scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Sea-surface pCO_2 maps for the Bay of Bengal based on advanced machine learning algorithms

A.P. Joshi[1], Prasanna Kanti Ghoshal[1,2], Kunal Chakraborty[1]✉ & V. V. S. S. Sarma[3]

Lack of sufficient observations has been an impediment for understanding the spatial and temporal variability of sea-surface pCO_2 for the Bay of Bengal (BoB). The limited number of observations into existing machine learning (ML) products from BoB often results in high prediction errors. This study develops climatological sea-surface pCO_2 maps using a significant number of open and coastal ocean observations of pCO_2 and associated variables regulating pCO_2 variability in BoB. We employ four advanced ML algorithms to predict pCO_2. We use the best ML model to produce a high-resolution climatological product (INCOIS-ReML). The comparison of INCOIS-ReML pCO_2 with RAMA buoy-based sea-surface pCO_2 observations indicates INCOIS-ReML's satisfactory performance. Further, the comparison of INCOIS-ReML pCO_2 with existing ML products establishes the superiority of INCOIS-ReML. The high-resolution INCOIS-ReML greatly captures the spatial variability of pCO_2 and associated air-sea CO_2 flux compared to other ML products in the coastal BoB and the northern BoB.

Background & Summary

Oceans play a significant role in regulating the amount of CO_2 in the atmosphere. Human-induced anthropogenic activities have increased atmospheric CO_2, counterbalanced by the increasing global ocean CO_2 uptake. Thus, the oceans become over-saturated, and as a result, the regional oceans have been increasingly becoming sources of atmospheric CO_2. An increase in ocean sink strength has been seen in the past decade ($\approx 2.5 \pm 0.6$ GtC per year[1]). The first two years of this decade are reported to have even higher ocean sink strength ($\approx 3.0 \pm 0.6$ GtC per year in 2020[2] and $\approx 2.9 \pm 0.6$ GtC per year in 2021[3]). Based on previous literature, the estimated ocean sink strength of the global coasts has decreased to ≈ 0.2 PgC per year[4,5]. On the other hand, current research shows an increase in the continental shelves' sink strength[6]. The wintertime CO_2 sink in the northern South China Sea behaves stronger after 2007, although this sea area still serves as a weak annual source of atmospheric CO_2[7–9]. These global studies have highlighted the importance and role of the ocean in modulating the atmospheric CO_2, and hence the environment. With the rise in importance of studying the sea-surface partial pressure of CO_2 (pCO_2), the paucity of measured data (especially on a regional scale) is an impediment for observational analysis and model validation[10–12].

This study aims to develop pCO_2 climatological data based on observation for the Bay of Bengal region (BoB). The BoB is recognized for having complex physical dynamics because of significant freshwater input and its distinctive geographical location. The Ganges-Brahmaputra river system, the second largest river system in the world, brings in high freshwater along with organic pollutants into the BoB region[13,14]. The freshwater influx increases stratification and reduces the vertical mixing (thick barrier layer), which influences the absorption and/or outgassing of atmospheric CO_2 in BoB[15]. The nutrients brought down by these rivers decrease the ocean-surface pCO_2 in the offshore region, but its influence diminishes away from the coast[14].

The BoB is influenced by the seasonal reversing coastal currents (East India Coastal Currents (EICC)). From February to March, the EICC brings high saline waters from south to north, which weakens stratification and initiates upwelling. The upwelling brings high subsurface dissolved inorganic carbon (DIC) to the surface, which increases the sea-surface pCO_2[16,17]. The EICC flows south from October to December, carrying less saline waters from the north towards the south. This results in low sea-surface pCO_2 ($\approx 320\ \mu$ atm) values during this period. The freshwater plume spread due to this southward motion of EICC results in low sea-surface pCO_2 values in

[1]Indian National Centre for Ocean Information Services, Ministry of Earth Sciences, Hyderabad, India. [2]Faculty of Ocean Science and Technology, Kerala University of Fisheries and Ocean Studies, Kochi, India. [3]CSIR-National Institute of Oceanography, Visakhapatnam, India. ✉e-mail: kunal.c@incois.gov.in

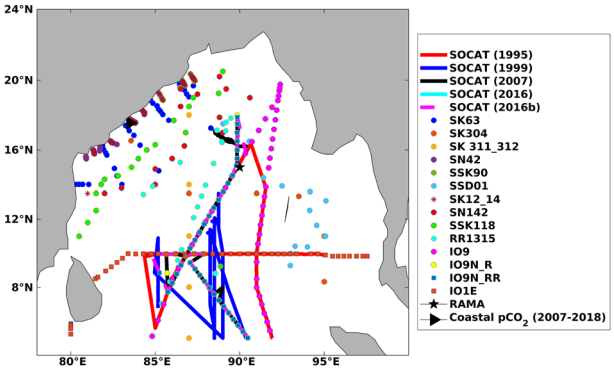

Fig. 1 Representation of the study region (BoB) and the *available of* pCO_2 observations included in this study.

the northern BoB[15]. The spatial pattern of the sea-surface pCO_2 is dominated by the biological and thermal mechanisms[14,18]. Temporal evolution is dominated by solubility, primarily increased by sea-surface temperature (SST) and decreased by DIC[18,19].

The sparse observations of sea-surface pCO_2 constitute a significant hindrance in validating the coupled bio-physical model simulated ocean carbon cycle. The studies[15,17–20] based on bio-physical models often validate the model with the BOBOA (Bay of Bengal Ocean Acidification) mooring[21] at 15° N, 90° E. Another popular observation data is the SOCAT (Surface Ocean Carbon-dioxide Surface Atlas) data[22], which has poor spatial and temporal coverage in the BoB region. These models are often compared with observation-based products like GLODAP[23], which is a spatial annual mean data, and Takahashi data[24], which has a very coarse resolution ($4° \times 5°$). The observation-based products suffer due to a lack of observations in the BoB region, specifically, the unavailability of data near the coast[25,26]. The high freshwater flux, affecting the physical dynamics, also affects these observation-based products as the general assumption (e.g., failure of linear relation assumption of potential alkalinity and sea-surface salinity (SSS)) often fails in the BoB[24].

Besides bio-physical models, the use of regression models[27–29] is popular to understand the carbonate dynamics of the BoB region. These regression models emulate the sea-surface pCO_2 with relatively larger errors. The linearity assumption between the dependent and independent variables is not always true. Region-specific Machine Learning (ReML) algorithms showed promising results for the central BoB[30]. Hence, this study attempts to construct spatiotemporal sea-surface pCO_2 maps for the BoB using observations and advanced ML techniques.

Methods

This study includes a significant number of open and coastal ocean pCO_2 observations and associated variables regulating pCO_2 variability in BoB to come up with a data set that could aid in training advanced ML models (Fig. 1). We assume that the sea-surface pCO_2 is a function of sea-surface temperature (SST), sea-surface salinity (SSS), mixed layer depth (MLD), atmospheric CO_2 mole fraction (xCO_2) and chlorophyll-a (CHL). The influence of the above-mentioned independent variables in regulating sea-surface pCO_2 variability has been included as a proxy of different mechanisms (thermal, solubility, mixing, air-sea interaction, and biology).

