Check for updates

ADVANCING EARTH AND SPACE SCIENCES

Earth and Space Science

RESEARCH ARTICLE
10.1029/2023EA003230

Key Points:

- A random forest based regression ensemble model is developed to estimate the sea surface salinity for the East China Sea from MODIS products
- The model is evaluated using an independent data set from the East China Sea and its sensitivity to input parameter errors are quantified
- The model is successfully applied in the Chinese marginal seas, indicating it is an efficient tool to monitor the spatiotemporal variations in SSS

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

R. G. J. Bellerby and Q. Zhu,
Richard.Bellerby@niva.no;
qz@xmu.edu.cn

Citation:

Liu, J., Bellerby, R. G. J., Zhu, Q., & Ge, J. (2023). Estimating sea surface salinity in the East China Sea using satellite remote sensing and machine learning. *Earth and Space Science*, *10*, e2023EA003230. https://doi.org/10.1029/2023EA003230

Received 8 AUG 2023
Accepted 10 OCT 2023

Author Contributions:

Conceptualization: Jing Liu, Richard G. J. Bellerby
Data curation: Qing Zhu, Jianzhong Ge
Funding acquisition: Richard G. J. Bellerby
Investigation: Jing Liu, Richard G. J. Bellerby, Jianzhong Ge
Methodology: Jing Liu

Estimating Sea Surface Salinity in the East China Sea Using Satellite Remote Sensing and Machine Learning

Jing Liu[1,2], Richard G. J. Bellerby[1,2,3], Qing Zhu[4], and Jianzhong Ge[1]

[1]State Key Laboratory of Estuarine and Coastal Research, East China Normal University, Shanghai, China, [2]Norwegian Institute for Water Research, Bergen, Norway, [3]Faculty of Applied Sciences, UCSI University, Kuala Lumpur, Malaysia, [4]State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen, China

Abstract Sea surface salinity (SSS) is a master variable in oceanography and important to understand marine biogeochemical and physical processes. In the East China Sea (ECS), a random forest based regression ensemble model (RF) was developed to estimate the SSS with a spatial resolution of ∼1 km based on a large synchronous data set of in situ SSS observations, MODIS-derived remote sensing reflectance (R_{rs}) and sea surface temperature (SST). The model showed the best performance when the $R_{rs}(412)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, SST and Julian day (JD) were used as inputs, with a root mean square error (RMSE) of 0.84, mean absolute error (MAE) of 0.31 and coefficient of determination (R^2) of 0.81 for model training ($N = 4{,}504$), and a RMSE of 0.77, MAE of 0.30 and R^2 of 0.86 for the model test ($N = 1{,}153$). The accuracy of the SSS model was examined using an independent data set during the period of 2020–2022 with a RMSE of 0.66 and MAE of 0.39 ($N = 2{,}151$). The interannual and seasonal signal of modeled SSS of the ECS, showed that important drivers of variability are the Changjiang discharge and the East-Asian monsoon. Applications of the model to other Chinese marginal seas (Yellow and Bohai seas) showed good agreement in distribution patterns when compared with the estimated SSS from NASA Soil Moisture Active Passive. Once more empirical oceanographic data is made available, this robust model can be applied to other regions retraining the model with informed local data sets.

1. Introduction

Sea surface salinity (SSS) is a crucial variable to understand ocean circulation, mixing between offshore seawater and freshwater, and influences many marine biogeochemical and physical processes (Font et al., 2010; Vandermeulen et al., 2014). Salinity, with temperature, determines the density of seawater, thus, salinity influences the thermohaline circulation and plays an important role in climate change (de Boyer Montégut et al., 2007; Zhu et al., 2014). In addition, SSS is also an essential variable to improve the accuracy of determining air-sea CO_2 fluxes, sea surface current fields, and El Niño-Southern Oscillation forecasts (Chakraborty et al., 2014; Köhl et al., 2014; Yang et al., 2010).

However, it is difficult to obtain SSS data with sufficient temporal and spatial resolution based on ship-based measurements. Here, remote sensing is being considered to use to estimate salinity, sourced from remote data from microwave satellite sensors launched to observed SSS from space (ESA Soil Moisture and Ocean Salinity (SMOS), NASA Aquarius/SAC–D, and NASA Soil Moisture Active-Passive (SMAP), launched in 2009, 2011 and 2015, respectively). Nevertheless, these satellite sensors are limited in monitoring SSS in the inshore waters as a result of the coarse spatial resolution (30–100 km), low revisit frequency (three days or longer) as well as the influence of land contamination (Bao et al., 2021; Entekhabi et al., 2010; Font et al., 2010; Koblinsky et al., 2003). Therefore, they are not considered effective tools for monitoring the SSS dynamics in the East China Sea (ECS).

In order to observe the SSS distributions continuously for the coastal waters, ocean color measurements have been widely used due to their high spatial-temporal resolution (Ahn et al., 2008; Bai et al., 2013; Chen & Hu, 2017; Choi et al., 2021; Geiger et al., 2013; Hu et al., 2003; Kim et al., 2020; Marghany & Hashim, 2011; Qing et al., 2013; Sun et al., 2019; Urquhart et al., 2012; Zhao et al., 2017). In these studies, there are two main routes to retrieve the SSS from space, developing the relationship between SSS and the colored dissolved organic matter (CDOM) absorption coefficient (a_{CDOM}, m^{-1}) and the a_{CDOM} can be estimated from ocean color measurements (Ahn et al., 2008; Bowers & Brett, 2008; Carder et al., 2003; Del Vecchio & Blough, 2004; Zhu

Project Administration: Richard G. J. Bellerby
Software: Jing Liu, Qing Zhu
Supervision: Richard G. J. Bellerby
Validation: Jing Liu, Qing Zhu
Writing – original draft: Jing Liu
Writing – review & editing: Richard G. J. Bellerby, Qing Zhu

et al., 2011); and directly developing the linear or nonlinear relationship between SSS and ocean color measurements, such as remote sensing reflectance (R_{rs}, sr^{-1}) (e.g., Chen & Hu, 2017; Kim et al., 2020; Qing et al., 2013; Sun et al., 2019). Either way, the principle is that SSS in the coastal waters can be tracked by CDOM (Del Vecchio & Blough, 2004; Hu et al., 2003) as CDOM is mainly from terrestrial sources, mostly imported through river discharge. However, CDOM has distinct local characteristics in different coastal waters, and is not a conservative variable, regulated by biogeochemical and physical processes (Bai et al., 2013). Thus, the relationship between SSS and a_{CDOM} may change seasonally and spatially. In addition, there are large uncertainties when CDOM is retrieved from $R_{\mathrm{rs}}(\lambda)$, especially in the turbid coastal waters.

Traditional regression methods, including multiple linear regression (MLR) and multiple nonlinear regression (MNR) have estimated SSS using satellite R_{rs} data in the Bohai Sea (BS), southern Yellow Sea (YS) and low-salinity water plumes in the ECS (Choi et al., 2021; Qing et al., 2013; Sun et al., 2019, respectively), and Marghany and Hashim (2011) retrieved SSS from MODIS data in the South China Sea (SCS) using a Box-Jenkins algorithm. Machine learning has also been applied to model SSS. For example, a multilayer perceptron neural network (MPNN) was used to predict the SSS using satellite-derived R_{rs} and sea surface temperature (SST) data in the northern Gulf of Mexico (Chen & Hu, 2017). To improve the predictive accuracy, a MPNN also used to monitor the hourly SSS spatial distribution around the Changjiang Estuary using the predictors, satellite-derived R_{rs} and SST and location (Kim et al., 2020). Moreover, Ahn et al. (2008) and Bai et al. (2013) developed the linear relationship between SSS and a_{CDOM} to map the Changjiang River plume. To our knowledge, too few in situ SSS data or only SMAP-derived SSS data were employed to establish the SSS model in the previous studies, and used different spectral bands although in the same region. Here we examined the traditional regression and machine learning methods using a large synchronous data set of in situ SSS and satellite ocean color products, aiming to select the best method to estimate the SSS with high spatial-temporal resolution for the whole ESC and covered all seasons.

The structure of this paper is as following: Section 2 describes the observed and satellite data and the development of random forest based regression ensemble (RF) model; Section 3 evaluates the models performance, independent validation, comparison with other method and interannual and seasonal variability of modeled SSS in the ECS; Section 4 discusses the model sensitivity, the influence of the Changjiang runoff on SSS, and application of RF model in other Chinese marginal seas. A conclusion is presented in the last section.

2. Data Sources and Methods

2.1. Study Region

The ECS is located on the western side of the Pacific Ocean and connects the YS and SCS. It is bounded by Korean Peninsula, Chinese mainland, Taiwan, and the Kyushu and Ryukyu Islands (Figure 1). The circulation of several major water masses (as shown in Figure 1) influences the hydrological environment of the ECS and forming distinctly different physical-biogeochemical sub-regions (Liu et al., 2023). The ECS inner shelf is the most heterogeneous region of the ECS as it is strongly affected by Changjiang discharge and coastal waters, receiving a large amount of freshwater, nutrients, organic matter, which is characterized by low SSS and highly active biogeochemical processes (Liu et al., 2010). The central region of the ECS shelf is affected by Taiwan Warm Current (TWC) and the mixing water between coastal water and offshore water, which is characterized by intermediate SSS, and has a strong seasonal cycle. The ECS outer shelf is influenced by the Kuroshio (KS) and is weakly influenced by material of terrestrial origin. It is featured by high SST and SSS, low nutrient concentrations and biological activity.

2.2. Data Sources

2.2.1. Field Data

We collected publicly available SSS data set in the ECS for the past 20 years (Tables 1 and 4). In Table 1 (for model development), 16 cruises were identified for the Changjiang Estuary and ECS inner shelf from March 2013, August 2013, February to March 2014, July 2014, March 2015, July 2015, March 2016, July 2016, July 2016, February 2017, May 2017, July to August 2017, March 2018, July 2018, October 2018, and July 2020. The field measured SSS profiles at each station were measured directly using a SeaBird conductivity

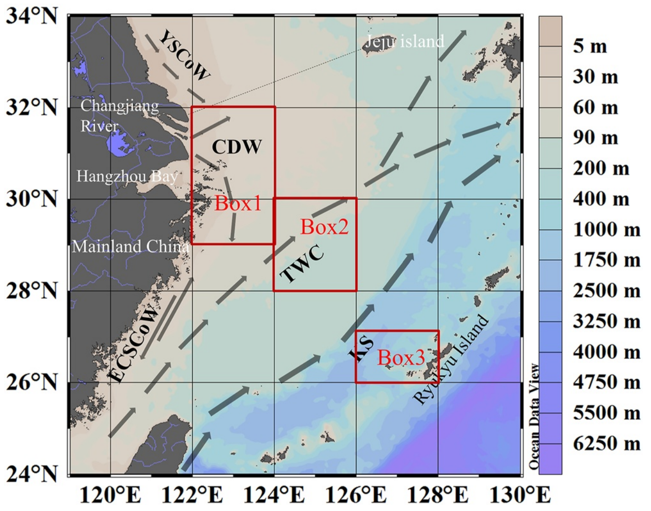

Figure 1. Diagram of the topography and major water currents of the ECS. The dark gray arrows indicate the circulations of major water masses, including Changjiang Diluted Water (CDW), Yellow Sea Coastal Water (YSCoW), East China Sea Coastal Water (ECSCoW), Taiwan Warm Current (TWC) and Kuroshio (KS) (Chen, 2009; Qi et al., 2014). Three subregions were selected for this study to highlight contrasting regimes of the ECS. Box 1 (29–32°N, 122–124°E) is near the Changjiang River Estuary, Box 2 (28–30°N, 124–126°E) is in the central ECS, Box 3 (26–27°N, 126–128°E) is in the ECS open water.

temperature-depth/pressure (CTD) recorders (SBE 25plus or 911plus). A cruise was conducted during "Vulnerabilities and Opportunities of the Coastal Ocean (VOCO)" on the ECS inner shelf during 12–24 May 2017. The discrete field SSS data were measured using a SeaBird CTD recorders (SBE 25plus). One additional cruise data set of the ECS was from Xiong et al. (2019) (these cruises were conducted from December 2005 to July 2017, for a total of 17 cruises, https://doi.org/10.6084/m9.figshare.9768215) and Xiong et al. (2020) (a cruise carried out on October 2018, https://doi.org/10.6084/m9.figshare.12630335). The discrete SSS data was measured with a calibrated WTW's TetrCon®925 probe or using a SeaBird CTD recorders (SBE 911plus). Additionally, two cruises were carried out for the whole ECS from 13 to 20 May 2018 and 14 to 24 May 2019. The field measured SSS profiles were obtained directly using a SeaBird CTD recorders (SBE 911plus).

In addition, field SSS data were obtained from Surface Ocean CO_2 Atlas (SOCAT) version 2022 from May 2003 to October 2019 (Bakker et al., 2022, https://www.socat.info/index.php/data-access/) and World Ocean Database (WOD) 2018 from February 2003 to June 2020 (Boyer et al., 2018, https://www.ncei.noaa.gov/products/world-ocean-database). Note that the field SSS data from SOCAT was ship-based underway data, but the SSS data from WOD was discrete data including CTD recorded data and fixed-location Argos. The spatial distribution of all the field-measured SSS data that were used for developing the SSS model is shown in Figure 2a with cover all seasons.

