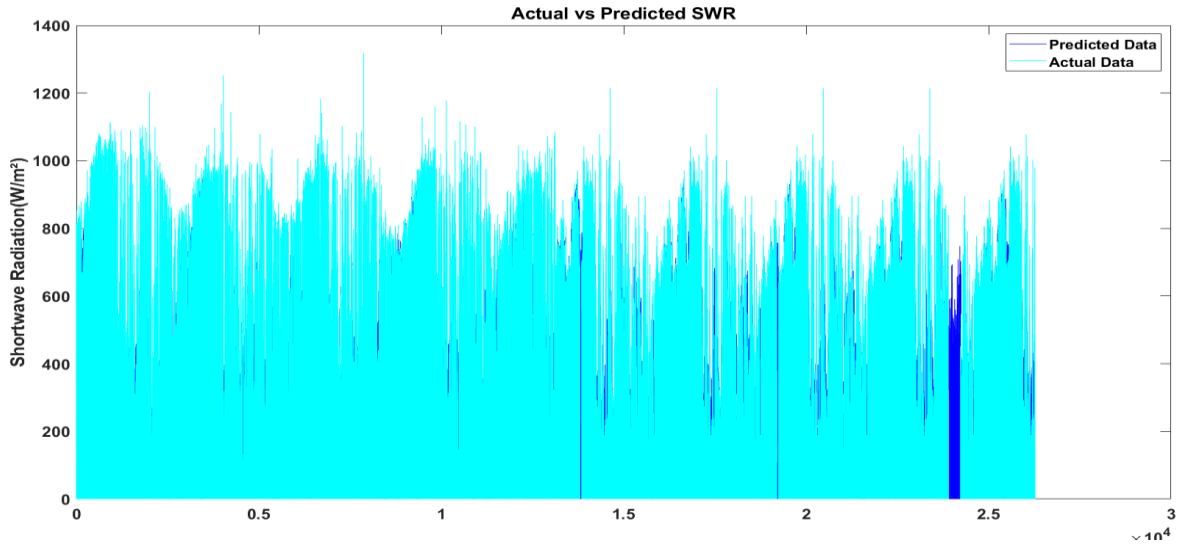


Fig 12. Predicted vs Observed Shortwave radiation (SWR) from 2013 to 2017 using an ensemble model on OMNI and RAMA buoy data.

The data collected from the OMNI and RAMA buoys provides a valuable foundation for developing a reliable machine learning model to predict SWR. Initially, we conducted thorough testing of the model using historical data to evaluate its performance and generalization capability. This testing phase helped validate the model's robustness and its ability to handle real-world variability in oceanic and atmospheric parameters. With the testing complete and the model showing high accuracy, particularly using Random Forest, which delivered the best performance, we are now confident in applying the trained model to both OMNI and RAMA buoy datasets.

Now, we are leveraging the trained model to predict SWR across the entire OMNI and RAMA buoy datasets, expanding its utility from theoretical development to practical application. This approach allows us to fill missing data gaps in SWR records, support climate research, and improve our understanding of ocean-atmosphere energy exchange. By applying the model to these buoys, we aim to generate a continuous, high-quality dataset that can be used for further analysis, modelling, and climate-related studies. The use of multiple buoys also enables a spatially richer estimation of radiation over the Indian Ocean region. The success of this model demonstrates the power of integrating machine learning with oceanographic observation