Data acquisition. SST and SSS observations, along with collocated sea-surface pCO_2, are available at the locations shown in Fig. 1. We obtain the synthesised SST, SSS, and pCO_2 observations from SOCAT (https://www.socat.info/index.php/data-access/)[22] and other locations shown in Fig. 1. Other than the observations at SOCAT and RAMA buoy locations, the available observations are addressed here as SAS (Sridevi and Sarma) data[28]. The data collection and quality control methods are elaborately available in the literature corresponding to each of these data[22,28]. The monthly data frequency of collocated SST, SSS, and pCO_2 from various sources is shown in Fig. 2. The maximum number of observations is sourced from SOCAT (Fig. 2b), but it does not uniformly cover all the months. Specifically, in the open ocean SOCAT and SAS data, the observations are unavailable for the winter monsoon season (Dec, Jan, Feb). But these data provide a very good spatial coverage in other seasons. Further, the winter monsoon season observations are available from two sources: firstly, from the RAMA mooring and secondly, from the coastal transects of SAS data (as shown in Fig. 1). All ship-based observations (available in the SOCAT database) from 1991 to 2020 were acquired for this study. In the SAS data, the observations were available from 1991 to 2019.

CHL concentration is not available in SOCAT and SAS (except in a few locations) database; hence, we use a merged satellite product OC-CCI (Ocean Color Climate Change Initiative, available at https://climate.esa.int/en/projects/ocean-colour/data/)[31]. This data has excellent spatial (1/12°) and temporal coverage (1997–2020). We extract collocated monthly CHL concentrations from OC-CCI at the available observation locations. Like CHL, MLD data cannot be obtained from SOCAT and SAS since temperature and salinity depth profiles are unavailable. So we obtain MLD from GLORYS12V1 product, which is a CMEMS eddy-resolving reanalysis product (data available at https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_001_030/download). The MLD product has a spatial resolution of 1/12°, and the data is available from 1993–2020. The xCO_2 is obtained from CAMS CO_2 atmospheric inversion product[32–34] (https://atmosphere.copernicus.eu/). The xCO_2 data has spatial coverage of 0.25° and is available from 1985–2020. We use the nearest-neighbor interpolation method to find collocated data at the available sea-surface pCO_2 observation locations.

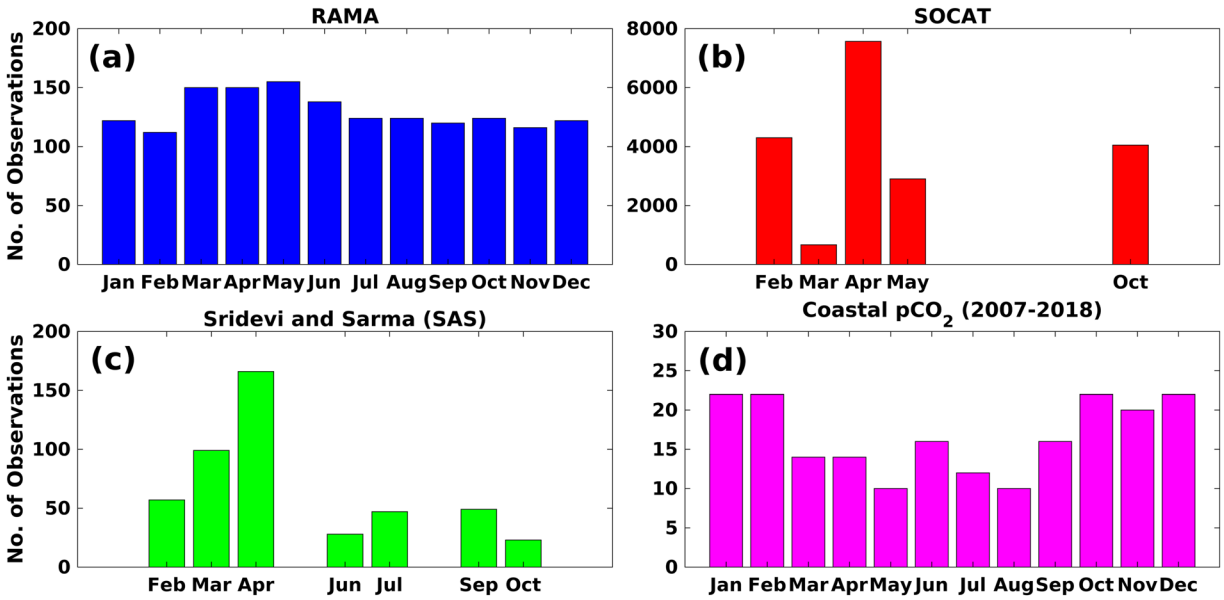

Fig. 2 Monthly observations of SST, SSS, and pCO_2 were acquired from various sources. The RAMA buoy (**a**) provides the sea-surface pCO_2 observations between November 2013 to December 2018. All ship-based observations (available in the SOCAT database) from 1991 to 2020 were used in this study (**b**). Further, additional ship-based observations available from 1991 to 2019 (denoted here as SAS data) were also included (**c**). The availability of pCO_2 data from coastal transects from 2007 to 2018 is shown in (**d**).

We checked the data distribution before using these data for training and predictions. The MLD and CHL data are converted to normal by taking their log transformation. Since ML models are known to be sensitive to outliers ($> 3\,\sigma$), these are removed from the available data.

Splitting and scaling data. To avoid data leakage, using a train-test split from the Scikit-Learn module, we divided the data into train-set (80%) and test-set (20%)[35]. The same training and test data are used for all ML models used in this study, which gives an advantage in testing model performance with respect to common test data. The K-fold (10 K-folds) technique is utilized for training each model, which aids in circumventing the over-training issue.

The data is then scaled using the StandardScaler method from the pre-processing libraries of the scikit learn. Scaling converts all the data between the range −1 to 1 with a mean of zero and a standard deviation of one. This process, called standardization, simplifies learning new things for ML models.

Models. The study tests four advanced ML algorithms, and the best among the four is used to create sea-surface pCO_2 maps for the BoB. The description of each of these algorithms is as follows:

- **Multiple Linear Regression (MLR)**
  Multiple linear regression is an analysis that builds the output variables from the input variables. The approach attempts to link the response and interpretation variables linearly. It extends the traditional least square strategy because it considers numerous pertinent variables.
  The use of multiple linear regression is evident and well-established for different applications in the literature[27–29]. It is to be noted that advanced ML can only be used if a significant number of observations are available. The multiple linear regression equation to predict sea-surface pCO_2 is as follows:

$$pCO_2 = 365.94 + 11.92 \times \mathrm{SST} + 7.45 \times \mathrm{SSS} - 1.23 \times \log(\mathrm{CHL}) + 0.86 \times \log(\mathrm{MLD}) + 19.29 \times \mathrm{xCO_2} \quad (1)$$

- **Artificial Neural Network (ANN)**
  The Artificial Neural Network (ANN) is a part of artificial intelligence based on the biological neural system. It has become common practise to establish the pCO_2 for regional scales[30,36–39]. The ANN comprises interconnected neurons that interpret incoming data like how the human brain learns. Each connection's signals are absolute values, and each neuron's output is calculated as the sum of its inputs, a nonlinear function. The edges are another name for the physical link that exists between neurons. Weights are allocated to the neurons and edges, and they self-adjust to get the best results. An input, an output, and at least one hidden layer compose an ANN. The neurons in the input layer equal the number of input parameters (independent variables) as the input layer is linked to the input data. Similarly, the output layer's neurons match the number of dependent variables. A signal can go through numerous hidden layers comprising several neurons, from the input to the output layer. The hidden layer's main objective is to establish a link between the output and input variables.