The field-measured SSS were used as an independent data set to evaluate the SSS model skill (not included in model development) is presented in Table 4. Five more cruises were conducted in the Changjiang Estuary and ECS inner shelf from 12 to 21 May 2020, 15 to 27 October 2020, 9 to 15 March 2021, 12 to 18 July 2021, and 14 to 19 October 2021. The field-measured SSS data were also measured by using a SeaBird CTD recorders (SBE 25plus or 911plus). Additionally, we also downloaded the new field SSS data from SOCAT version 2022 from January 2020 to October 2021 and WOD 2018 from August 2020 to July 2022. The distribution of the observed SSS data is shown in Figure 6a.

2.2.2. Satellite Data

To obtain high quality and sufficient data set, the observed SSS was matched-up with daily satellite products to represent the in situ situation. Moderate Resolution Imaging Spectroradiometer (MODIS) level-2 daily R_{rs} and SST data (spatial resolution of ~1 km) between 2003 and 2022 were used in this study, and obtained from NASA Goddard Space Flight Center (GSFC) (https://oceancolor.gsfc.nasa.gov/). Specifically, the spectral bands of R_{rs} data we adopted were 412, 443, 469, 488, 531, 547, 555, 645, 667, and 678 nm. The MODIS daily R_{rs} and SST were used to matchup with the observed SSS data sets, which within a 3 × 3-pixel window centered on the

Table 1
The Sources and Ranges of SSS Measurements, and the Number of These Measurements Match-Up With Daily MODIS L2 R_{rs} and SST Data

Data source	Surveying time	Range of SSS	Range of SSS with satellite matchups	Number of SSS observations	Number of SSS observations with satellite matchups
Changjiang Estuary cruises	03.2013–07.2020	0.12–34.52	14.68–34.26	1,003	174
201705 VOCO cruise	05.2017	18.12–30.38	18.12–30.25	57	16
Xiong et al. (2019, 2020)	12.2005–10.2018	0.1–34.52	27.89–34.16	347	22
ECS cruises	05.2018–05.2019	24.75–34.58	28.90–34.49	83	16
SOCAT	05.2003–10.2019	28.95–35.05	28.95–34.93	16,201	3,724
WOD	02.2003–06.2020	21.48–35.44	21.48–35.37	9,575	1,705

Note. These SSS data were collected at depths ≤5 m covered all seasons. These SSS data with satellite matchups were used to develop the RF model.

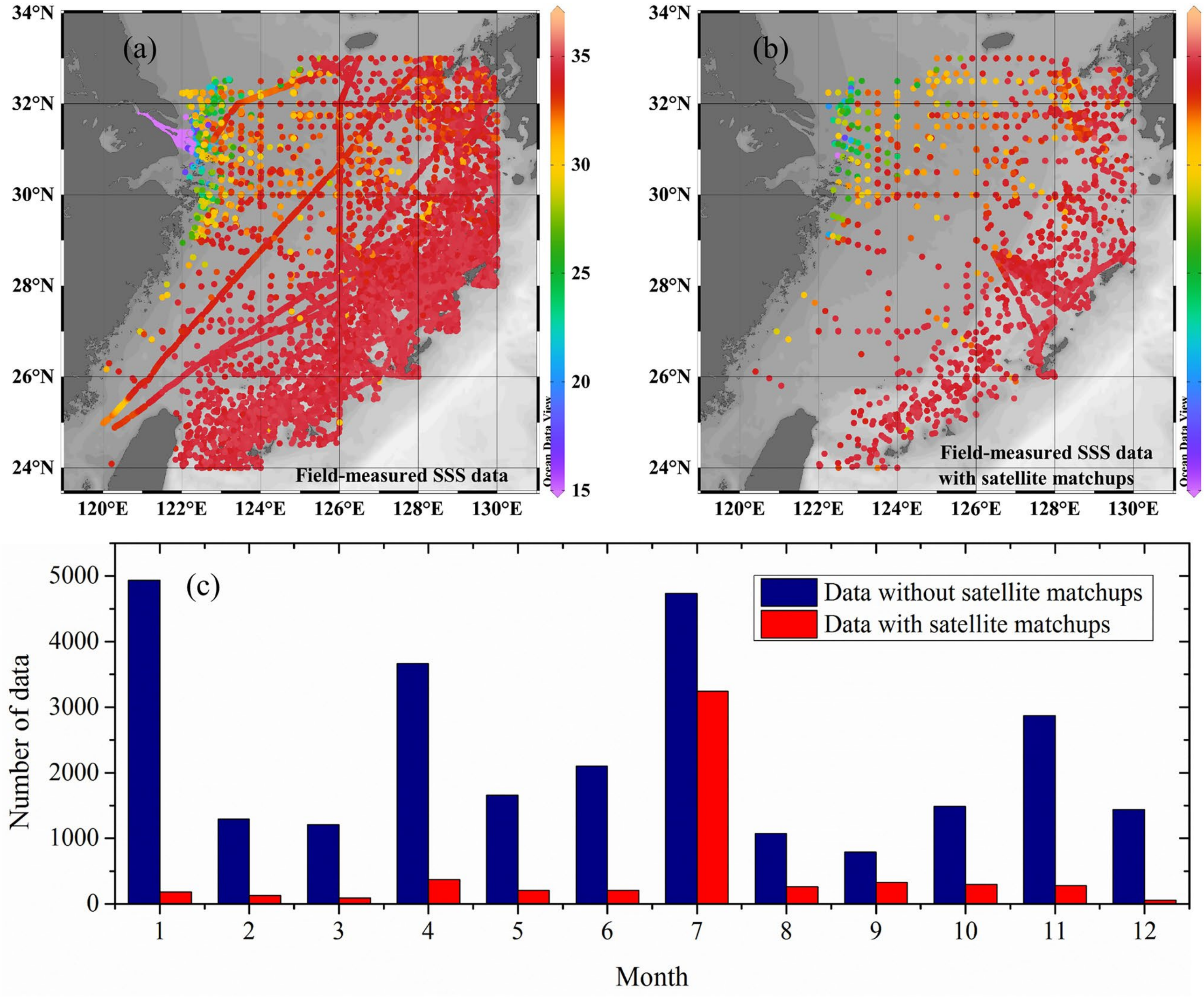

Figure 2. Spatial distributions of the observed SSS in the ECS from February 2003 to July 2020. (a) Locations of underway and discrete SSS observations (b) observed SSS data after match-up with daily MODIS L2 R_{rs} and SST data. (c) Number of observed SSS data in each month between 2003 and 2020 shown in (a) and (b) with and without MODIS matchups. Note that the large data gaps in Figure (a) and (b) are due to quality-controlled by various flags such as cloud coverage, stray light and sun glint to remove invalid satellite signals.

position of each observation were extracted and averaged. A total number of 5,657 and 2,151 conjugate observation data were available for model development and independent validation respectively, the spatial distributions of the observed SSS data after match up with MODIS R_{rs} and SST are shown in Figures 2b and 6b.

2.3. Model Development

2.3.1. Model Selection and Structure

Machine-learning approaches (e.g., MPNN, support vector machine (SVM) regression) have been widely used to retrieve marine environmental variables, such as sea surface nitrate (SSN), chlorophyll *a*, and carbonate parameters (pH, total alkalinity, surface seawater partial pressure of CO_2) (Chen, Hu, Barnes, Wanninkhof, et al., 2019; Hu et al., 2021; Li, Bellerby, Ge, et al., 2020; Li, Bellerby, Wallhead, et al., 2020; Sauzède et al., 2015; Yu et al., 2022), these methods are characterized by the ability to approximate non-linear relationships between model inputs and target variables without explicit knowledge of their functional dependencies (Chen & Hu, 2017).

In this study, the traditional regression approaches (MLR, MNR) and machine-learning based regression approaches (such as decision tree, random forest, SVM and MPNN) were first tested including R_{rs} in 9 spectral

Table 2
Different Combinations of Input Variables and Their Performance When Training the RF Model

		Model training			Model validation		
Model	Model inputs	R^2	RMSE	MAE	R^2	RMSE	MAE
1	$R_{rs}(412)$, $R_{rs}(488)$, $R_{rs}(555)$	0.69	1.13	0.52	0.67	1.19	0.53
2	$R_{rs}(412)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$	0.74	1.03	0.48	0.72	1.09	0.49
3	$R_{rs}(412)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, SST	0.78	0.95	0.39	0.79	0.95	0.40
4	**$R_{rs}(412)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, SST, JD**	**0.84**	**0.81**	**0.31**	**0.86**	**0.77**	**0.30**
5	$R_{rs}(412)$, $R_{rs}(443)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, SST, JD	0.83	0.82	0.31	0.85	0.80	0.31
6	$R_{rs}(412)$, $R_{rs}(443)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, $R_{rs}(678)$, SST, JD	0.84	0.81	0.31	0.86	0.77	0.30
7	$R_{rs}(412)$, $R_{rs}(443)$, $R_{rs}(469)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, $R_{rs}(678)$, SST, JD	0.84	0.81	0.31	0.86	0.77	0.31
8	$R_{rs}(412)$, $R_{rs}(443)$, $R_{rs}(469)$, $R_{rs}(488)$, $R_{rs}(547)$, $R_{rs}(555)$, $R_{rs}(667)$, $R_{rs}(678)$, SST, JD	0.83	0.82	0.31	0.86	0.79	0.31
9	$R_{rs}(412)$, $R_{rs}(443)$, $R_{rs}(469)$, $R_{rs}(488)$, $R_{rs}(531)$, $R_{rs}(547)$, $R_{rs}(555)$, $R_{rs}(667)$, $R_{rs}(678)$, SST, JD	0.84	0.81	0.32	0.85	0.79	0.32

Note. Skill statistics are coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE). The bold values represent the best performance of the RF model.

bands at 412, 443, 469, 488, 531, 547, 555, 667, and 678 nm, SST and Julian day (JD) as model inputs, all these approaches were implemented in Matlab (R2018a). Note that the R_{rs} in 645 nm was not used as too few data could be matched with field-measured SSS data. Among these tested approaches, RF had the best performance (Table S1). Thus, the RF was then selected to model the SSS in the ECS. Then the RF model was applied to test different combinations of input variables to determine which variables to use (Table 2). The optimal performance of the RF model was achieved when R_{rs} in 4 spectral bands (412, 488, 555, 667 nm), SST and JD were used as inputs. Note that SST was used to present the water mixing between freshwater and seawater, and upwelling (Chen & Hu, 2017; Palacios et al., 2009), the JD was selected to denote the seasonal cycle.

RF is an ensemble regression algorithm (the structure of RF was shown in Figure 3), which is composed of multiple decision trees (Liaw & Wiener, 2002). An individual decision tree grows depending on the characteristics of the input data during the training process. Each decision tree comprises multiple decision nodes and a leaf node, the decision nodes assess the characteristics of the input data and pass their subsets to different branches, the leaf node represents the end of the splitting process (Krzywinski & Altman, 2017; Li et al., 2018). Therefore, decision tree is very sensitive to how and where to split, thence, although a minor change in input data may cause large differences in the tree structure, and the predicted results are usually prone to overfit (James et al., 2013). Instead, the RF consisted of many decision trees by bagging to decrease the influence of overfitting and improve the predictive accuracy and generalization of the model. In the training of the RF model, the training data set is resampled by bootstrapping and random subsets of the training data are used to train decision trees, in which way, the independence of each decision tree can be greatly improved (Chen, Hu, Barnes, Xie, et al., 2019). The final model result of RF is a weighted average of the output from each decision tree.

In this study, developed the relationship between the predictors (R_{rs} in 412, 488, 555, 667 nm, SST and JD) and target variable (SSS) by using a function of fitrensemble in Matlab (R2018a) with the approach designated as "bag." For the model development, the 5,657 conjugate observation data were divided into two subsamples randomly, with 80% ($N = 4{,}504$) were used to train the RF model, and the remaining 20% ($N = 1{,}153$) were used to validate the model. During the training process, there are two critical parameters to tune the structure of RF model, one is the minimum leaf size (MLS), and another one is number of learning cycles (NLC). The leaf size represents the number of data used in each decision node of decision trees, thus, the MLS determines the splits and depth of a decision tree. The NLC represents the number of decision trees in the RF model. In this study, the NLC were varied from 2 to 50, and the MLS was varied from 1 to 20. After several iterations, when the NLC was set to 30 and MLS was set to 8, the RF model was shown to be stable and with the best performance. With these settings, the RF model was developed to predict SSS in the ECS.