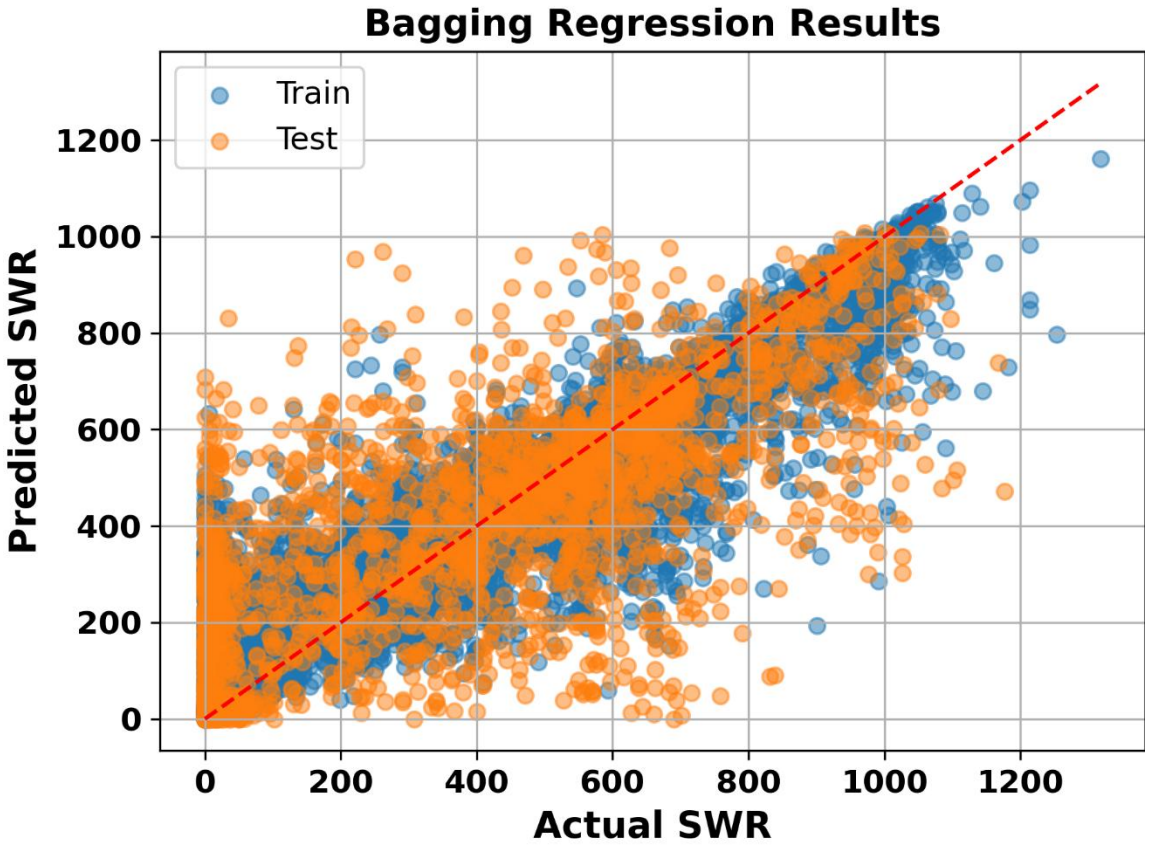


Fig 13. Bagging Model Prediction Accuracy

This scatter plot uses a Bagging Regression model to compare the actual and anticipated Shortwave Radiation values. Predictions for training and testing data are shown by blue and orange dots, respectively. Model performance is shown by the red dashed line, which is the optimal 1:1 prediction line.

Operational Implementation and Scientific Impact

The operationalization of this framework involves applying the trained Random Forest ensemble across the OMNI and RAMA buoy records to generate a seamless SWR time series.

Operational Workflow

1. **Data Partitioning:** Allocation of the 10/10/80 split to prioritize model testing on unseen, real-world data.
2. **Normalization:** Scaling features to stabilize the learning process across disparate units (e.g., coordinates vs. radiation flux).
3. **Gap Identification:** Systematic detection of NaN values within the 3-hourly buoy records.
4. **Ensemble Reconstruction:** Applying the trained model to predict missing SWR based on available meteorological and temporal proxies.