Hidden Layer	Number of neurons
Layer 1	32
Layer 2	24
Layer 3	64
Layer 4	44
Layer 5	80
Layer 6	24
Layer 7	22
Layer 8	30
Layer 9	42
Layer 10	78
Layer 11	34
Layer 12	72
Layer 13	38
Layer 14	72
Layer 15	50
Layer 16	62
Layer 17	46
Layer 18	26

Table 1. Neurons in each hidden layer.

Hyper-parameters	Range or Options	Optimized Value
lambda	0–1.0	0.8634
alpha	0-1	0.2574
subsample	0-1	0.6920
Booster	gbtree/gblinear/dart	gbtree
colsample_bytree	0-1	0.6460
max_depth	10–100 (step = 1)	93
min_child_weight	1–100	36
learning_rate	1 × e-08-1	0.0001
gamma	1 × e-08-1	5.546 × e-07
n_estimators	100–150 (step = 1)	131
grow_policy	depthwise/lossguide	lossguide

Table 2. Optimized values of the XGB hyper-parameters.

The ANN hyper-parameters are tuned using KerasTuner[40] class from the Keras library. Rectified Linear Unit (ReLU)[41] activation function is used for the hidden layers and the Linear activation function for the output layer. The network is optimized using the Adam optimizer[42]. The loss function, Mean Absolute Error, is employed and must be minimized. Two executions per trial are allowed with the parameters set for 100 trials. There are 18 hidden layers in the optimized ANN used in this study. Table 1 displays the neurons associated with each hidden layer. The model operates most well at a 0.0001 learning rate.

- **Xtreme Gradient Boosting (XGB)**
  Extreme Gradient Boosting (XGB)[43] is one of the members of the family of boosting algorithms built on decision trees. The gradient-boosted algorithm's performance and computational speed were both expanded to produce the XGBoost algorithm. Since it performed well for the central BoB region[30], the model's great speed and accuracy motivate us to compare its performance to that of other models. Only the residuals are supplied to the following weaker learners once the trees or vulnerable learners have been added in sequential order. This method helps to cut down on errors. Contrary to gradient descent, the Newton boosting based on the Newton Raphson method accelerates the approach to global minima.
  Similar to the ANN, the XGBoost model also has tunable hyperparameters. Following previous literature[30], we employ the Optuna optimization framework[44] to fine-tune the hyper-parameters. At https://xgboost.readthedocs.io/en/stable/parameter.html one may find the description for each of the XGB hyper-parameters. The hyper-parameters range and final optimized values are shown in Table 2.
- **Random Forest(RF)**
  As XGB belongs to a family of boosting algorithms, Random Forest (RF)[45] belongs to a family of bagging algorithms. RF is also built on decision trees. RF uses with-replacement random samples from the training data to generate decision trees, and the results of these decision trees are averaged to get the final output. The combined output from several trees tends to smooth out the volatility between trees and improves the ability to generalize the model as a whole. One appealing aspect of RF is its ability to estimate error using

Hyper-parameters	Range	Optimized Value
min_samples_split	2–150	17
min_samples_leaf	1–100	11
max_depth	4–100	27
n_estimators	10–2000	355

Table 3. Optimized values of the RF hyper-parameters.

out-of-bag error estimates without needing a set-aside testing dataset[46]. Like the ANN and XGB, RF also had tunable hyper-parameters optimized using the Optuna optimization framework. The list of the range and optimized hyper-parameter are provided in Table 3.

Mapping method. After selecting the best algorithm from the four algorithms described in the previous section, we employ the best algorithm to build spatial maps. To build these maps, we select SST and SSS from different products, and the rest of the input variables are chosen from the same data used for acquiring collocated data at SOCAT cruise locations. The SST is taken from the GLORYS12V1 product, which is a CMEMS eddy-resolving reanalysis product (data available at https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_001_030). The SST has a spatial resolution of 1/12° and is available from 1993 to 2020. We obtained the SSS from ESA-CCI (ESA stands for European Space Agency, and CCI stands for Climate Change Initiative), a merged product of three satellite data (SMOS, Aquarius, and SMAP). This ESA-CCI (having a spatial resolution of 0.25°) is reported to perform excellently for the BoB region[47]. The ESA-CCI SSS is available at https://catalogue.ceda.ac.uk/uuid/fad2e982a59d44788eda09e3c67ed7d5.

Since ESA-CCI is available only for the period from 2010–2020, we predict sea-surface pCO_2 for the previous decade (2010–2020) and then average it to form a climatology. The mean of the common period (2010–2020) is centered around 2015. Thus, 2015 is the climatological reference year for the INCOIS-ReML sea-surface pCO_2 climatology. The reason for making climatology is to reduce the uncertainty caused by extreme events. All the independent data are interpolated to 1/12° resolution (same as SST, CHL, and MLD) and provided to the model for prediction. Further, we compare our product with the climatology produced by averaging pCO_2 of RAMA (which is available from November 2013 to November 2018) and the gridded SOCAT data (having a spatial resolution of 1° and temporally available from 2010 to 2020). This product is expected to help in evaluating high-resolution bio-physical model simulated ocean carbon cycle as only a limited number of spatial pCO_2 observations are available in the BoB across different time scales.

CO_2 flux calculation. After preparing the climatological sea-surface pCO_2 for the BoB region, we calculate air-sea CO_2 flux to examine the sink and source regions of the BoB. The flux is calculated using the following equation.

$$CO_2 \text{ flux} = kw \times L \times \Delta pCO_2 \tag{2}$$

where kw is the piston velocity calculated as a function of wind speed[48]. We use ERA5[49] winds (https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5) to calculate the piston velocity[50]. L represents solubility of CO_2[51], and $\Delta\, pCO_2$ is the difference between sea-surface pCO_2 and atmospheric pCO_2.

Data Records

The high-resolution sea-surface pCO_2 maps and associated CO_2 flux data produced for the BoB (reported in this paper) could be accessed from https://zenodo.org/record/8375320. [52] The dataset contains two products, the first being sea-surface pCO_2 and the second being air-sea CO_2 flux for the BoB region. It is a monthly climatological data. Each of these data has a spatial resolution of 1/12°. A positive value of CO_2 flux indicates outgassing of CO_2, and the negative value shows uptake of atmospheric CO_2.