2.3.2. Data Preprocessing of RF Model

Based on Section 2.3.1, the optimal combination of model inputs is MODIS-derived R_{rs} in the 4 spectral bands (412, 488, 555, 667 nm), SST, and JD. The benefit of using contemporaneous satellite products to develop the

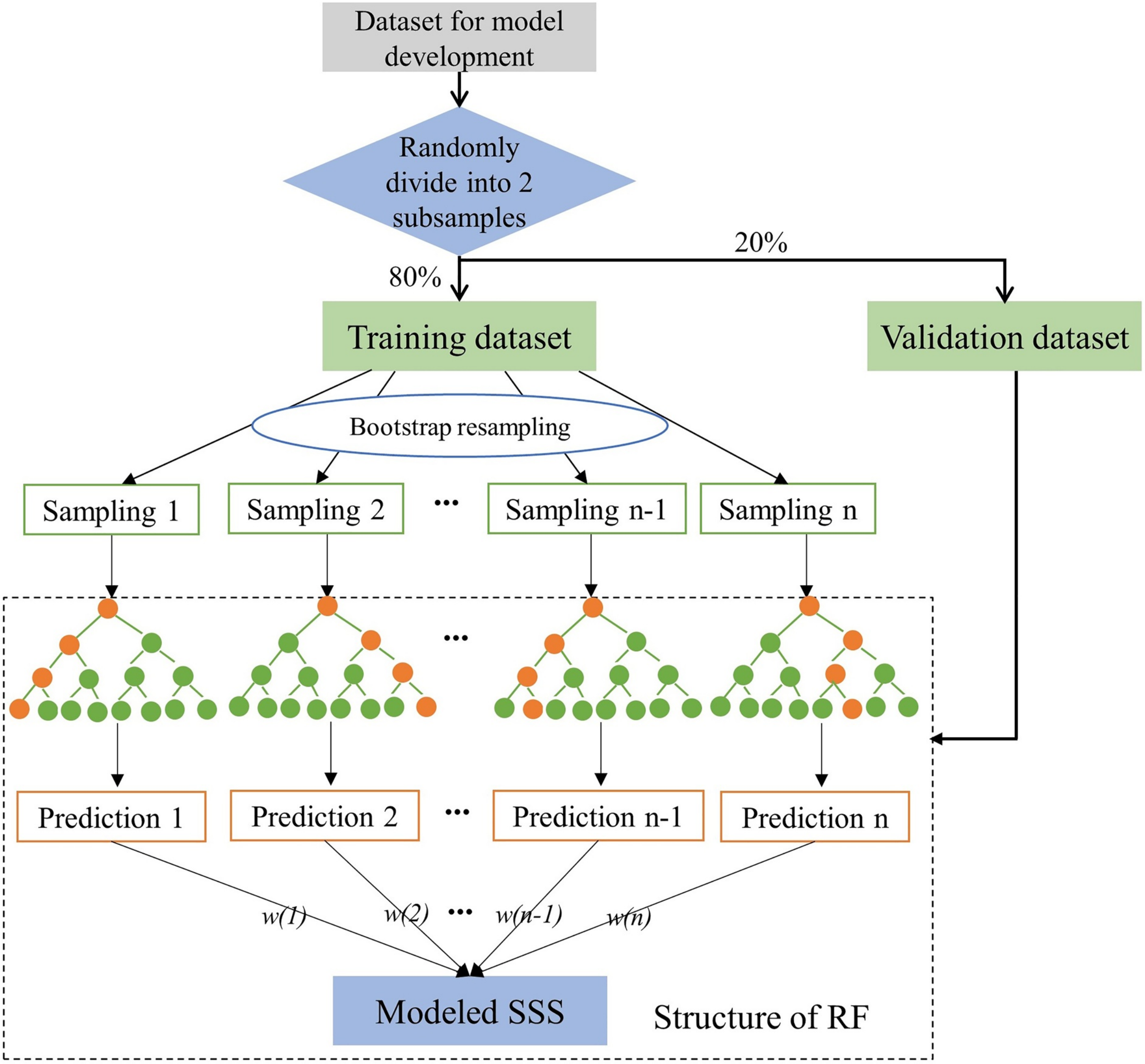

Figure 3. Schematic diagram of the RF model. Note that *w*(*i*) (where *i* = 1, 2, …, n − 1, n) is the weight of the predicted SSS from the corresponding decision tree.

SSS model due to that the errors from satellite products will be minimized (Chen & Hu, 2017). Since we tested several traditional regression approaches and machine-learning based regression approaches, some methods are sensitive to the inputs and output data (e.g., SVM, MPNN), thus, both the inputs and output data were normalized to eliminate the influence of dimensional of the data on the model (Li, Bellerby, Ge, et al., 2020). All the data were normalized using following equations (Chen & Hu, 2017; Xi et al., 2020):

$$x_{i,j} = \frac{x_{i,j} - \text{mean}(x_{i,j})}{\sigma(x_{i,j})} \tag{1}$$

with σ refers to the standard deviation (SD) of the input or output variables. For the date inputs, the JD was transformed as following equation based on Sauzède et al. (2015) due to the periodicity:

$$\text{sJD} = \sin\left(\frac{\text{Day} \cdot \pi}{182.625}\right) \tag{2}$$

$$\text{cJD} = \cos\left(\frac{\text{Day} \cdot \pi}{182.625}\right) \tag{3}$$

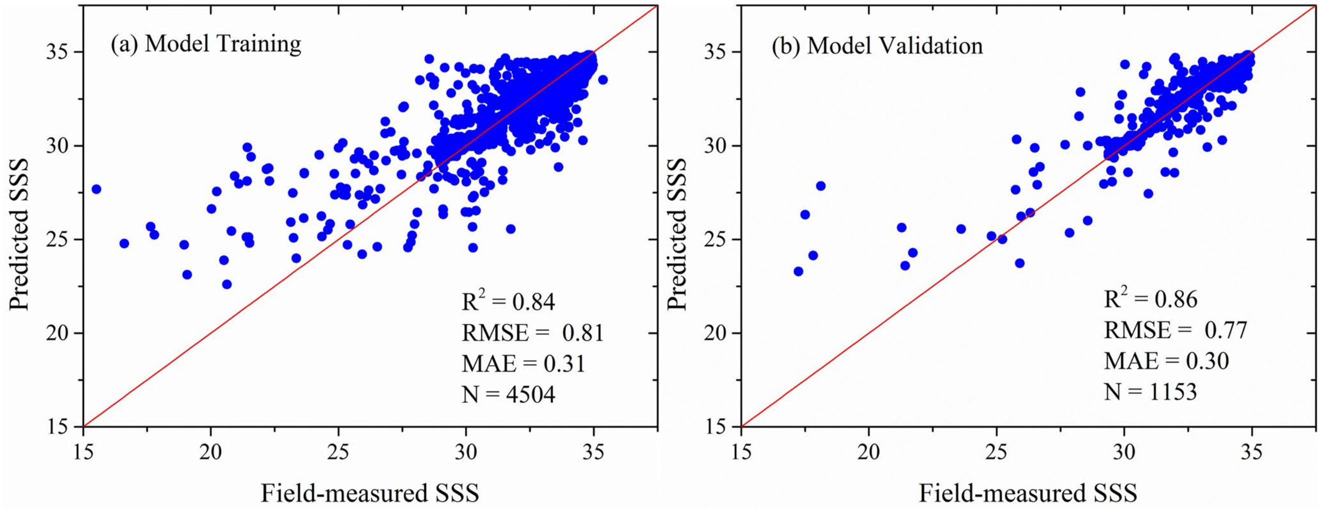

Figure 4. Comparison of the estimated SSS by RF model with corresponding observed SSS during (a) model training, and (b) model validation in the ECS.

The coefficient 182.625 refers to half of the number of days per year (365.25).

3. Results

3.1. Model Performance

To assess the performance of the RF-based SSS model in the ECS, the SSS estimated by RF model were compared with the corresponding observed SSS (Figure 4) using three statistical indices including root mean square error (RMSE), coefficient of determination (R^2), and mean absolute error (MAE). During the model training (80% of all the field-measured SSS data with satellite matchups), the RMSE was 0.81, MAE was 0.31 and R^2 was 0.84 (Figure 4a). Similarly, for the mode validation (remaining 20% of all the field-measured SSS data with satellite matchups), the estimated SSS with a RMSE of 0.77, MAE of 0.30 and R^2 of 0.86 (Figure 4b). Additionally, the RF model performed better at higher SSS during model development, with RMSE of 1.50 and 0.47, MAE of 0.63 and 0.22 for SSS ≤ 31 and SSS > 31 in the model training, and RMSE of 1.42 and 0.47, MAE of 0.60 and 0.22 for SSS ≤ 31 and SSS > 31 in the model validation. The performance of the RF model was comparable with previous reported studies, for example, Bai et al. (2013) developed the linear relationship between SSS and a_{CDOM} (355 nm) to estimate the SSS with 87.1% of the data were within the absolute error of ±1.5 in the ECS, Qing et al. (2013) developed a MLR model to estimate the SSS with a RMSE of 0.833 in the BS, and Chen and Hu (2017) developed a MPNN model to predict the SSS with a RMSE of 1.2 in the northern Gulf of Mexico. For reference, the performances of other traditional empirical approaches (MLR, MNR) and machine-learning based empirical approaches with different kernel functions (decision tree, random forest, SVM regression, and MPNN) were shown in Table 3. Obviously, the RF (bagged trees) showed the best performance compared to the other tested methods with the same input variables, thus, was selected in this study.

The histogram of residuals between observed and modeled SSS in the model training and validation data sets showed that 91% of the residuals were within the RMSE of ±0.81 (Figure 5c). Figure 5a showed the spatial distribution of SSS residuals (observed SSS minus estimated SSS) to further explore the RF uncertainties spatially. The ECS inner shelf had the largest residuals, while the outer ECS had the narrowest range of SSS residuals. Figure 5b showed that the most points of residuals beyond ±2RMSE are concentrated in areas near the Changjiang Estuary, where are intensely affected by Changjiang River discharge during spring and summer due to the flooding of the Changjiang River (Li, Bellerby, Wallhead, et al., 2020).

3.2. Independent Validation and Method Comparison

An independent data set (described in Section 2.2.1) was applied to the RF model to further examine the predictability of the RF model in predicting SSS in the ECS (Figures 6a and 6b, Table 4). The comparison between

Table 3
Comparison of Model Performances Between Traditional Empirical Approaches (MLR and MNR) and Machine-Learning Based Empirical Approaches With Different Kernel Functions (Decision Tree, Random Forest, SVMs and MPNN)

		Model training			Model validation		
Model	Kernel function	R^2	RMSE	MAE	R^2	RMSE	MAE
MLR	–	0.68	1.13	0.73	0.70	1.14	0.72
MNR	–	0.72	1.06	0.70	0.73	1.08	0.70
Decision tree	Simple tree	0.75	1.01	0.42	0.78	0.97	0.40
	Medium tree	0.77	0.97	0.38	0.80	0.92	0.36
	Complex tree	0.76	0.98	0.36	0.81	0.90	0.35
Random forest	Boosted trees	0.81	0.88	0.43	0.84	0.83	0.44
	Bagged trees	**0.84**	**0.81**	**0.31**	**0.86**	**0.77**	**0.30**
SVM	Linear	0.67	1.15	0.72	0.69	1.16	0.71
	Quadratic	0.71	1.08	0.63	0.70	1.13	0.67
	Cubic	−3.66	4.35	2.12	−0.11	2.18	1.24
	Fine Gaussian	0.67	1.16	0.42	0.61	1.29	0.45
	Medium Gaussian	0.78	0.94	0.42	0.78	0.97	0.41
	Coarse Gaussian	0.74	1.03	0.55	0.76	1.01	0.53
MPNN	Scaled conjugate gradient optimization and Bayesian regularization	0.76	0.94	0.49	0.79	0.93	0.47

Note. The statistics were calculated based on the model development data set (model training and validation data) using the same input variables including $R_{rs}(412)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, SST and JD. The bold values represent the RF (bagged trees) model has the best performance. Note that the negative R^2 means there was a strong bias in the estimated SSS.

observed SSS and corresponding RF-derived SSS had a RMSE of 0.66 and MAE of 0.39 for all independent data (Figure 6c). Similarly, the RF model still showed a better performance at higher SSS and larger uncertainties were seen for SSS ≤ 31, with RMSE of 3.24 and 0.50, MAE of 2.49 and 0.35 for SSS ≤ 31 and SSS > 31, mainly owing to the RF model was more sensitive to the errors of R_{rs} and SST at lower salinity (Figure 11).

Furthermore, Bai et al. (2013) proposed the SSS-CDOM relationship for the ECS, which was also applied to the same independent data set, with a RMSE of 1.78, MAE of 1.41 (Figure 6d). Similar to the RF, the SSS-CDOM relationship also showed a better predictive performance in higher SSS range, with RMSE of 5.36 and 1.65, MAE of 3.77 and 1.36 for SSS ≤ 31 and SSS > 31. The comparison of the RF model and SSS-CDOM relationship, the RF model performed much better than the latter in all SSS range, this may because the semi-analytical method is able to interpret some physical mechanisms, but it is difficult to quantify these physical mechanisms accurately and also subject to unknown factors. On the contrary, these unknown factors driving the uncertainties in RF model could be canceled out by tuning of the model structures and empirical coefficient (Chen & Hu, 2017).