The scientific impact of this gap-filling methodology is substantial. A continuous 3-hourly SWR dataset allows for high-fidelity research into the air-sea energy exchange that fuels the Indian Ocean's climate. For forecasting centers, this means improved monsoon onset predictions and more accurate cyclogenesis modeling. By maintaining the "data pulse" of the Bay of Bengal, we can provide the maritime and agricultural sectors with more reliable lead times for extreme weather events.

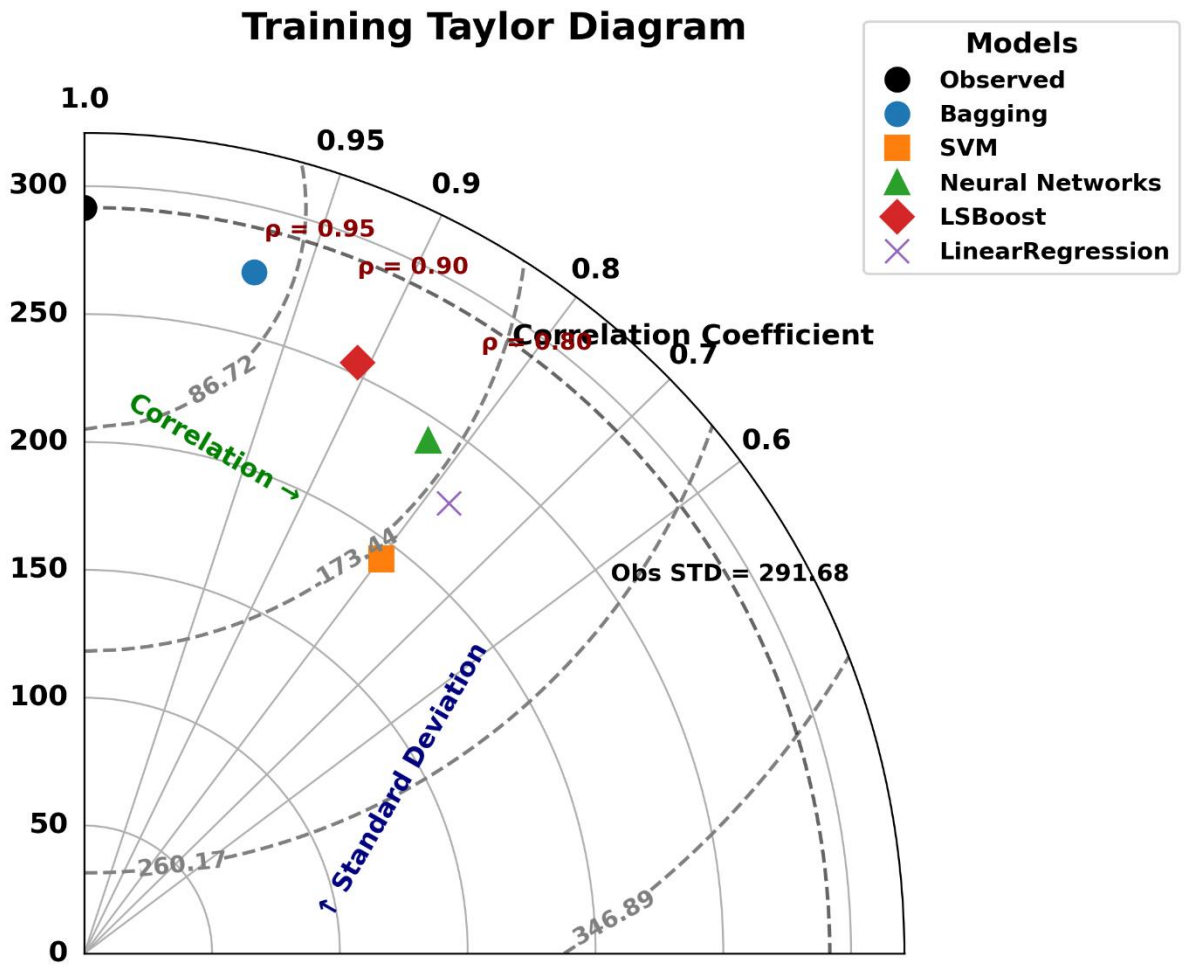


Fig 14. Comparison of various Machine Learning models on training datasets.