Technical Validation

In this study, we use the Taylor diagram representation[53] to evaluate the performance of the models. The Taylor diagram provides a summarized graphical view of the model performance with respect to the available observation data. Three statistics, namely Correlation Coefficient (r), Standard Deviation (STD), and Centred Root Mean Square Difference (CRMSD), are used to create the Taylor Diagram. The correlation coefficient ranges between −1 and 1; higher negative or positive values represent a strong inverse or in-sync relation between prediction and observation. Ideally, the STD of predicted values should be the same as observed, and lower CRMSD represents better model performance.

Model selection. Figure 3 represents the performance of all four models against a common test data. The performance of multiple linear regression is the worst, whereas the ANN, RF, and XGB perform almost closely to each other. The CRMSD (centered root-mean-square difference) of ANN, XGB, and RF is 6.26, 4.52, and 5.71 μatm, respectively. At the same time, the correlation of ANN, XGB, and RF is, respectively, 0.978, 0.988, and 0.982. Based on the statistics, XGB seems to have a slight edge over the other two ML models. The STD of the test data is 30.38 μatm , and all three models (ANN, XGB, and RF) are very close to this STD. Hence, from Fig. 3, it is clear that the XGB performs best among the four ML models chosen in this study. Thus, we employ the XGB model to build sea-surface pCO_2 maps for the BoB. Henceforth, we refer to the XGB-based climatological data

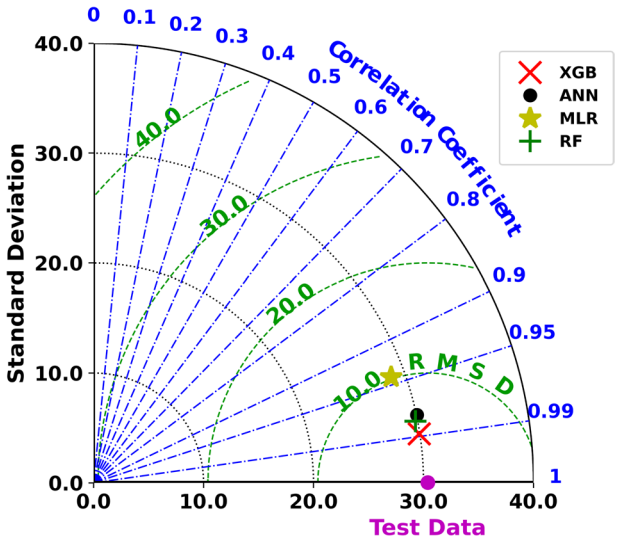

Fig. 3 Comparison of model performance with respect to the test data.

product as INCOIS-ReML (Indian National Centre for Ocean Information Services-Regional Machine Learning model).

Creating sea-surface pCO_2 maps. INCOIS-ReML is a high-resolution monthly climatological data product (Fig. 4). The temporal evolution of the INCOIS-ReML pCO_2 climatology has been compared with BOBOA mooring-based pCO_2 climatology (averaging over the available observation from 2014–2018) using correlation, root mean square error (RMSE), and Willmott skill score (WSS)[54]. The monthly variability of sea-surface pCO_2 is satisfactorily captured by the INCOIS-ReML (correlation (r) = 0.93; Fig. 5). This comparison shows that INCOIS-ReML underestimates the sea-surface pCO_2 (particularly in April and May). However, the RMSE between the observed and modeled values is 7.40, which indicates that the error is within acceptable bounds (Fig. 5). The capability of INCOIS-ReML pCO_2 is also evident from its WSS of 0.885.

Using the available observations from BOBOA mooring (location-specific data), we validated the temporal variability of INCOIS-ReML pCO_2. However, a limited number of observations makes it difficult to validate spatial variability of INCOIS-ReML pCO. Therefore, we use observations-based gridded (1° × 1°) SOCAT product (available from the 1990s to date) to compare spatial variability of pCO_2. As a first step, we generate a climatology of SOCAT data product with reference to the year 2015 for comparison. Before comparison, we interpolate the high-resolution INCOIS-ReML data product (Fig. 6a) to the spatial resolution of SOCAT gridded data product (Fig. 6b) using the nearest-neighbor interpolation method. Here, the reader must understand that the unavailability of a sufficient number of temporally varying observations in the BoB impacts the magnitude of the sea-surface pCO_2 climatology derived from SOCAT. INCOIS-ReML satisfactorily captures the spatial pattern, i.e., lower sea-surface pCO_2 in the north and higher sea-surface pCO_2 in the south. Figure 6c,d provide spatial statistics to evaluate the performance of the INCOIS-ReML data product. A high correlation is seen in the central BoB region (Fig. 6c). A few grids show negative to low correlation in the south of the Sri Lankan coast. Figure 6d shows overestimation in the region east of 92° E, but low negative bias persists in the rest of the region. The domain average bias is approximately 0.92 μatm. The overestimation of the INCOIS-ReML can be attributed to the discontinuous time-series data from SOCAT in a large part of BoB.

We compare the INCOIS-ReML pCO_2 with the results of existing studies, carried out using *in-situ* observations, available in the literature to validate the spatial variability of pCO_2 more rigorously. The spatial monthly variation of INCOIS-ReML is shown in Fig. 4. The northern BoB (approximately above 15° N) is seen to have lower sea-surface pCO_2 than the southern BoB region[55,56]. The EICC (East India Coastal Current) is known to have dominant control over the sea-surface pCO_2, especially in the western coast of BoB[16] due to the spreading of river-influenced water along the coast. The northward-moving EICC is primarily strong from March to May when high salinity and pCO_2 levels are observed. In contrast, southward-moving EICC during October to December brings river-influenced low saline and pCO_2 water along the coast[16]. The INCOIS-ReML well reproduces the coastal pattern of pCO_2 levels with the lowest during November and the highest pCO_2 levels during May (Fig. 4). Overall, the spatial and temporal patterns are well captured by INCOIS-ReML.

Further, we compare our climatological product with six widely used ML-based pCO_2 products (listed in Table 4). Figure 7 shows that based on the Willmott skill score (WSS), INCOIS-ReML performs better than all the other six products. This is due to two primary reasons: a) the inclusion of a significant number of open and coastal ocean observations from SAS leads to an improvement in model prediction, and b) the high spatial resolution of INCOIS-ReML. Figure 7 (based on WSS) shows that CMEMS performs as good as INCOIS-ReML. Hence, we further compare the two products spatially and explain the advantages of high-resolution INCOIS-ReML (Fig. 8).