3.3. Seasonal and Interannual Variations of Modeled SSS

The seasonal and spatial variability of monthly SSS distribution estimated by the RF model in the ECS from August 2002 to July 2022 are shown in Figure 7, based on the monthly MODIS L3 R_{rs} and SST data. The estimated SSS values range from 25.76 to 34.77 in spring, 23.02 to 34.48 in summer, 27.39 to 34.73 in autumn, and 28.83 to 34.77 in winter, with average SSS values of 33.41, 32.34, 33.44, and 33.84, respectively. Regardless of the season, the SSS generally increased from nearshore to offshore regions. The lowest SSS generally dominated the Changjiang Estuary and nearshore coast due to the import of Changjiang River freshwater, the intermediate SSS is mainly distributed in the transitional region because of mixing of the coastal low-salinity water and offshore seawater, and the high SSS is distributed in the open water, influenced by the high-salinity KS. The patterns of the monthly average SSS over the 20-years period in Figure 7 were similar to the observed results of Chen et al. (2006) and Lie et al. (2003) and were consistent with studies that modeled the SSS of the ECS from satellite (Ahn et al., 2008; Bai et al., 2013; Kim et al., 2020; Sasaki et al., 2008). From October to the following April (in the dry season), the northerly East-Asian monsoon prevails over the ECS (Lie et al., 2003). The CDW

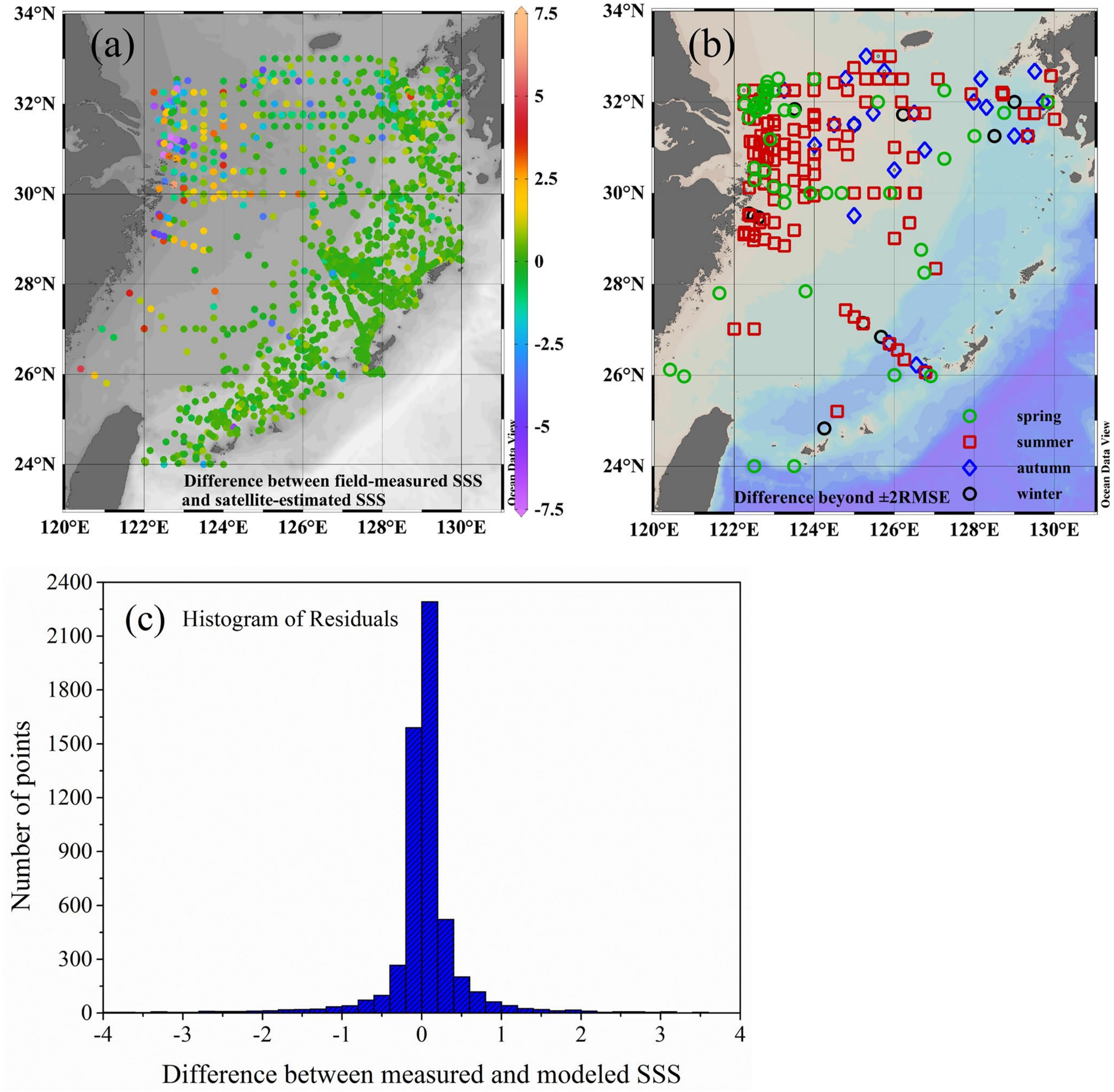

Figure 5. (a) Spatial distributions of the SSS residuals (observed SSS minus SSS estimated by RF model) and (c) histogram of residuals for the model development data set (99.15% of the residuals were within −4–4). (b) Seasonal distribution of stations with SSS residuals greater than ±2RMSE.

flows southward along the Zhejiang-Fujian coast and the low-salinity plume is confined to the Changjiang Estuary and a narrow coastal band (Figures 7a–7d and 7j–7l). However, during the wet season, the monsoon switches to the southern from May to August, the strong CDW extends farther offshore to northeast and appears as a tongue-shaped plume of low salinity (Figures 7e–7h).

The seasonal cycles of estimated SSS for the ECS during the period of August 2002 to July 2022 are shown in Figure 8a. SSS decreases gradually from spring to summer, and reaching a minimum value in August, then slowly increases from autumn to winter, remaining almost constant in winter. These seasonal variations in the SSS were in good agreement with the Changjiang River discharge (Figure 8). Moreover, the SST showed the opposite trend to SSS, indicating that from spring to summer, the SST increases gradually resulting in strong

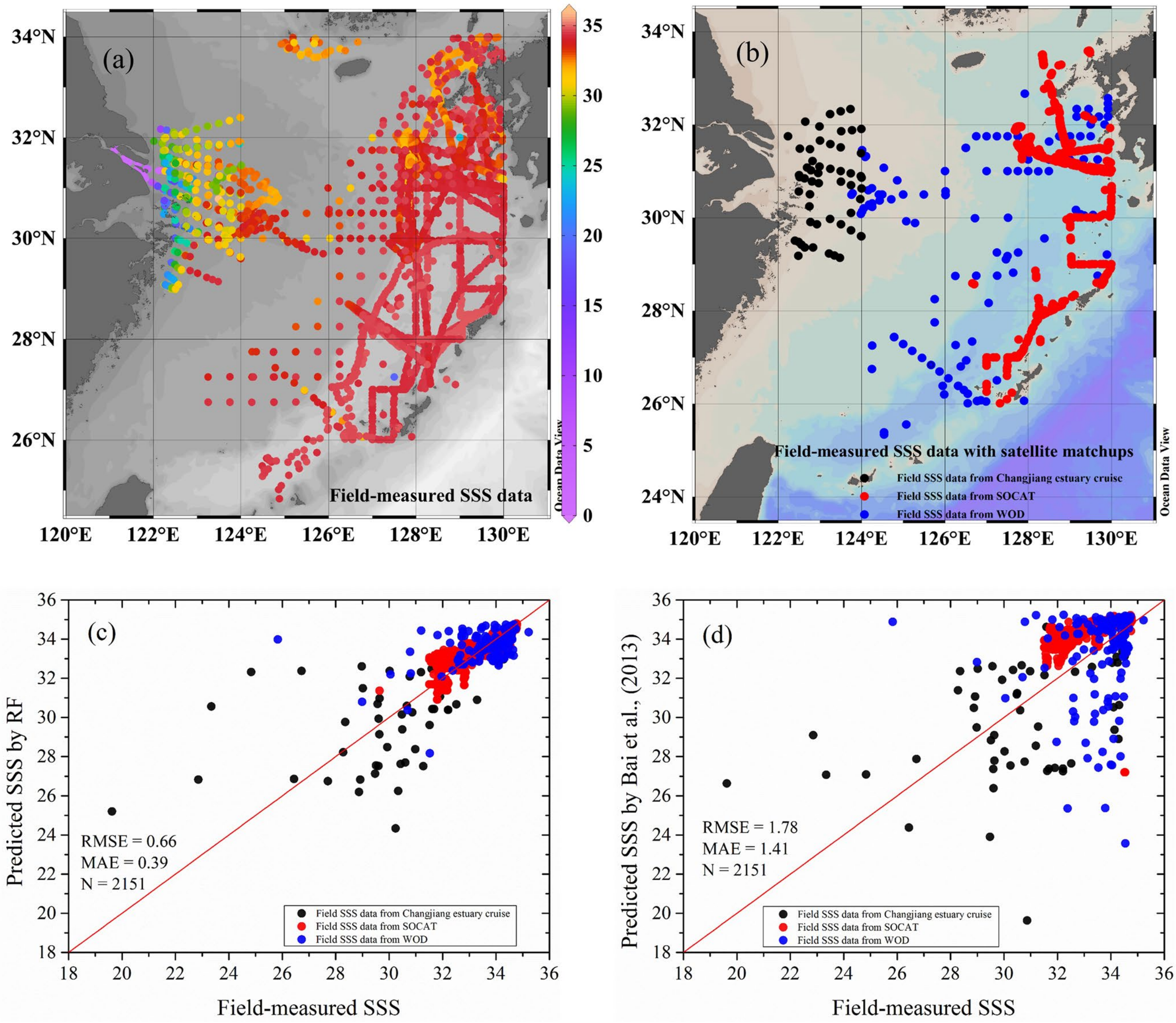

Figure 6. (a) Spatial distributions of the observed SSS in the ECS from 2020 to 2022; (b) locations of observed SSS data after match-up with daily MODIS L2 R_{rs} and SST data; (c) Comparison of the estimated SSS by RF model with corresponding observed SSS using the independent data set shown in (b); (d) Comparison of the estimated SSS by Bai et al. (2013) with corresponding observed SSS using the independent data set shown in (b). The black, red and blue dots represent the observed SSS were from Changjiang Estuary cruises, SOCAT and WOD.

Table 4
The Sources and Ranges of SSS Measurements, and the Number of These Measurements Match-Up With Daily MODIS L2 R_{rs} and SST Data

Data source	Surveying time	Range of SSS	Range of SSS with satellite matchups	Number of SSS observations	Number of SSS observations with satellite matchups
Changjiang Estuary cruises	05.2020–10.2021	0.11–34.52	19.63–34.52	420	53
SOCAT	01.2020–10.2021	31.04–34.78	31.50–34.78	7,798	1,932
WOD	08.2020–07.2022	4.16–35.22	25.83–35.22	1,775	166

Note. These SSS data were collected at depths ≤5 m. These SSS data with satellite matchups as an independent data set was used to validate the RF model.

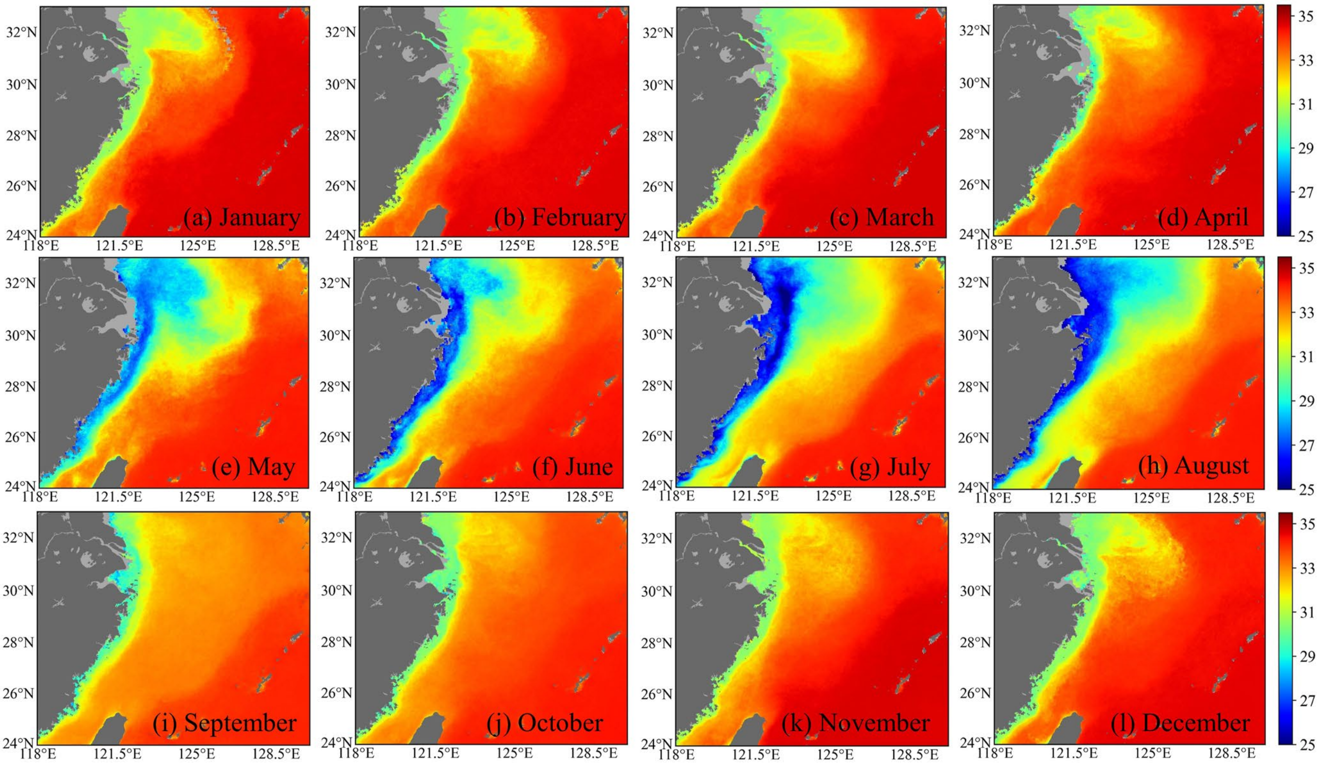

Figure 7. Distributions of monthly average RF-derived SSS in the ECS from August 2002 to July 2022.

surface stratification, and the fresh water may stay on the sea surface, the SSS continues to decrease. During the cold season, with the intensification of water mixing, the high-salinity bottom and subsurface waters mixed into the surface water, which is a contributor to the increase in SSS (Zhang et al., 2010). The $R_{rs}(412)$ had the opposite trend to SSS, while the $R_{rs}(488)$, $R_{rs}(555)$, and $R_{rs}(667)$ showed a similar trend with SSS, suggesting that a negative correlation between $R_{rs}(412)$ and SSS, and there are positive relationships between $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, and SSS.