This Taylor diagram uses statistical similarity to observed data to assess each model's accuracy throughout the training phase. Once more, the LSBoost and bagging models exhibit a high correlation and good agreement with the observed standard deviation. Effective model training with little overfitting is indicated by the close clustering of models close to the reference point.

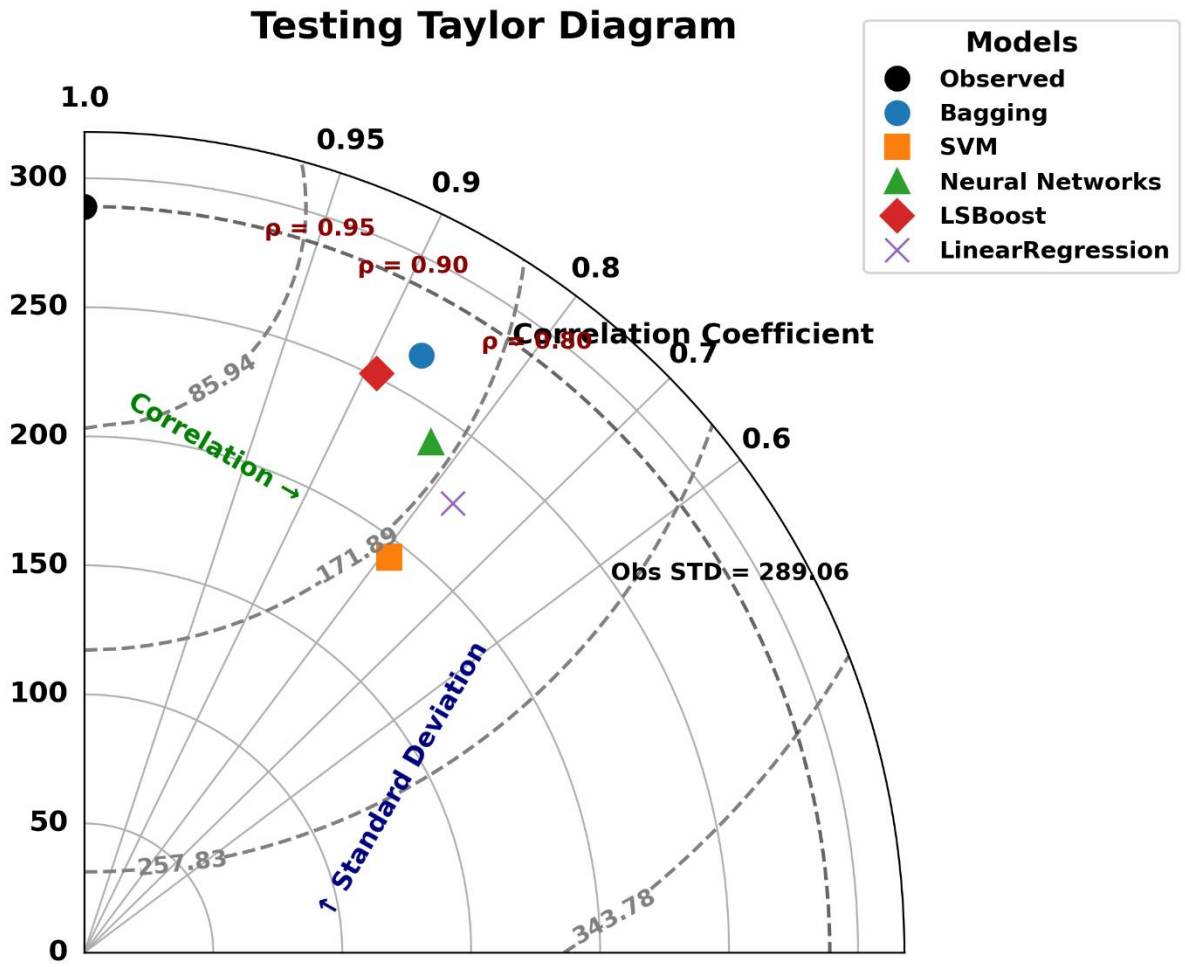


Fig 15. Comparison of various Machine Learning models on testing datasets.