The first observation from Fig. 8 is that INCOIS-ReML is capable of capturing spatio-temporal variability of pCO_2 in the coastal waters of the BoB. Since BoB receives high freshwater flux from rivers and precipitation

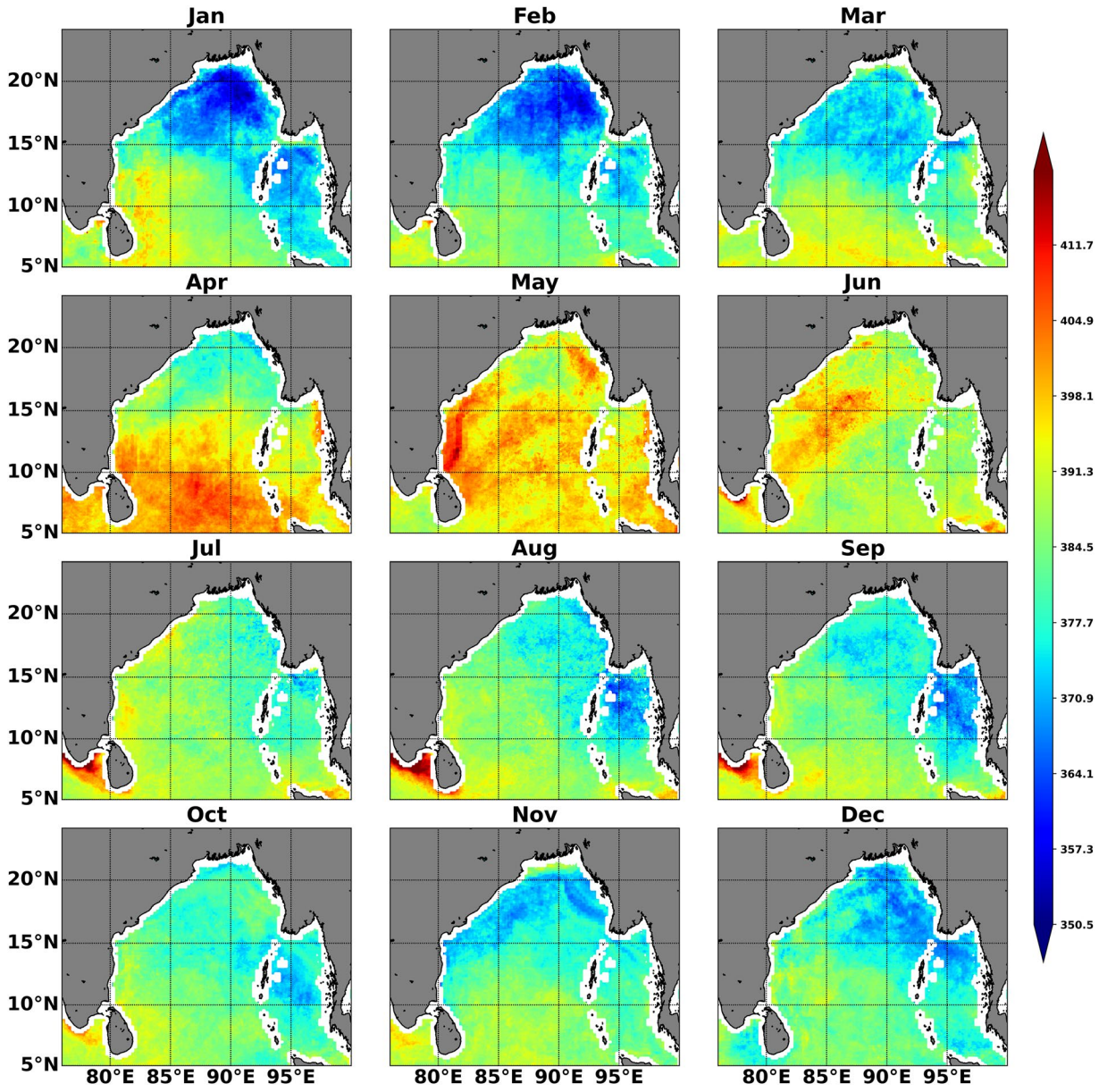

Fig. 4 Climatological monthly variability of the sea-surface pCO_2 produced by INCOIS-ReML. The climatological reference year for this dataset is 2015.

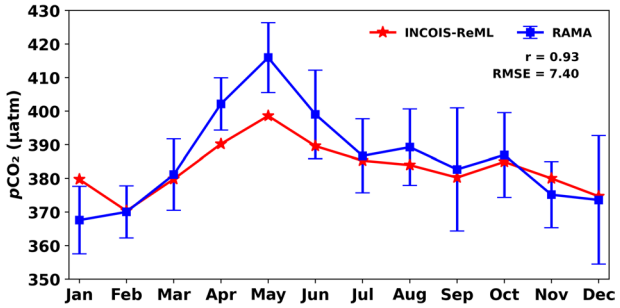

Fig. 5 Climatological monthly variability of the sea-surface pCO_2 produced by INCOIS-ReML is compared with the climatology created by RAMA mooring buoy. The climatological reference year for this dataset is 2015.