The interannual variability of the area-averaged monthly SSS for the ECS and three subregions from August 2002 to July 2022 are shown in Figure 9. Over the whole ECS (blue line in each panel of Figure 9), the monthly SSS exhibited similar seasonal fluctuations between 31.73 and 34.20 from 2002 to 2022, the lowest SSS value was

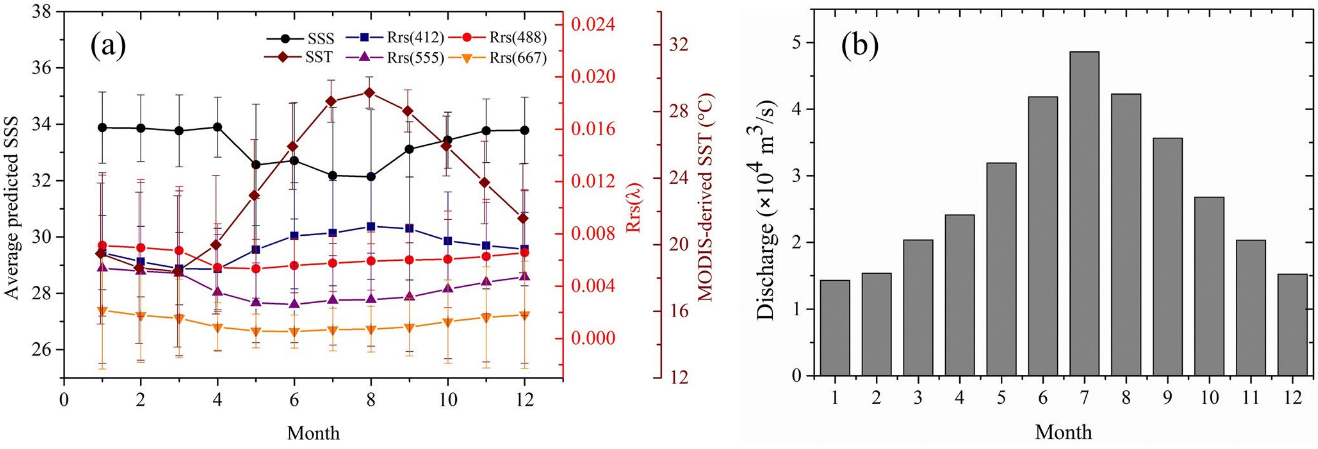

Figure 8. (a) Seasonal variations in modeled SSS, MODIS-derived SST and $R_{rs}(\lambda)$. (b) Monthly average Changjiang River discharge from August 2002 to July 2022 (Datong station, data obtained from the Hydrological Information Center of China, http://www.mwr.gov.cn/).

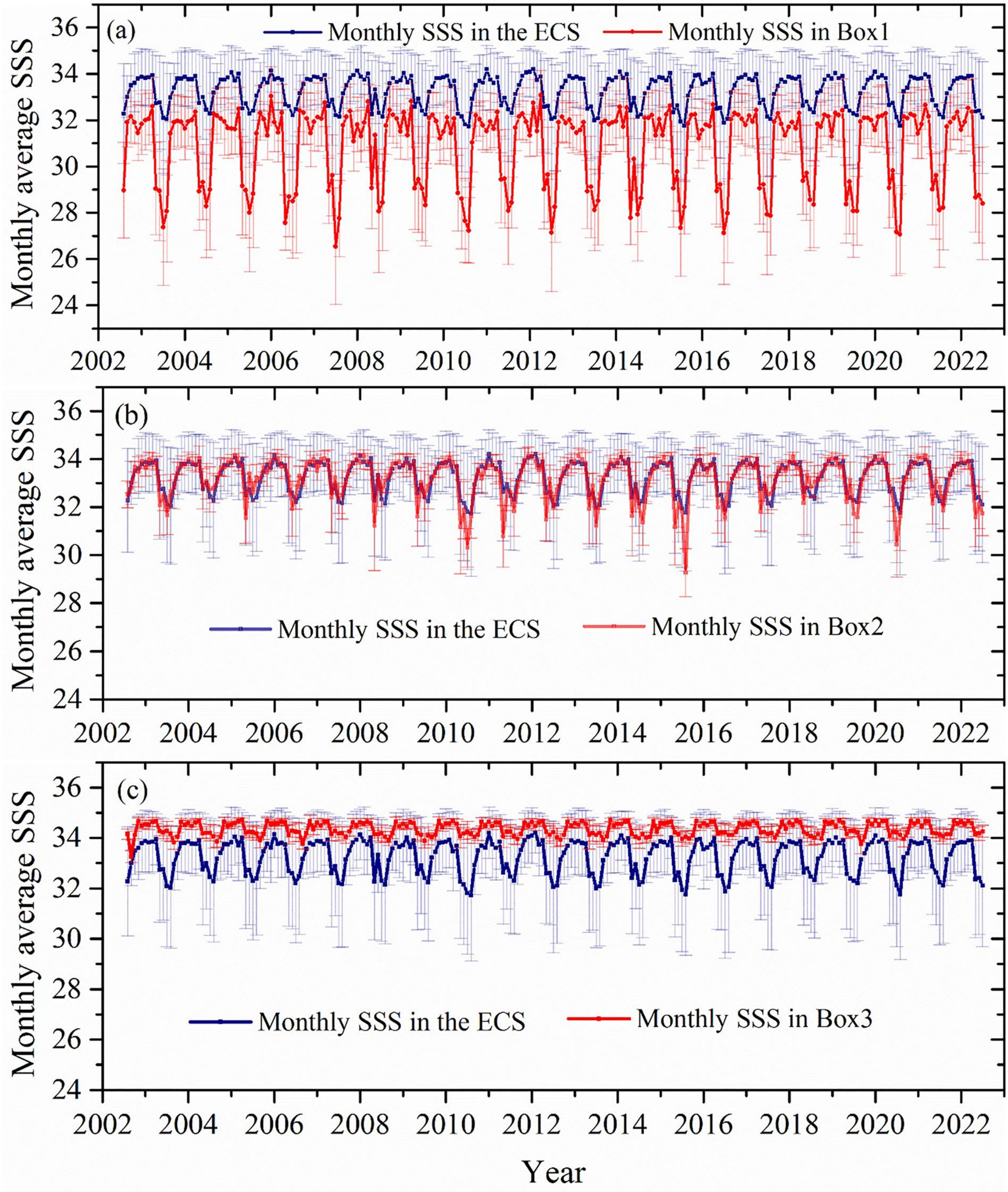

Figure 9. Interannual variations of MODIS-derived SSS in the whole ECS and three subregions annotated in Figure 1 during August 2002 to July 2022. Error bars refer to the SD of monthly average SSS in each region.

found in summer while the highest SSS value was found in winter with a SD of ~±1.49 on average. The monthly SSS in three subregions (Box 1, Box 2, and Box 3, Figure 1) also showed a similar pattern of interannual variability to that of the entire ECS (Figure 9). For example, Box 1 is located near the Changjiang River Estuary, due to the effects of seasonal variation in the Changjiang discharge, the SSS in Box 1 was the lowest (26.55–33.07, SD of ~± 1.40 on average) and showed largest seasonal amplitudes compare with other two boxes and the entire

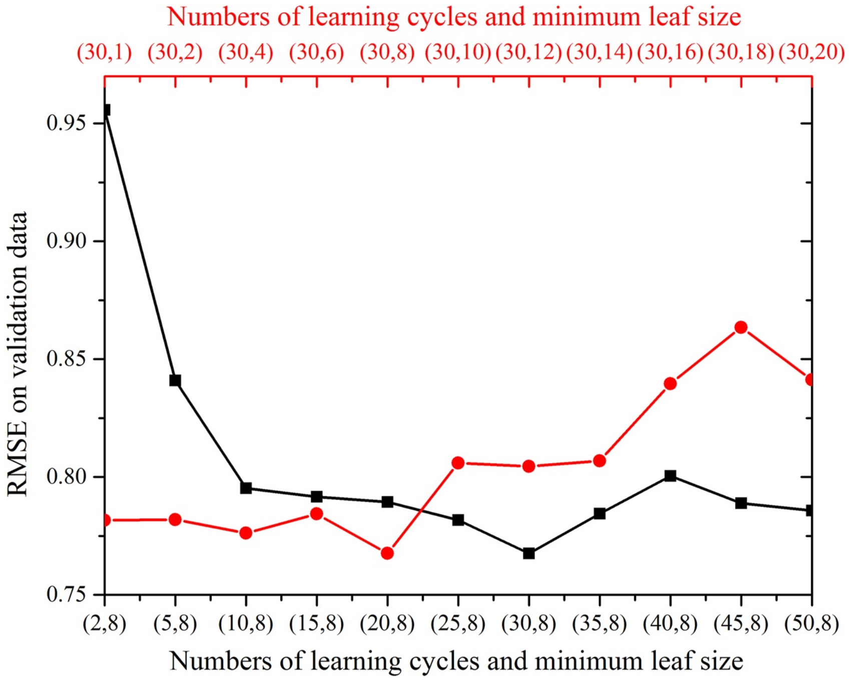

Figure 10. Comparison of the model performance of different NLC and MLS on validation data ($N = 1{,}153$). The red line represents the model performance response to an increase in MLS with the NLC was fixed at 30. The black line represents the model performance response to increasing NLC with the MLS was fixed at 8.

ECS (Figure 9a). Box 2 is located in the central ECS, which is a transitional region affected by water mixing of coastal water and offshore water, thus, the SSS showed the intermediate value (29.27–34.14, SD of ~±0.51 on average) and its interannual variation has similar magnitudes as those in the entire ECS (Figure 9b). In additional, the lower SSS in Box 2 generally corresponds to the Changjiang River floods (Figures 9b and 12b), indicating that the transitional region may be influenced by Changjiang River discharge, especially in flood years. Box 3 is located in the ECS offshore water, where the hydrological environment is stable and rarely affected by freshwater input. Therefore, the SSS in Box 3 showed highest value (33.26–34.71, SD of ~±0.20 on average) and weakest seasonal amplitudes (Figure 9c).

4. Discussion

4.1. Model Sensitivity

The NLC and MLS are important parameters of the RF model, we examined the model performances on validation data ($N = 1{,}153$) when the NLC ranged from 2 to 50, and the MLS ranged from 1 to 20 (Figure 10). Specifically, when the MLS was fixed at 8, the model performance showed significant improvement with the NLC was increased from 2 to 10, the model performance showed slight improvement with the NLC was increased from 10 to 30, while after the NLC was more than 30, there was no improvement of the model performance. On the contrary, when the NLC was fixed at 30, the model performance was relatively stable with the MLS was increased from 1 to 8, and the model performed best when the MLS was 8, but after the MLS was more than 8, the model performance gradually decreased with the increase of the MLS. Thus, the best performance of the SSS model was obtained with the NLC was 30 and the MLS was 8.