This Taylor diagram uses centred RMS difference, correlation, and standard deviation to show model performance during testing. Because of its closeness to the reference point, the Bagging model exhibits the best agreement with data. Strong performance is also demonstrated by other models, such as LSBoost and Neural Networks, which have moderate variances and high correlation.

5.0 CONCLUSION

This study aimed to address missing data gaps in Buoy Data 10 (BD10), particularly in Shortwave Radiation (SWR) observations, using machine learning models. The inconsistencies in BD10 data, caused by sensor failures and transmission errors, posed significant challenges for oceanic and atmospheric studies in the Bay of Bengal. To mitigate these issues, we integrated data from multiple sources, including RAMA buoys, ISCCP satellite data, and reanalysis datasets, and employed machine learning techniques to reconstruct missing values.

Among the machine learning models tested—including Random Forest, Support Vector Machines, Neural Networks, Regression Trees, and Least-Squares Boosting—the Bagging approach with the Random Forest model achieved the best performance. It produced the lowest Root Mean Square Error of 71.05 W/m^2 and a high R-Squared value of 0.97030, indicating superior predictive accuracy and robustness. The ensemble learning strategy effectively reduced variance, improved model stability, and prevented overfitting, making it the most reliable approach for gap-filling.

The reconstructed dataset will be useful for weather forecasting, monsoon predictions, and long-term climate assessments by utilising advanced machine learning techniques to ensure a continuous and accurate dataset for BD10.

This enhances the understanding of shortwave radiation variability, air-sea interactions, and climate modelling in the Bay of Bengal. Moreover, the same process can be applied to fill gaps in other buoy datasets, making it a scalable and generalizable solution.

Future research could concentrate on integrating deep learning models, incorporating additional meteorological parameters, and improving data assimilation techniques for further improvements in SWR prediction. Additionally, this methodology can be expanded to other regions like the Arabian Sea and Indian Ocean, expanding its impact and applicability. The scalability of this approach is a significant frontier for physical oceanography. The methodology is directly transferable to other parameters such as Sea Surface Temperature (SST), salinity, and wind speed. Furthermore, there is immense potential for **Transfer Learning**; the models trained here could be used to fill gaps in historical datasets (pre-2013) where sensor reliability was significantly lower. The integration of deep learning and more sophisticated data assimilation remains the next objective, moving toward a future of truly dynamic and continuous oceanic monitoring.

6.0 REFERENCE

1. Gerrit, Cécile G., Almuth Arneth, Nicolas B., Laurent B., Philip W. Boyd, Hugo AC Denier van der Gon et al. "Ocean–atmosphere interactions of particles." Ocean-Atmosphere Interactions of Gases and Particles (2014): 171-246.

2. Gurney, Kevin. An introduction to neural networks. CRC press, 2018.

3. Hamilton, Glenn D. "National data buoy center programs." Bulletin of the American Meteorological Society 67, no. 4 (1986): 411-415.

4. Hersbach, Hans, Bill Bell, Paul B., Shoji H., András H., Joaquín Muñoz-Sabater, Julien Nicolas. "The ERA5 global reanalysis." Quarterly Journal of the Royal Meteorological Society 146, no. 730 (2020): 1999-2049.

5. Karnan, C, and Sreedharan G. "Seasonal enhancement of phytoplankton biomass in the southern tropical Indian Ocean: Significance of meteorological and oceanography parameters." Oceanologia 66, no. 2 (2024): 196-219.