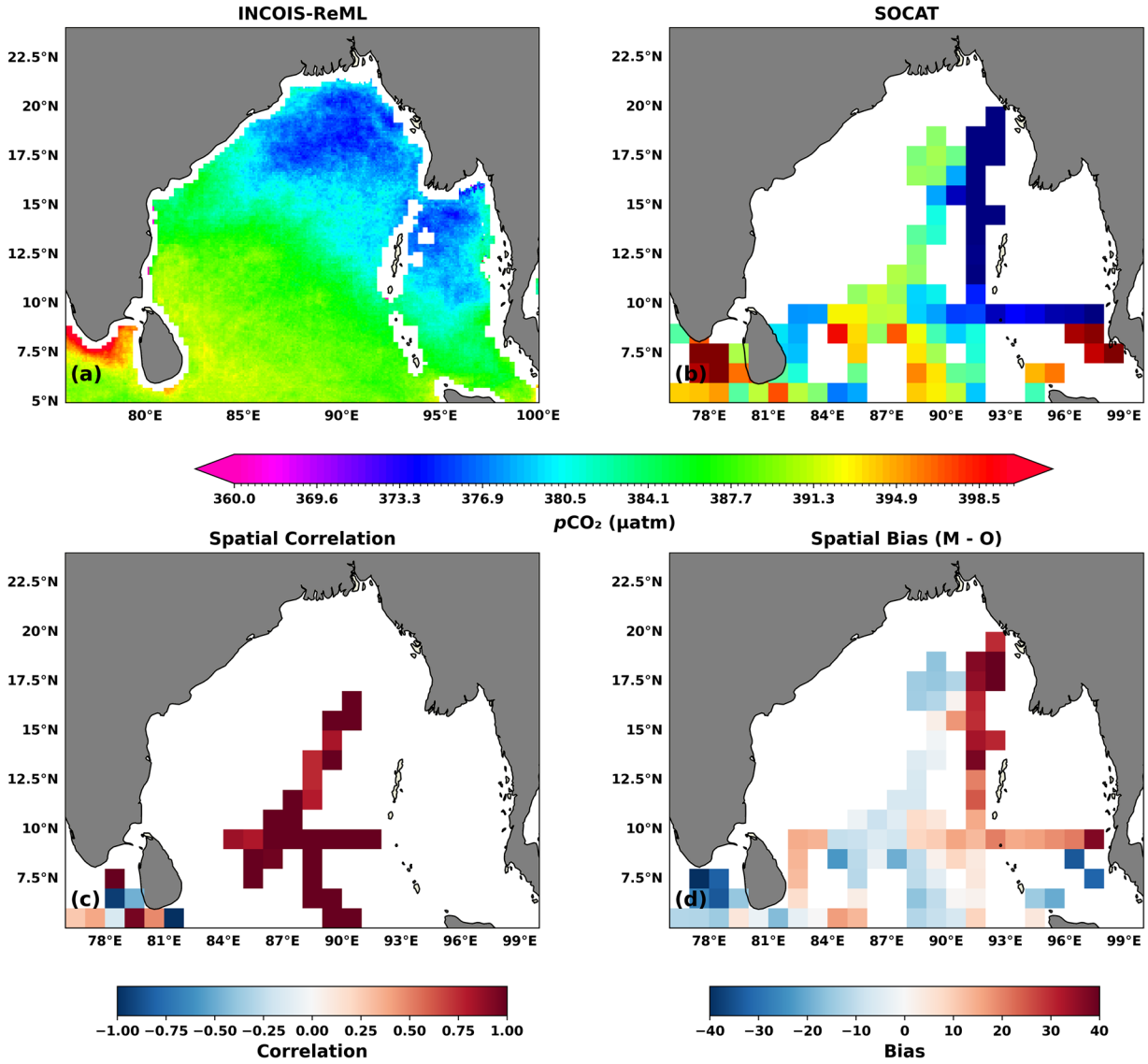

Fig. 6 Comparison (annual mean of the climatological year) between the (**a**) INCOIS-ReML produced sea-surface pCO_2 and (**b**) SOCAT. The spatial correlation and spatial bias (difference (Model - Observation (M-O)) in an annual mean of the climatological year) are shown in figures (**c**) and (**d**). The climatological reference year for this dataset is 2015.

Abbreviations	Full Form
CMEMS	CMEMS-LSCE-FFNN[25]
LAND	SOMFNN[62]
SODA	OceanSODAETHZ[63]
LDEO_HPD	$SpCO_2$_LDEO_HPD[64]
JMA	JMAMLR[65]
CSIR	CSIRML6[26]

Table 4. List of ML-based models with which we compare INCOIS-ReML.

during the southwest monsoon (June-September), low salinity water is found in the north that spreads to the south by monsoon currents[57]. This freshwater plume spreads to the BoB by fall monsoon (ON)0[15,58,59]. This plume first spreads in the eastern Bay, followed by the western Bay, with minimal impact on freshwater during spring inter-monsoon (March to May). CMEMS and INCOIS-ReML performed well in capturing spatial variations of low pCO_2 primarily driven by the low saline waters in the BoB. However, the spatial variations were not well captured by CMEMS compared to INCOIS-ReML during spring monsoon (MAM). Perennial occurrence

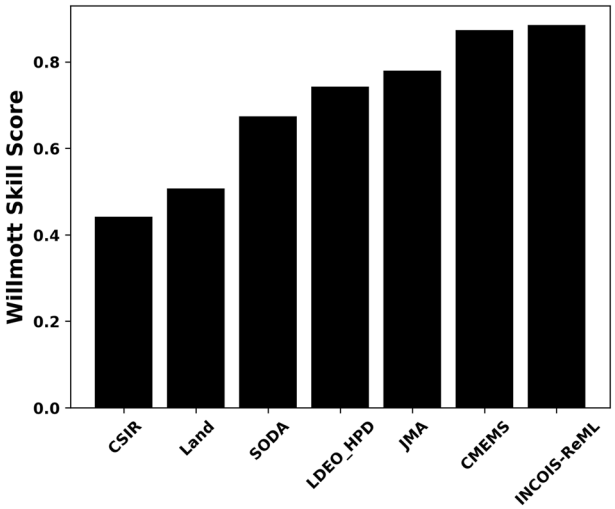

Fig. 7 Willmott Skill Score of the comparison between INCOIS-ReML and other six widely used ML-based products' climatological pCO_2 with BOBOA based climatological pCO_2 observations.

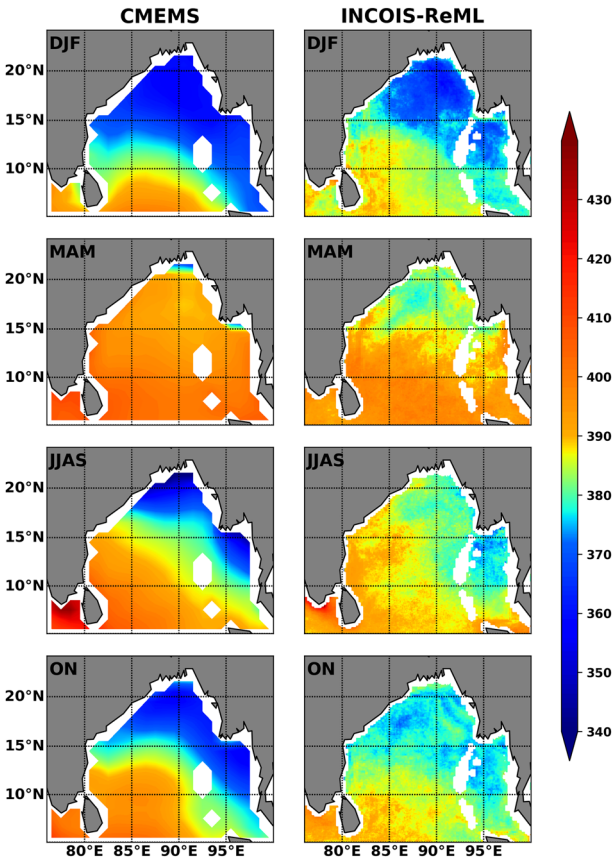

Fig. 8 Seasonal spatial comparison of pCO_2 between CMEMS and INCOIS-ReML.

of low pCO_2 due to low salinity during the summer monsoon season was reported in the northern BoB[60], that was not well captured by CMEMS (Fig. 8). In addition, the pCO_2 levels in the low salinity plume region were underestimated by CMEMS compared to INCOIS-ReML[61]. The presence of low-saline freshwater and associated strong stratification lower the sea-surface pCO_2 values in the northern BoB[16]. These physical processes play a significant role in regulating the seasonality of sea-surface pCO_2 in the BoB[15,17,19]. It is evident that the seasonality of sea-surface pCO_2 is well captured by the INCOIS-ReML. Therefore, the high resolution INCOIS-ReML data product is an improved version of the climatological mean state of sea-surface pCO_2 in the BoB region.

Hence, we provide a high-resolution sea-surface pCO_2 maps and associated air-sea CO_2 flux (calculated using the equation mentioned in the earlier section) data product, which would immensely aid in validating not only high-resolution bio-physical model simulated ocean carbon cycle but also coarser-resolution CMIP6 models. Further, it is worth mentioning that the inclusion of SAS data makes it possible for this high-resolution product to capture the coastal pCO_2 dynamics better, which is missing in other observation-based data products. We understand that the product can still be improved, and we will keep on updating the product as the number of observations increases. This product is expected to be extremely helpful in validating models (especially spatial variability) used to understand the future scenarios of the sea-surface pCO_2 in the BoB.