To evaluate the sensitivity of the RF model to errors in the input variables, the errors of each variable were added into the RF model separately. In order to obtain the MODIS-derived SST uncertainties in the ECS, we compared the MODIS-derived SST and field-measured SST by using both the model development and independent validation data set matchup with satellite SST (Figure 11a), with a RMSE of 1.27°C and R^2 of 0.87. Therefore, errors of ±1.5°C was added in the SST to do the sensitivity analysis (Figures 11b and 11c). Specifically, added +1.5°C for SST uncertainties, the RF model showed very slightly underestimated, with a RMSE of 0.31 and MB of −0.10. Conversely, added −1.5°C for SST uncertainties, the RF model showed slightly overestimated, with a RMSE of

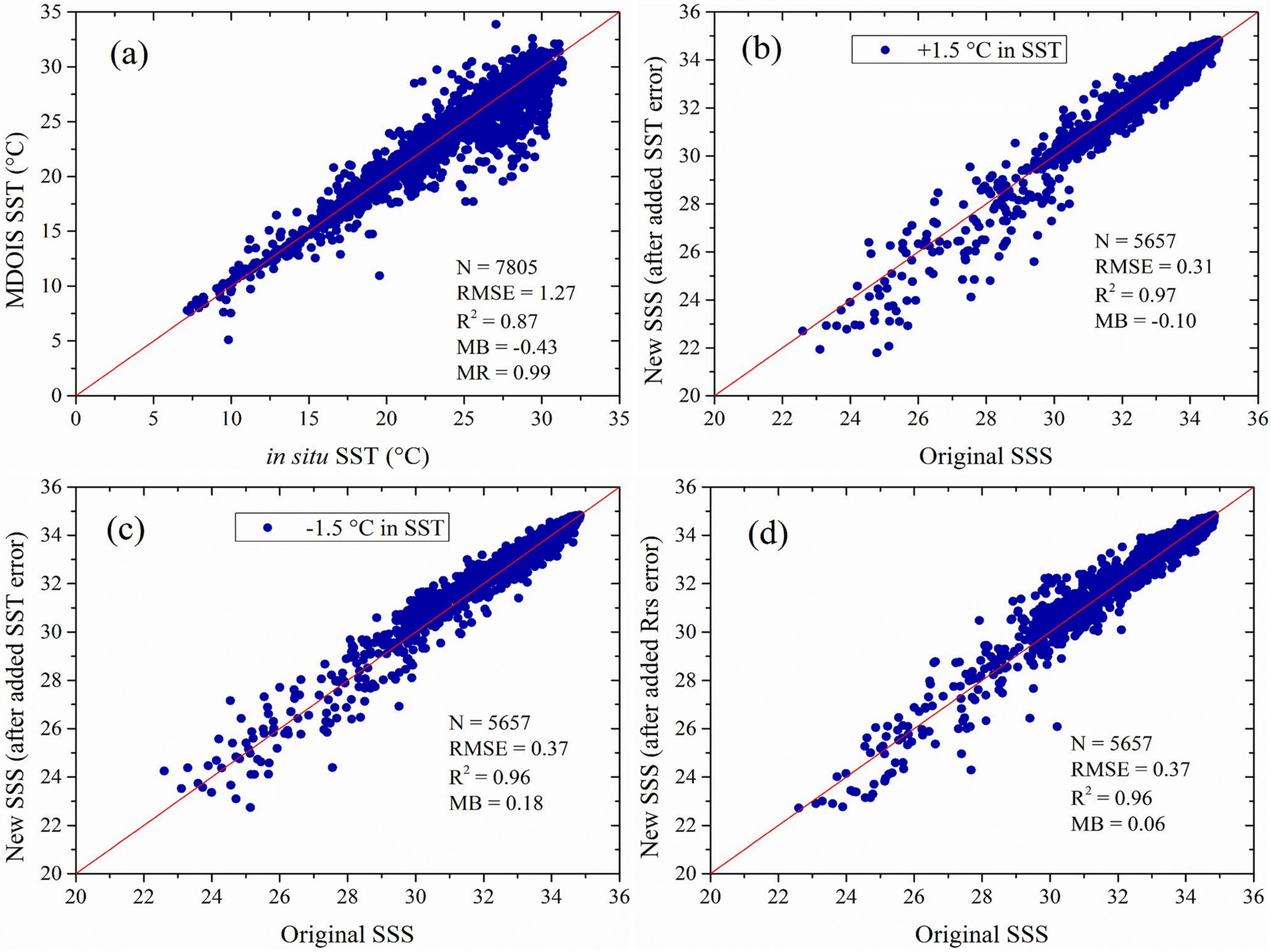

Figure 11. (a) Scatterplot of MODIS-derived SST and in situ SST. Sensitivity of the RF-based SSS model to added (d) R_{rs} errors and SST errors (b) +1.5°C in SST, (c) −1.5°C in SST, based on the model development data set.

0.37 and MB of 0.18. These results indicated that the RF model was not sensitive to the SST errors but responded to SST errors in a negative way. This may be because the strong Changjiang discharge leads to the lower salinity in the warm season, while in cold season, the Changjiang discharge decreases significantly, coupled with upwelling and vertical water mixing bring cold and high-salinity subsurface water into the surface, resulting in an increase in SSS.

Unlike SST, assessing the sensitivity of the RF model to the R_{rs} errors is more complex since the errors in MODIS-derived R_{rs} are not independent spectrally, which means that errors in one band are associated with errors in other bands (Chen & Hu, 2017; Hu et al., 2013). The errors in independent MODIS-derived R_{rs} were simulated following the method of Chen and Hu (2017). Briefly, simulated 5,657 R_{rs}(667) errors in the data set used for model development according to Hu et al. (2013) and then calculated the corresponding errors in R_{rs}(412), R_{rs}(488), and R_{rs}(555) (Chen & Hu, 2017; Hu et al., 2013). The RF model showed a slight overestimation after propagating R_{rs} errors to the model, with a RMSE of 0.37 and MB of 0.06. These results showed that the RF model was insensitive to the R_{rs} errors. However, the RF model was more sensitive to the R_{rs} errors at lower SSS range (Figure 11d).

Overall, the estimated SSS variations resulted from errors in SST and R_{rs} were all within the RMSE of the development of RF model, indicating a high tolerance of the RF model to inaccuracies in MODIS-derived SST and R_{rs}. Moreover, as the SST and R_{rs} data used to develop the RF model were obtained from the same satellite, these errors were, likely, already canceled implicitly by the model (Chen, Hu, Barnes, Wanninkhof, et al., 2019).

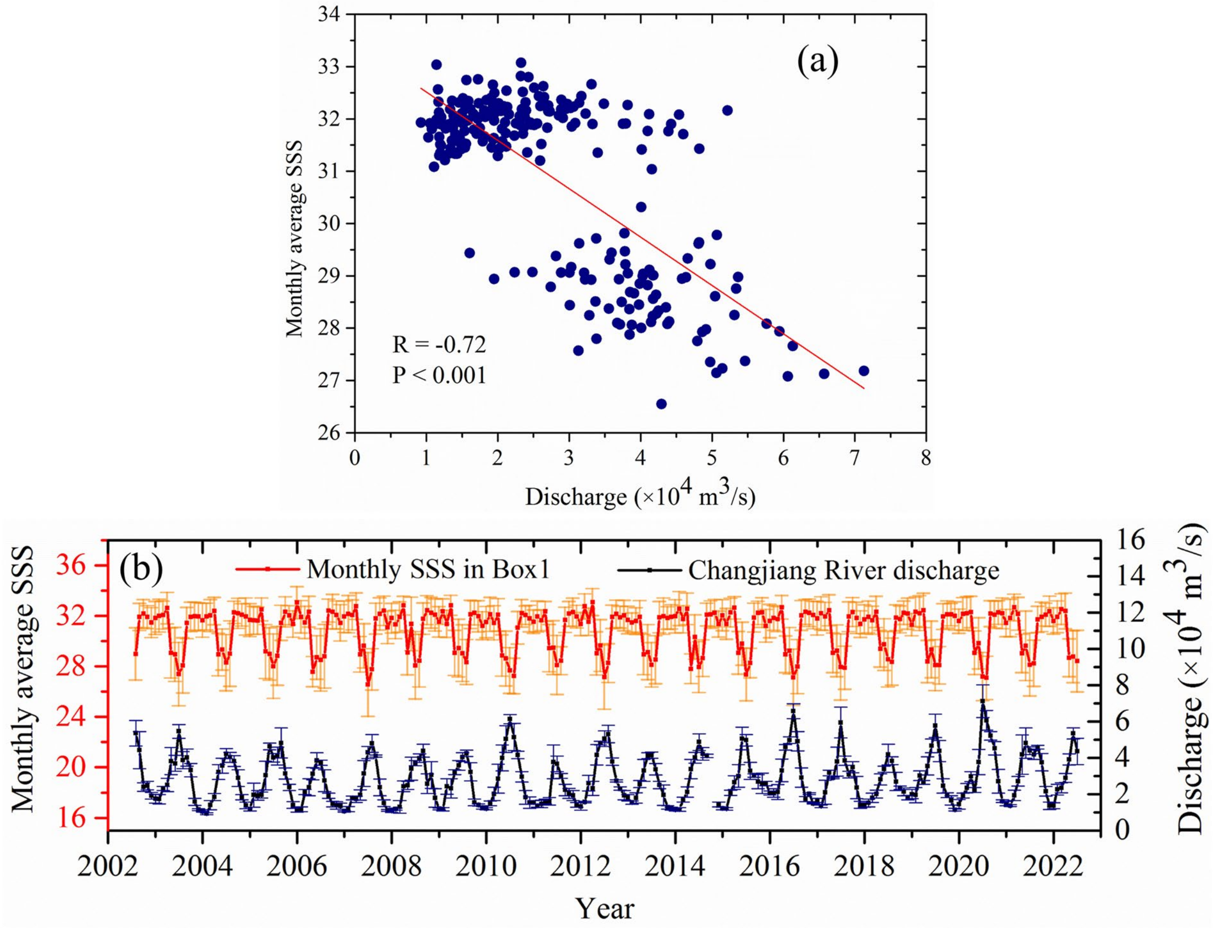

Figure 12. (a) Correlation between Changjiang River discharge and MODIS-derived SSS in the subregion (Box 1), with $N = 238$. (b) Interannual variations of MODIS-derived SSS in the Box 1 and Changjiang discharge during August 2002 to July 2022.

4.2. Effects of Changjiang Discharge on SSS

The distribution of low-salinity plume of the ECS was primarily influenced by the Changjiang discharge mentioned in Section 3.3. The study region near the Changjiang River Estuary (Box 1) was selected to analyze the relationship between Changjiang River discharge and variations of SSS during the period of August 2002 to July 2022. As shown in Figure 12a, there is a negative correlation between the monthly SSS and monthly Changjiang River discharge ($R = -0.72$, $p < 0.001$, $N = 238$), indicating that fluvial input is an important contributor to the formation of the low-salinity plum. Our result is consistent with that of Sun et al. (2019), they also found the high negative relationship between monthly SSS and monthly Changjiang River discharge, and the lower salinity region was caused by freshwater input. The interannual variation of monthly average SSS in Box 1 was in the opposite trend to that of monthly Changjiang discharge (Figure 12b). During the dry season, SSS values were higher, while the SSS gradually decreased during the wet season, reaching the lowest SSS value in July or August. In additional, from the time series, the SSS showed the extreme low values when Changjiang River flood events occurred, such as in the summers in 2010, 2016 and 2020. It suggests that our SSS model was able to capture the significant interannual variations of the Changjiang River plume, the RF-based SSS model can be considered as an efficient tool for mapping and monitoring the SSS in the ECS, especially for the low-salinity plume.

4.3. Model Application to Other Chinese Marginal Seas

In order to examine the generally applicability of the SSS model for Chinese marginal seas, we applied the model to the YS and BS that are adjacent seas of the ECS. The current circulations of these two seas are similar with that in the ECS, especially, also affected by freshwater input (Lin et al., 2005; Zhang et al., 2004). The seasonal

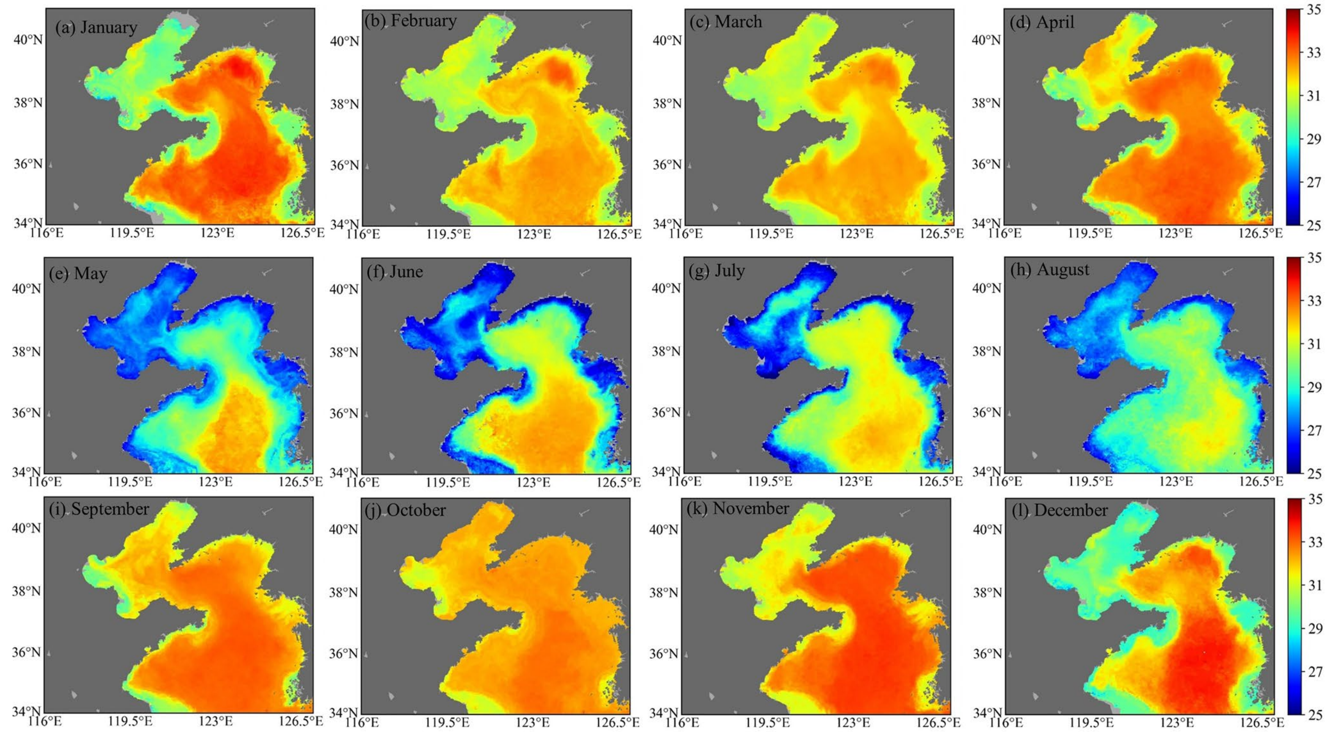

Figure 13. Distributions of monthly MODIS-derived SSS in the Yellow sea and Bohai sea during April 2015–July 2022 using the RF model developed based on the ECS in situ SSS data.

and spatial variations of monthly SSS distribution in the YS and BS from April 2015 to July 2022 were shown in Figure 13, based on the monthly MODIS L3 R_{rs} and SST data. Similar to the ECS, the low SSS generally distributed near the inshore coast (Figure 7), whereas the high SSS are mainly dominated in the central YS and BS. Specifically, the BS comprises three major bays and many rivers discharge into the BS, among which the Yellow River is the largest river. The YS not only has frequent water transportation with the BS and ECS, but also receives large quantities of Changjiang River discharge.