6. Klaus K., Casey H., and Vivek Srikrishnan. "Climate risk management." Annual Review of Earth and Planetary Sciences 49, no. 1 (2021): 95-116.

7. Kolukula, Siva Srinivas, and P. L. N. Murty. "Enhancing observations data: A machine-learning approach to fill gaps in the moored buoy data." Results in Engineering 26 (2025): 104708.

8. Philander, S. George. "A review of tropical ocean–atmosphere interactions." Tellus B 51, no. 1 (1999): 71-90.

9. Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.

10. Sarafanov, Mikhail, Eduard K., Nikolay O. Nikitin, and Anna V. Kalyuzhnaya. "A machine learning approach for remote sensing data gap-filling with open-source implementation: An example regarding land surface temperature, surface albedo and NDVI." Remote Sensing 12, no. 23 (2020): 3865.

11. Showman, Adam P., and Timothy E. Dowling. "Earth as a planet: Atmosphere and oceans." In Encyclopedia of the Solar System, pp. 423-444. Elsevier, 2014.

12. Smola, Alex J., and Bernhard S., "A tutorial on support vector regression." Statistics and computing 14 (2004): 199-222. Vapnik, Vladimir. The nature of statistical learning theory. Springer science & business media, 1999.

13. Tsuang, Ben-Jei., Ming-Dah Chou., Yuanchong Zhang., Andreas Roesch., and Kun Yang. "Evaluations of land–ocean skin temperatures of the ISCCP satellite retrievals and the NCEP and ERA reanalyses." Journal of Climate 21, no. 2 (2008): 308-330.

14. Tutz, Gerhard, and Shahla Ramzan. "Improved methods for the imputation of missing data by nearest neighbor methods." Computational Statistics & Data Analysis 90 (2015): 84-99.

15. Venkatesan, R., Shamji, V.R., Latha, G., Simi M., Rao, R. R., Arul M., and Atmanand, M. A. "In situ ocean subsurface time-series measurements from OMNI buoy network in the Bay of Bengal." Current Science (2013): 1166-1177.

16. Venkatesan, R., J. K. Lix, A. Phanindra Reddy, M. Arul Muthiah, and M. A. Atmanand. "Two decades of operating the Indian moored buoy network: significance and impact." Journal of Operational Oceanography 9, no. 1 (2016): 45-54.

17. Vishnu, V., Balaji B., Siva Srinivas K., Nithish Ragav, N., and J. Pavan Kumar, J., "Forecasting of sea surface temperature using machine learning and its applications." Journal of Earth System Science 134, no. 1 (2025): 25.

List of Figures

Fig1. Deployed buoy as part of the National Buoy Program 7

Fig2. Locations of moored buoy BD10 and RAMA buoy at which the Shortwave Radiation is analysed 10

Fig3. RAMA and OMNI buoy network 13

Fig4. ISSCP CloudAmount distribution with BD10 buoy 14

Fig5. Time series of solar irradiance (W/m^2) from 2013 to 2017 15

Fig6. Relationship between Irradiance and total cloud cover 16

Fig7. Feature selection 20

Fig8. Time series of SWR plot of Irradiance from RAMA buoy 20

Fig9. SWR prediction stepwise flowchart 21

Fig10. Bagging Model vs Observed 27

Fig11. Comparing of ML Models 28

Fig12. Predicted vs Observed Shortwave radiation (SWR) from 2013 to 2017 using an ensemble model on OMNI and RAMA buoy data. 28

Fig13. Bagging Model Prediction Accuracy 29

Fig 14 Training Taylor 30

Fig 15 Testing Taylor 31

List of Tables

Table 1: Parameters considered in the study and their units 16

Table 2. Description of input features used for shortwave radiation prediction 23

Table 3. Machine learning model training and testing performance comparison. 26