Code availability

The code used to create the final product (different machine learning models) is available at https://github.com/APJ1812/INCOIS_pCO2. The study uses general machine learning codes available in Python.

Received: 2 November 2023; Accepted: 8 April 2024;
Published online: 13 April 2024

References

1. Friedlingstein, P. *et al*. Global carbon budget 2020. *Earth System Science Data* **12**, 3269–3340 (2020).
2. Friedlingstein, P. *et al*. Global carbon budget 2021. *Earth System Science Data Discussions* 1–191 (2021).
3. Friedlingstein, P. *et al*. Global carbon budget 2022. *Earth System Science Data Discussions* **2022**, 1–159 (2022).
4. Chen, C.-T. *et al*. Air–sea exchanges of CO_2 in the world's coastal seas. *Biogeosciences* **10**, 6509–6544 (2013).
5. Laruelle, G. G., Lauerwald, R., Pfeil, B. & Regnier, P. Regionalized global budget of the CO_2 exchange at the air-water interface in continental shelf seas. *Global biogeochemical cycles* **28**, 1199–1214 (2014).
6. Laruelle, G. G. *et al*. Continental shelves as a variable but increasing global sink for atmospheric carbon dioxide. *Nature communications* **9**, 454 (2018).
7. Dai, M. *et al*. Why are some marginal seas sources of atmospheric CO_2? *Geophysical Research Letters* **40**, 2154–2158 (2013).
8. Zhai, W.-D. *et al*. Seasonal variations of the sea–air CO_2 fluxes in the largest tropical marginal sea (South China sea) based on multiple-year underway measurements. *Biogeosciences* **10**, 7775–7791 (2013).
9. Li, Q., Guo, X., Zhai, W., Xu, Y. & Dai, M. Partial pressure of CO_2 and air-sea CO_2 fluxes in the South China sea: Synthesis of an 18-year dataset. *Progress in Oceanography* **182**, 102272 (2020).
10. Borges, A. V. Do we have enough pieces of the jigsaw to integrate CO_2 fluxes in the coastal ocean? *Estuaries* **28**, 3–27 (2005).
11. Anderson, T. R. Plankton functional type modelling: running before we can walk? *Journal of Plankton Research* **27**, 1073–1081 (2005).
12. Anderson, T. R. Progress in marine ecosystem modelling and the "unreasonable effectiveness of mathematics". *Journal of Marine Systems* **81**, 4–11 (2010).
13. Sarma, V., Krishna, M. & Srinivas, T. Sources of organic matter and tracing of nutrient pollution in the coastal Bay of Bengal. *Marine Pollution Bulletin* **159**, 111477 (2020).
14. Sarma, V., Prasad, M. & Dalabehera, H. Influence of phytoplankton pigment composition and primary production on pCO_2 levels in the Indian ocean. *Journal of Earth System Science* **130**, 1–16 (2021).
15. Joshi, A., Chowdhury, R. R., Warrior, H. & Kumar, V. Influence of the freshwater plume dynamics and the barrier layer thickness on the CO_2 source and sink characteristics of the Bay of Bengal. *Marine Chemistry* **236**, 104030 (2021).
16. Sarma, V. *et al*. East India coastal current controls the Dissolved Inorganic Carbon in the coastal Bay of Bengal. *Marine Chemistry* **205**, 37–47 (2018).
17. Joshi, A., Roychowdhury, R., Kumar, V. & Warrior, H. Configuration and skill assessment of the coupled biogeochemical model for the carbonate system in the Bay of Bengal. *Marine Chemistry* 103871 (2020).
18. Joshi, A. & Warrior, H. Comprehending the role of different mechanisms and drivers affecting the sea-surface pCO_2 and the air-sea CO_2 fluxes in the Bay of Bengal: A modelling study. *Marine Chemistry* **243**, 104120 (2022).
19. Chakraborty, K., Valsala, V., Bhattacharya, T. & Ghosh, J. Seasonal cycle of surface ocean pCO_2 and pH in the northern Indian ocean and their controlling factors. *Progress in Oceanography* **198**, 102683 (2021).
20. Chakraborty, K., Valsala, V., Gupta, G. & Sarma, V. Dominant biological control over upwelling on pCO_2 in sea east of sri lanka. *Journal of Geophysical Research: Biogeosciences* **123**, 3250–3261 (2018).
21. Sutton, A. J. *et al*. A high-frequency atmospheric and seawater pCO_2 data set from 14 open-ocean sites using a moored autonomous system. *Earth System Science Data* **6**, 353–366 (2014).
22. Bakker, D. C. *et al*. Surface ocean CO_2 atlas database version 2022 (SOCATv2022)(ncei accession 0253659). *Earth System Science Data* (2022).
23. Lauvset, S. K. *et al*. GLODAPv2. 2022: the latest version of the global interior ocean biogeochemical data product. *Earth System Science Data Discussions* **2022**, 1–37 (2022).
24. Takahashi, T. *et al*. Climatological distributions of pH, pCO_2, total CO_2, alkalinity, and CaCO3 saturation in the global surface ocean, and temporal changes at selected locations. *Marine Chemistry* **164**, 95–125 (2014).
25. Chau, T. T. T., Gehlen, M. & Chevallier, F. A seamless ensemble-based reconstruction of surface ocean pCO_2 and air–sea CO_2 fluxes over the global coastal and open oceans. *Biogeosciences* **19**, 1087–1109 (2022).
26. Gregor, L., Lebehot, A. D., Kok, S. & Scheel Monteiro, P. M. A comparative assessment of the uncertainties of global surface ocean CO_2 estimates using a machine-learning ensemble (csir-ml6 version 2019a)–have we hit the wall? *Geoscientific Model Development* **12**, 5113–5136 (2019).
27. Dixit, A., Lekshmi, K., Bharti, R. & Mahanta, C. Net sea–air CO_2 fluxes and modeled partial pressure of CO_2 in open ocean of Bay of Bengal. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **12**, 2462–2469 (2019).
28. Sridevi, B. & Sarma, V. Role of river discharge and warming on ocean acidification and pCO_2 levels in the Bay of Bengal. *Tellus B: Chemical and Physical Meteorology* **73**, 1–20 (2021).
29. Mohanty, S., Raman, M., Mitra, D. & Chauhan, P. Surface pCO_2 variability in two contrasting basins of north Indian ocean using satellite data. *Deep Sea Research Part I: Oceanographic Research Papers* **179**, 103665 (2022).
30. Joshi, A., Kumar, V. & Warrior, H. Modeling the sea-surface pCO_2 of the central Bay of Bengal region using machine learning algorithms. *Ocean Modelling* **178**, 102094 (2022).
31. Sathyendranath, S. *et al*. An ocean-colour time series for use in climate studies: the experience of the ocean-colour climate change initiative (oc-cci). *Sensors* **19**, 4285 (2019).
32. Chevallier, F. *et al*. Inferring CO_2 sources and sinks from satellite observations: Method and application to tovs data. *Journal of Geophysical Research: Atmospheres* **110** (2005).
33. Chevallier, F. *et al*. CO_2 surface fluxes at grid point scale estimated from a global 21 year reanalysis of atmospheric measurements. *Journal of Geophysical Research: Atmospheres* **115** (2010).