From autumn to early spring, when the northeastern monsoon winds prevail over the YS and BS, the saline YS warm current (YSWC) intrudes into the central BS through the BS Strait, resulting in the BS central has higher SSS than that in the coastal waters. Furthermore, the less saline BS coastal water (BSCoW) also outflows from southern BS Strait and flows southward along the coastal to the YS, therefore, the low SSS is mainly distributed in the nearshore waters of the YS (Figures 13a–13d and 13i–13l; Chen, 2009; Lin et al., 2001). From the late spring to summer, the weak northeastern monsoon wind turns to the southwest wind, which is unfavorable to the development of YSWC, coupled with the strong influences of the freshwater input, the SSS was the lowest in summer of the BS and YS, especially in the coastal waters. However, the relative higher SSS is always distributed in the central of southern YS all year around, which may be affected by the YS cold water (YSCW) (Figures 13e–13h; Sun et al., 2019). Overall, the distributed patterns of monthly MODIS-derived SSS in the YS and BS by using the RF-based SSS model developed in the ECS are in good agreement with the results of reported studies (Qing et al., 2013; Sun et al., 2019, 2022; Yu et al., 2017; Zhang et al., 2018), except the SSS in the central BS from May to August was lower. It may be because the RF model was developed based on the observed SSS of the ECS, and the observed SSS was lower from May to August.

SMAP was designed to monitor the SSS through L-band passive microwave sensing (Entekhabi et al., 2010). Due to the absence of publicly available observational data for the BS and YS, the estimated SSS by RF model based MODIS products and SMAP-estimated SSS from April 2015 to July 2022 were compared to assess the performance of the RF-based SSS model applied to the YS and BS. The monthly L3 SMAP-estimated SSS with a resolution of 0.25° × 0.25° was downloaded from Jet Propulsion Laboratory (JPL, 2020. https://doi.org/10.5067/SMP50-3TMCS). The spatial distributions of SMAP-estimated monthly SSS in the YS and BS were shown in Figure 14. Clearly, the spatial distribution patterns of SSS were similar with MODIS-derived SSS (Figure 13), but the SMAP-estimated SSS has coarse spatial resolution and no coverage in very nearshore waters, probably may not provide details in SSS spatial

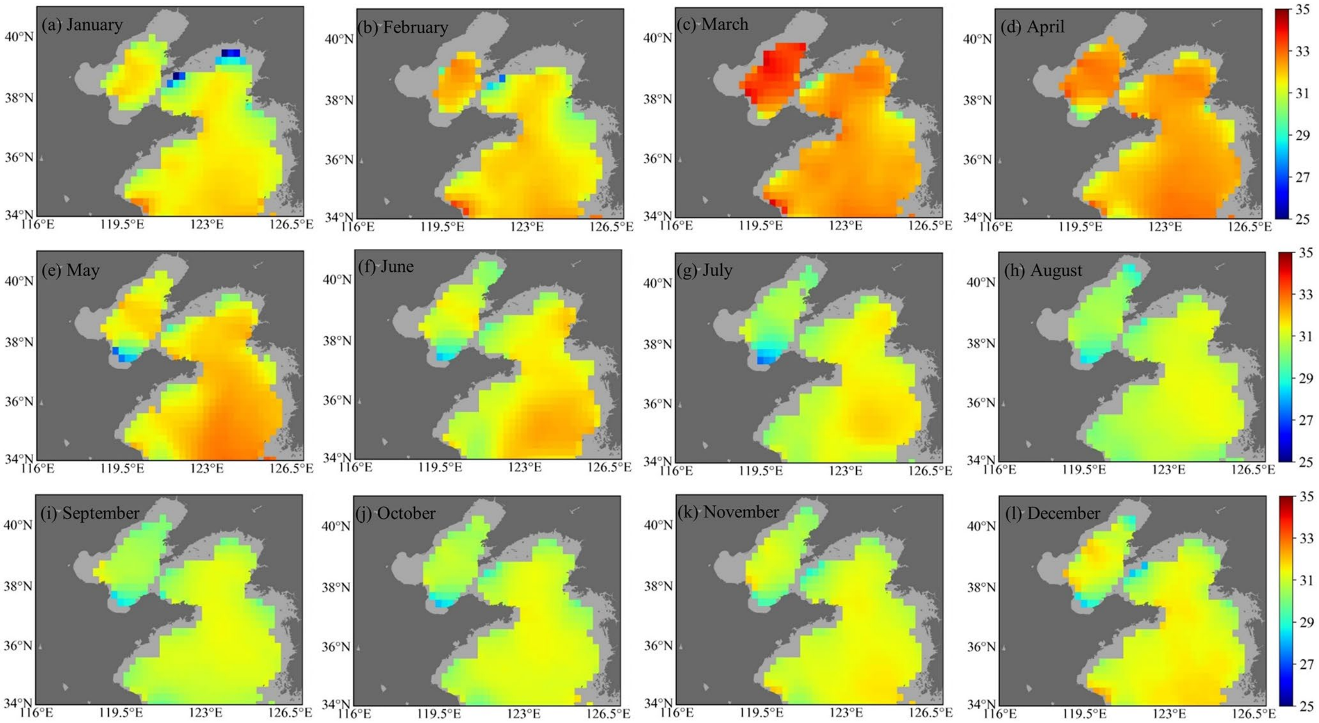

Figure 14. Distributions of monthly SMAP-derived SSS in the Yellow sea and Bohai sea during April 2015–July 2022.

variations. In additional, compare with the MODIS-derived SSS, the SMAP-estimated SSS was lower in the YS in cold season, and higher in central BS in summer, which may be attributed to the SMAP performance is worse with a high uncertainty in the Chinese marginal seas due to the land-sea contamination and low SST (Jang et al., 2021).

5. Conclusions

Although microwave satellite sensors have been deployed for SSS observation, SSS products with high temporal and spatial resolution remain a challenge due to the coarse spatial resolution and limited skill in nearshore waters. In this study, a RF model was first developed to accurately estimate the SSS for the ECS with a spatial resolution of ∼1 km based on a large synchronous data set of in situ SSS and MODIS-derived $R_{rs}(\lambda)$ and SST. The SSS model performed best with the $R_{rs}(412)$, $R_{rs}(488)$, $R_{rs}(555)$, $R_{rs}(667)$, SST and JD as model inputs. The accuracy of the SSS model was also validated with a RMSE of 0.66 using an independent data set, which was in agreement with the performance of the model training. The SSS model responded to uncertainties of each input showed that the estimated SSS variations induced by errors in SST and R_{rs} were all within the uncertainty of the RF model. The RF-based SSS model has been successfully applied in the Chinese marginal seas, indicating that RF model is an efficient tool to monitor the temporal and spatial variations in SSS and the formation of Changjiang River plume in the ECS. This study shows that the RF-based SSS model is a robust approach for SSS modeling from space in the coastal oceans although it is a challenging mission due to the high uncertainties of the satellite products in coastal oceans. This model could be adapted to other regions where the hydrology environment is completely different from the ECS by retraining the model with a local data set.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The raw sea surface salinity data for RF model development are available at https://doi.org/10.5281/zenodo.8019553 (Liu, 2023).

Acknowledgments

This study was financially supported by the National Thousand Talents Program for Foreign Experts (Grant WQ20133100150), Vulnerabilities and Opportunities of the Coastal Ocean (Grant SKLEC-2016RCDW01), Marginal Seas (MARSEAS) (Grants SKLEC Taskteam project), and Innovative Talents International Cooperation Training Project (Grant China Scholarship Council-202106140007). RB was also supported by the Norwegian Research Council Project CE2COAST—Downscaling Climate and Ocean Change to Services: Thresholds and Opportunities (Project No. 321890) through the 2019 "Joint Transnational Call on Next Generation Climate Science in Europe for Oceans" initiated by JPI Climate and JPI Oceans. QZ was supported by the Outstanding Postdoctoral Scholarship of the State Key Laboratory of Marine Environmental Science at Xiamen University. We sincerely thank the people who worked on these cruises of the "Shiptime Sharing Project of National Natural Science Foundation of China" and in the laboratory. We acknowledge SOCAT and WOD for providing sea surface salinity data set, NASA for providing MODIS satellite data and SMAP-derived SSS data. We also deeply thank Xiong et al. (2019, 2020) providing multiple cruise data from 2005 to 2018 in the ECS.

References

Ahn, Y. H., Shanmugam, P., Moon, J. E., & Ryu, J. H. (2008). Satellite remote sensing of a low-salinity water plume in the East China Sea. *Annales Geophysicae*, *26*(7), 2019–2035. https://doi.org/10.5194/angeo-26-2019-2008

Bai, Y., Pan, D. L., Cai, W. J., He, X. Q., Wang, D. F., Tao, B. Y., & Zhu, Q. K. (2013). Remote sensing of salinity from satellite-derived CDOM in the Changjiang River dominated East China Sea. *Journal of Geophysical Research: Oceans*, *118*(1), 227–243. https://doi.org/10.1029/2012JC008467

Bakker, D. C. E., Alin, S. R., Becker, M., Bitting, H. C., Castaño-Primo, R., Feely, R. A., et al. (2022). Surface ocean CO_2 Atlas database version 2022 (SOCATv2022) (NCEI Accession 0253659) Dataset. *NOAA National Centers for Environmental Information*. [indicate subset used]. https://doi.org/10.25921/1h9f-nb73

Bao, S. L., Zhang, R., Wang, H. Z., Yan, H. Q., Chen, J., & Wang, Y. J. (2021). Correction of satellite sea surface salinity products using ensemble learning method. *IEEE Access*, *11*, 17870–17881. https://doi.org/10.1109/ACCESS.2021.3057886

Bowers, D. G., & Brett, H. L. (2008). The relationship between CDOM and salinity in estuaries: An analytical and graphical solution. *Journal of Marine Systems*, *73*(1–2), 1–7. https://doi.org/10.1016/j.jmarsys.2007.07.001

Boyer, T. P., Baranova, O. K., Coleman, C., Garcia, H. E., Grodsky, A., Locarnini, R. A., et al. (2018). World ocean database 2018. In A. V. Mishonov, (Eds.), *NOAA Atlas NESDIS 87*. Retrieved from https://www.ncei.noaa.gov/sites/default/files/2020-04/wod_intro_0.pdf

Carder, K. L., Chen, F. R., Lee, Z. P., Hawes, S. K., & Cannizzaro, J. P. (2003). MODIS ocean science team algorithm theoretical basis document. *ATBD*, *19*, 67. Retrieved from https://modis.gsfc.nasa.gov/data/atbd/atbd_mod19.pdf

Chakraborty, A., Sharma, R., Kumar, R., & Basu, S. (2014). A seek filter assimilation of sea surface salinity from Aquarius in an OGCM: Implication for surface dynamics and thermohaline structure. *Journal of Geophysical Research: Oceans*, *119*(8), 4777–4796. https://doi.org/10.1002/2014JC009984

Chen, C. T. A. (2009). Chemical and physical fronts in the Bohai, Yellow and East China seas. *Journal of Marine Systems*, *78*(3), 394–410. https://doi.org/10.1016/j.jmarsys.2008.11.016

Chen, S. L., & Hu, C. M. (2017). Estimating sea surface salinity in the northern Gulf of Mexico from satellite ocean color measurements. *Remote Sensing of Environment*, *201*, 115–132. https://doi.org/10.1016/j.rse.2017.09.004

Chen, S. L., Hu, C. M., Barnes, B. B., Wanninkhof, R., Cai, W. J., Barbero, L., & Pierrot, D. (2019). A machine learning approach to estimate surface ocean $p\mathrm{CO}_2$ from satellite measurements. *Remote Sensing of Environment*, *228*, 203–226. https://doi.org/10.1016/j.rse.2019.04.019

Chen, S. L., Hu, C. M., Barnes, B. B., Xie, Y. Y., Lin, G., & Qu, Z. F. (2019). Improving ocean color data coverage through machine learning. *Remote Sensing of Environment*, *222*, 286–302. https://doi.org/10.1016/j.rse.2018.12.023

Chen, X. Y., Wang, X. H., & Guo, J. S. (2006). Seasonal variability of the sea surface salinity in the East China Sea during 1990–2002. *Journal of Geophysical Research*, *111*(C5), C05008. https://doi.org/10.1029/2005JC003078

Choi, J. K., Son, Y. B., Park, M. S., Hwang, D. J., Ahn, J. H., & Park, Y. G. (2021). The applicability of the geostationary ocean color imager to the mapping of sea surface salinity in the East China Sea. *Remote Sensing*, *13*(14), 2676. https://doi.org/10.3390/rs13142676

de Boyer Montégut, C., Mignot, J., Lazar, A., & Cravatte, S. (2007). Control of salinity on the mixed layer depth in the world ocean: 1. General description. *Journal of Geophysical Research*, *112*, C06011. https://doi.org/10.1029/2006JC003953