34. Chevallier, F. On the parallelization of atmospheric inversions of CO_2 surface fluxes within a variational framework. *Geoscientific Model Development* **6**, 783–790 (2013).
35. Pedregosa, F. *et al*. Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).
36. Friedrich, T. & Oschlies, A. Neural network-based estimates of north Atlantic surface pCO_2 from satellite data: A methodological study. *Journal of Geophysical Research: Oceans* **114** (2009).
37. Jo, Y.-H., Dai, M., Zhai, W., Yan, X.-H. & Shang, S. On the variations of sea surface pCO_2 in the northern South China sea: A remote sensing based neural network approach. *Journal of Geophysical Research: Oceans* **117** (2012).
38. Moussa, H., Benallal, M., Goyet, C. & Lefèvre, N. Satellite-derived CO_2 fugacity in surface seawater of the tropical atlantic ocean using a feedforward neural network. *International Journal of Remote Sensing* **37**, 580–598 (2016).
39. Wang, Y. *et al*. Carbon sinks and variations of pCO_2 in the southern ocean from 1998 to 2018 based on a deep learning approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**, 3495–3503 (2021).
40. O'Malley, T. *et al*. Keras tuner. *Retrieved May* **21**, 2020 (2019).
41. Agarap, A. F. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).
42. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *Anon. International Conference on Learning Representations. SanDego: ICLR 7* (2015).
43. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).
44. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631 (2019).
45. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).
46. Lawrence, R. L., Wood, S. D. & Sheley, R. L. Mapping invasive plants using hyperspectral imagery and breiman cutler classifications (randomforest). *Remote Sensing of Environment* **100**, 356–362 (2006).
47. Akhil, V. P. *et al*. Bay of Bengal sea surface salinity variability using a decade of improved smos re-processing. *Remote Sensing of Environment* **248**, 111964 (2020).
48. Wanninkhof, R. Relationship between wind speed and gas exchange over the ocean. *Journal of Geophysical Research: Oceans* **97**, 7373–7382 (1992).
49. Hersbach, H. *et al*. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* **146**, 1999–2049 (2020).
50. Wanninkhof, R. Relationship between wind speed and gas exchange over the ocean revisited. *Limnology and Oceanography: Methods* **12**, 351–362 (2014).
51. Weiss, R. Carbon dioxide in water and seawater: the solubility of a non-ideal gas. *Marine chemistry* **2**, 203–215 (1974).
52. Joshi, A., Ghoshal, K., Prasanna, Chakraborty, K. & Sarma, V. Sea-surface pCO_2 maps for the Bay of Bengal based on machine learning algorithms. *Zenodo* https://doi.org/10.5281/zenodo.8375320 (2024).
53. Taylor, K. E. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres* **106**, 7183–7192 (2001).
54. Willmott, C. J. On the validation of models. *Physical geography* **2**, 184–194 (1981).
55. Sabine, C., Wanninkhof, R., Key, R., Goyet, C. & Millero, F. Seasonal CO_2 fluxes in the tropical and subtropical Indian ocean. *Marine Chemistry* **72**, 33–53 (2000).
56. Bates, N. R., Pequignet, A. C. & Sabine, C. L. Ocean carbon cycling in the Indian ocean: 1. spatiotemporal variability of inorganic carbon and air-sea CO_2 gas exchange. *Global Biogeochemical Cycles* **20** (2006).
57. Schott, F. A. & McCreary, J. P. Jr The monsoon circulation of the Indian ocean. *Progress in Oceanography* **51**, 1–123 (2001).
58. Jana, S., Gangopadhyay, A. & Chakraborty, A. Impact of seasonal river input on the Bay of Bengal simulation. *Continental Shelf Research* **104**, 45–62 (2015).
59. Jana, S. *et al*. Sensitivity of the Bay of Bengal upper ocean to different winds and river input conditions. *Journal of Marine Systems* **187**, 206–222 (2018).
60. Sarma, V., Krishna, M., Paul, Y. & Murty, V. Observed changes in ocean acidity and carbon dioxide exchange in the coastal Bay of Bengal–a link to air pollution. *Tellus B: Chemical and Physical Meteorology* **67**, 24638 (2015).
61. Sarma, V. *et al*. Impact of eddies on dissolved inorganic carbon components in the Bay of Bengal. *Deep Sea Research Part I: Oceanographic Research Papers* **147**, 111–120 (2019).
62. Landschützer, P., Gruber, N. & Bakker, D. C. Decadal variations and trends of the global ocean carbon sink. *Global Biogeochemical Cycles* **30**, 1396–1417 (2016).
63. Gregor, L. & Gruber, N. OceanSODA-ETHZ: a global gridded data set of the surface ocean carbonate system for seasonal to decadal studies of ocean acidification. *Earth System Science Data* **13**, 777–808 (2021).
64. Gloege, L., Yan, M., Zheng, T. & McKinley, G. A. Improved quantification of ocean carbon uptake by using machine learning to merge global models and pCO_2 data. *Journal of Advances in Modeling Earth Systems* **14**, e2021MS002620 (2022).
65. Iida, Y., Takatani, Y., Kojima, A. & Ishii, M. Global trends of ocean CO_2 sink and ocean acidification: an observation-based reconstruction of surface ocean inorganic carbon variables. *Journal of Oceanography* **77**, 323–358 (2021).

Acknowledgements

We are grateful to the anonymous reviewers for their careful reading, constructive comments, and helpful suggestions, which have helped us to significantly improve the presentation of this work. INCOIS-ReML data product has been developed as a part of the 'Development of Climate Change Advisory Services' project of the Indian National Centre for Ocean Information Services, Hyderabad, India, under the 'Deep Ocean Mission' programme of the Ministry of Earth Sciences (MoES), Govt. of India. The Surface Ocean CO_2 Atlas (SOCAT) is an international effort endorsed by the International Ocean Carbon Coordination Project (IOCCP), the Surface Ocean Lower Atmosphere Study (SOLAS), and the Integrated Marine Biosphere Research (IMBeR) program to deliver a uniformly quality-controlled surface ocean CO_2 database. The many researchers and funding agencies responsible for collecting data and quality control are thanked for their contributions to SOCAT. Sincere gratitude is extended to the scientists, funding organizations, and SOCAT data collection and quality-control process organizers. The field programs for making ship-based observations (presented in this paper as SAS data) were funded by several Indian funding agencies (Ministry of Earth Sciences, Ministry of Science and Technology, Department of Space) of the Govt. of India. The authors acknowledge the efforts of scientists towards developing OC-CCI data. This is INCOIS contribution number 519.

Author contributions

A.P. Joshi: Methodology, Investigation, Validation, Formal Analysis, Writing - Original Draft; Prasanna Kanti Ghoshal: Data Curation, Methodology, Software, Investigation; Kunal Chakraborty: Conceptualization, Formal Analysis, Visualization, Resources, Writing - Review and Editing, Supervision; V.V.S.S. Sarma: Data Curation, Writing - Review and Editing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.