Del Vecchio, R., & Blough, N. V. (2004). Spatial and seasonal distribution of chromophoric dissolved organic matter and dissolved organic carbon in the Middle Atlantic Bight. *Marine Chemistry*, *89*(1–4), 169–187. https://doi.org/10.1016/j.marchem.2004.02.027

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., et al. (2010). The soil moisture active passive (SMAP) mission. *Proceedings of the IEEE*, *98*(5), 704–716. https://doi.org/10.1109/JPROC.2010.2043918

Font, J., Camps, A., Borges, A., Martín-Neira, M., Boutin, J., Reul, N., et al. (2010). SMOS: The challenging sea surface salinity measurement from space. *Proceedings of the IEEE*, *98*(5), 649–665. https://doi.org/10.1109/JPROC.2009.2033096

Geiger, E. F., Grossi, M. D., Trembanis, A. C., Kohut, J. T., & Oliver, M. J. (2013). Satellite-derived coastal ocean and estuarine salinity in the Mid-Atlantic. *Continental Shelf Research*, *63*, S235–S242. https://doi.org/10.1016/j.csr.2011.12.001

Hu, C. M., Feng, L., & Guan, Q. (2021). A machine learning approach to estimate surface chlorophyll *a* concentrations in global oceans from satellite measurements. *IEEE Transactions on Geoscience and Remote Sensing*, *59*(6), 4590–4607. https://doi.org/10.1109/TGRS.2020.3016473

Hu, C. M., Feng, L., & Lee, Z. P. (2013). Uncertainties of SeaWiFS and MODIS remote sensing reflectance: Implications from clear water measurements. *Remote Sensing of Environment*, *133*, 168–182. https://doi.org/10.1016/j.rse.2013.02.012

Hu, C. M., Muller-Karger, F. E., Biggs, D. C., Carder, K. L., Nababan, B., Nadeau, D., & Vanderbloemen, J. (2003). Comparison of ship and satellite bio-optical measurements on the continental margin of the NE Gulf of Mexico. *International Journal of Remote Sensing*, *24*(13), 2597–2612. https://doi.org/10.1080/0143116031000067007

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Tree-based methods. In *An introduction to statistical learning Springer texts in Statistics* (Vol. 103, pp. 303–328). Springer. https://doi.org/10.1007/978-1-4614-7138-7_8

Jang, E., Kim, Y. J., Im, J., & Park, Y. G. (2021). Improvement of SMAP sea surface salinity in river-dominated oceans using machine learning approaches. *GIScience and Remote Sensing*, *58*(1), 138–160. https://doi.org/10.1080/15481603.2021.1872228

JPL. (2020). JPL CAP SMAP sea surface salinity products. Ver. 5.0. PO.DAAC, CA, USA Dataset https://doi.org/10.5067/SMP50-3TMCS

Kim, D. W., Park, Y. J., Jeong, J. Y., & Jo, Y. H. (2020). Estimation of hourly sea surface salinity in the East China Sea using geostationary ocean color imager measurements. *Remote Sensing*, *12*(5), 755. https://doi.org/10.3390/rs12050755

Koblinsky, C. J., Hildebrand, P., LeVine, D., Pellerano, F., Chao, Y., Wilson, W., et al. (2003). Sea surface salinity from space: Science goals and measurement approach. *Radio Science*, *38*(4), 8064. https://doi.org/10.1029/2001RS002584

Köhl, A., Martins, M. S., & Stammer, D. (2014). Impact of assimilating surface salinity from SMOS on ocean circulation estimates. *Journal of Geophysical Research: Oceans*, *119*(8), 5449–5464. https://doi.org/10.1002/2014JC010040

Krzywinski, M., & Altman, N. (2017). Classification and regression trees. *Nature Methods*, *14*(8), 757–758. https://doi.org/10.1038/nmeth0917-928a

Li, X. S., Bellerby, R. G. J., Ge, J. Z., Wallhead, P., Liu, J., & Yang, A. Q. (2020). Retrieving monthly and interannual total-scale pH (pH_T) on the East China Sea shelf using an artificial neural network: ANN-pH_T-v1. *Geoscientific Model Development*, *13*(10), 5103–5117. https://doi.org/10.5194/gmd-13-5103-2020

Li, X. S., Bellerby, R. G. J., Wallhead, P., Ge, J. Z., Liu, J., Liu, J., & Yang, A. Q. (2020). A neural network-based analysis of the seasonal variability of surface total alkalinity on the East China Sea shelf. *Frontiers in Marine Science*, *7*, 219. https://doi.org/10.3389/fmars.2020.00219

Li, Y., Zou, C. F., Berecibar, M., Nanini-Maury, E., Chan, J. C.-W., van den Bossche, P., et al. (2018). Random forest regression for online capacity estimation of lithium-ion batteries. *Applied Energy*, *232*, 197–210. https://doi.org/10.1016/j.apenergy.2018.09.182

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*, 18–22.

Lie, H. J., Cho, C. H., Lee, J. H., & Lee, S. (2003). Structure and eastward extension of the Changjiang River plume in the East China Sea. *Journal of Geophysical Research*, *108*(C3), 3077. https://doi.org/10.1029/2001JC001194

Lin, C., Ning, X., Su, J., Lin, Y., & Xu, B. (2005). Environmental changes and the responses of the ecosystems of the Yellow Sea during 1976–2000. *Journal of Marine Systems*, *55*(3–4), 223–234. https://doi.org/10.1016/j.jmarsys.2004.08.001

Lin, C. L., Su, J. L., Xu, B. R., & Tang, Q. S. (2001). Long-term variations of temperature and salinity of the Bohai Sea and their influence on its ecosystem. *Progress in Oceanography*, *49*(1–4), 7–19. https://doi.org/10.1016/S0079-6611(01)00013-1

Liu, J. (2023). The raw sea surface salinity dataset for RF model development [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.8019553

Liu, J., Bellerby, R. G. J., Zhu, Q., & Ge, J. Z. (2023). Estimation of sea surface pCO_2 and air–sea CO_2 flux in the East China Sea using in-situ and satellite data over the period 2000–2016. *Continental Shelf Research*, *254*, 104879. https://doi.org/10.1016/j.csr.2022.104879

Liu, K. K., Chao, S. Y., Lee, H. J., Gong, G. C., & Teng, Y. C. (2010). Seasonal variation of primary productivity in the East China Sea: A numerical study based on coupled physical-biogeochemical model. *Deep-Sea Research II*, *57*(19–20), 1762–1782. https://doi.org/10.1016/j.dsr2.2010.04.003

Marghany, M., & Hashim, M. (2011). Retrieving seasonal sea surface salinity from MODIS satellite data using a Box-Jenkins algorithm. In *2011 IEEE international geoscience and remote sensing symposium* (pp. 2017–2020). https://doi.org/10.1109/IGARSS.2011.6049526

Palacios, S. L., Peterson, T. D., & Kudela, R. M. (2009). Development of synthetic salinity from remote sensing for the Columbia River plume. *Journal of Geophysical Research*, *114*(C2), C00B05. https://doi.org/10.1029/2008JC004895

Qi, J. F., Yin, B. S., Zhang, Q. L., Yang, D. Z., & Xu, Z. H. (2014). Analysis of seasonal variation of water masses in East China Sea. *Chinese Journal of Oceanology and Limnology*, *32*(4), 958–971. https://doi.org/10.1007/s00343-014-3269-1

Qing, S., Zhang, J., Cui, T. W., & Bao, Y. H. (2013). Retrieval of sea surface salinity with MERIS and MODIS data in the Bohai Sea. *Remote Sensing of Environment*, *136*, 117–125. https://doi.org/10.1016/j.rse.2013.04.016

Sasaki, H., Siswanto, E., Nishiuchi, K., Tanaka, K., Hasegawa, T., & Ishizaka, J. (2008). Mapping the low salinity Changjiang Diluted Water using satellite retrieved colored dissolved organic matter (CDOM) in the East China Sea during high river flow season. *Geophysical Research Letters*, *35*(4), L04604. https://doi.org/10.1029/2007GL032637

Sauzède, R., Claustre, H., Jamet, C., Uitz, J., Ras, J., Mignot, A., & D'Ortenzio, F. (2015). Retrieving the vertical distribution of chlorophyll-*a* concentration and phytoplankton community composition from in situ fluorescence profiles: A method based on a neural network with potential for global-scale applications. *Journal of Geophysical Research: Oceans*, *120*(1), 451–470. https://doi.org/10.1002/2014JC010355

Sun, D. Y., Su, X. P., Qiu, Z. F., Wang, S. Q., Mao, Z. H., & He, Y. J. (2019). Remote sensing estimation of sea surface salinity from GOCI measurements in the southern Yellow Sea. *Remote Sensing*, *11*(7), 775. https://doi.org/10.3390/rs11070775

Sun, Y. G., Sun, W. F., & Zhao, Y. J. (2022). Sea surface salinity dynamics in the Bohai Sea using MODIS data. In *IGARSS 2022–2022 IEEE international geoscience and remote sensing symposium* (pp. 7099–7102). https://doi.org/10.1109/IGARSS46834.2022.9883976

Urquhart, E. A., Zaitchik, B. F., Hoffman, M. J., Guikema, S. D., & Geiger, E. F. (2012). Remotely sensed estimates of surface salinity in the Chesapeake Bay: A statistical approach. *Remote Sensing of Environment*, *123*, 522–531. https://doi.org/10.1016/j.rse.2012.04.008

Vandermeulen, R. A., Arnone, R., Ladner, S., & Martinolich, P. (2014). Estimating sea surface salinity in coastal waters of the Gulf of Mexico using visible channels on SNPP-VIIRS. *Proceedings of SPIE*, *9111*, 911109. https://doi.org/10.1117/12.2053417

Xi, H. Y., Losa, S. N., Mangin, A., Soppa, M. A., Garnesson, P., Demaria, J., et al. (2020). Global retrieval of phytoplankton functional types based on empirical orthogonal functions using CMEMS GlobColour merged products and further extension to OLCI data. *Remote Sensing of Environment*, *240*, 111704. https://doi.org/10.1016/j.rse.2020.111704

Xiong, T. Q., Liu, P. F., Zhai, W. D., Bai, Y., Liu, D., Qi, D., et al. (2019). Export flux, biogeochemical effects, and the fate of a terrestrial carbonate system: From Changjiang (Yangtze River) Estuary to the East China Sea. *Earth and Space Science*, *6*(11), 2115–2141. https://doi.org/10.1029/2019EA000679

Xiong, T. Q., Wei, Q. S., Zhai, W. D., Li, C. L., Wang, S. Y., Zhang, Y. X., et al. (2020). Comparing subsurface seasonal deoxygenation and acidification in the Yellow Sea and northern East China Sea along the north-to-south latitude gradient. *Frontiers in Marine Science*, *7*, 686. https://doi.org/10.3389/fmars.2020.00686

Yang, S. C., Rienecker, M., & Keppenne, C. (2010). The impact of ocean data assimilation on seasonal-to-interannual forecasts: A case study of the 2006 El Niño event. *Journal of Climate*, *23*(15), 4080–4095. https://doi.org/10.1175/2010JCLI3319.1

Yu, X., Xiao, B., Liu, X. Y., Wang, Y. B., Cui, B. L., & Liu, X. (2017). Retrieval of remotely sensed sea surface salinity using MODIS data in the Chinese Bohai Sea. *International Journal of Remote Sensing*, *38*(23), 7357–7373. https://doi.org/10.1080/01431161.2017.1375570

Yu, X. L., Chen, S. L., & Chai, F. (2022). Remote estimation of sea surface nitrate in the California current system from satellite ocean color measurements. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 4203017. https://doi.org/10.1109/TGRS.2021.3095099

Zhang, J., Yu, Z. G., Raabe, T., Liu, S. M., Starke, A., Zou, L., & Brockmann, U. (2004). Dynamics of inorganic nutrient species in the Bohai seawaters. *Journal of Marine Systems*, *44*(3–4), 189–212. https://doi.org/10.1016/j.jmarsys.2003.09.010

Zhang, L. J., Xue, L., Song, M. Q., & Jiang, C. B. (2010). Distribution of the surface partial pressure of CO_2 in the southern Yellow Sea and its controls. *Continental Shelf Research*, *30*(3–4), 293–304. https://doi.org/10.1016/j.csr.2009.11.009

Zhang, Z. X., Qiao, F. L., Guo, J. S., & Guo, B. H. (2018). Seasonal changes and driving forces of inflow and outflow through the Bohai Strait. *Continental Shelf Research*, *154*, 1–8. https://doi.org/10.1016/j.csr.2017.12.012

Zhao, J., Temimi, M., & Ghedira, H. (2017). Remotely sensed sea surface salinity in the hyper-saline Arabian Gulf: Application to landsat 8 OLI data. *Estuarine, Coastal and Shelf Science*, *187*, 168–177. https://doi.org/10.1016/j.ecss.2017.01.008

Zhu, J. S., Huang, B. H., Zhang, R. H., Hu, Z. Z., Kumar, A., Balmaseda, M. A., et al. (2014). Salinity anomaly as a trigger for ENSO events. *Scientific Reports*, *4*(1), 6821. https://doi.org/10.1038/srep06821

Zhu, W., Yu, Q., Tian, Y. Q., Chen, R. F., & Gardner, G. B. (2011). Estimation of chromophoric dissolved organic matter in the Mississippi and Atchafalaya river plume regions using above-surface hyperspectral remote sensing. *Journal of Geophysical Research*, *116*(C2), C02011. https://doi.org/10.1029/2010JC006523