Multivariable integration method for estimating sea surface salinity in coastal waters from in situ data and remotely sensed data using random forest algorithm

Meiling Liu a, Xiangnan Liu a,*, Da Liu b, c, Chao Ding a, Jiale Jiang a

a China University of Geosciences, School of Information Engineering, 29 Xueyuan Road, Beijing 100083, China
b Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China
c University of Chinese Academy of Sciences, Beijing 100049, China

A R T I C L E I N F O

Article history:
Received 18 November 2013
Received in revised form 29 October 2014
Accepted 30 October 2014
Available online 10 November 2014

Keywords:
Random forest algorithm
Remote sensing
Multivariable integration
Sea surface salinity

A B S T R A C T

A random forest (RF) model was created to estimate sea surface salinity (SSS) in the Hong Kong Sea, China, by integrating in situ and remotely sensed data. Optical remotely sensed data from China’s HJ-1 satellite and in situ data were collected. The prediction model of salinity was developed by in situ environmental variables in the ocean, namely sea surface temperature (SST), pH, total inorganic nitrogen (TIN) and Chl-a, which are strongly related to SSS according to Pearson’s correlation analysis. The large-scale SSS was estimated using the established salinity model with the same input parameters. The ordinary kriging interpolation using in situ data and the retrieval model based on remotely sensed data were developed to obtain the large-scale input parameters of the model. The different number of trees in the forest (ntree) and the number of features at each node (mtry) were adjusted in the RF model. The results showed that an optimum RF model was obtained with mtry = 32 and ntree = 2000, and the most important variable of the model for SSS prediction was SST, followed by TIN, Chl-a and pH. Such an RF model was successful in evaluating the temporal-spatial distribution of SSS and had a relatively low estimation error. The root mean square error (RMSE) was less than 2.0 psu, the mean absolute error (MAE) was below 1.5 psu, and the absolute percent error (APE) was lower than 5%. The final RF salinity model was then compared with a multiple linear regression model (MLR), a back-propagation artificial neural network model, and a classification and regression trees (CART) model. The RF had a lower estimation error than the other three models. In addition, the RF model was used extensively under different periods and could be universal. This demonstrated that the RF algorithm has the capability to estimate SSS in coastal waters by integrating in situ and remotely sensed data.

& 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Ocean salinity is a key parameter in oceanic and climate studies, and it is a state variable that is essential for the determination of ocean circulation in response to the release of potential energy and forcings. Ocean salinity influences the density of water masses and actively participates in their formation and circulation (Yueh et al., 2001; Zine et al., 2008; Huang et al., 2008; Brassington and Divalaran, 2009). A few researchers have used microwave remote sensing to retrieve SSS with the launch of the Soil Moisture and Ocean Salinity (SMOS) and Aquarius (Banks et al., 2012; Camps et al., 2012; Reul et al., 2012; Boutin et al., 2012; Lee et al., 2012; Le Vine et al., 2014). The results showed that the large-scale estimation of SSS has accurate precision over the open ocean (Yin et al., 2012). However, a relatively high SSS bias appeared when the two satellites were used to investigate SSS of coastal waters and sub-mesoscale water (Zinc et al., 2007; Brassington and Divalaran, 2009). Other studies have explored the empirical relationships between colored dissolved organic matter (CDOM) and salinity using ocean color obtained by Sea-viewing Wide Field-of-view Sensor (SeaWiFS) and Moderate Resolution Imaging Spectroradiometer (MODIS) (Bowers et al., 2000; Binding and Bowers, 2003; Palacios et al., 2009; Del Vecchio and Blough, 2004; Bowers and Brett, 2008; Maisonet et al., 2009). To capture any additional information about the salinity of coastal water, other physicochemical parameters in the ocean such as SST and nitrogen should also be considered. In fact, few studies demonstrated that some physicochemical parameters in the ocean have an important effect on salinity (Brassington and Divalaran, 2009; Ravichandran et al., 2012; Duan et al., 2012). For example, Davies (2004) found that temperature and nitrogen strongly affect salinity. Therefore, the multi-physicochemical parameters rather

* Corresponding author. Fax: +86 10 82322095.
E-mail address: liuxncugb163.com (X. Liu).

http://dx.doi.org/10.1016/j.cageo.2014.10.016
0098-3004/© 2014 Elsevier Ltd. All rights reserved.
than a univariate parameter in the ocean as a proxy of salinity were considered to retrieve the SSS. To detect SSS accurately and extensively, it is important to develop a multivariate model for evaluating SSS. For example, some researchers developed multi-linear retrieval algorithms for SSS based on MODIS and MERIS data (Wong et al., 2007; Qing et al., 2013). Urquhart et al. (2012) used an artificial neural network (ANN) algorithm to estimate SSS in the Chesapeake Bay based on MODIS-Aqua data. To explore a range of statistical modeling options, Urquhart et al. (2012) have tried eight statistical models (such as ANN, random forest model) to predict SSS in the Chesapeake Bay. The existing work suggested that it is feasible to estimate SSS using statistical models. In this study, the random forest (RF) algorithm was selected to integrate multivariate parameters to detect SSS. It is well known that RF is a new method of data mining, and it has some advantages over statistical modeling approaches, such as the ability of modeling non-linear relationships, and the handling of continuous and categorical predictors (Breiman, 2001). Moreover, compared with other artificial intelligence models, such as the artificial neural network (ANN) model, the RF algorithm is quite robust to noise in its predictors, resistance to overfitting, implemented unbiased measure of error rate and implemented measures of variable importance (Grimm et al., 2008). Recently, research on RF has become an important topic and has a wide application in estimation (Prasad et al., 2006; Grimm et al., 2008; Oliveira et al., 2012; Urquhart et al., 2012). Therefore, the objective of this research was to apply the RF model to estimate SSS in coastal water based on multivariate parameters as a proxy of salinity.

2. Materials and methods

2.1. Study area

The Hong Kong Sea, Hong Kong Special Administrative Region, China (22°08′N–22°34′N, 113°49′E–114°30′E) was selected as the study area, which is surrounded by the land area of Hong Kong (Fig. 1). It is located in the northern area of the South China Sea, to the east of the Pearl River Estuary (PRE), China. It is a well-developed economic district with numerous industrial operations that have polluted the coastal areas. The study area has a warm and humid subtropical climate, and the Pearl River flows through large catchment areas into the PRE before finally reaching the South China Sea, which is the largest marginal sea on the western boundary of the Pacific Ocean.

2.2. Data collection and processing

2.2.1. In situ data

The dataset was collected from 76 monitoring stations along the Hong Kong Sea (Fig. 1). The study utilized monthly field surveys. In addition, 10 water control zones were identified in the study area; a detailed description of the water control zones and sample points of the study area is shown in Fig. 1. Sea surface measurements included nine types of physicochemical parameters (Table 1), namely sea surface salinity (SSS), total inorganic nitrogen (TIN), total suspended particles (TSP), dissolved oxygen (DO), sea surface temperature (SST), total volatile solids (TVS), chlorophyll-a (Chl-a), pH and total nitrogen (TN). We obtained the above dataset during 2003–2011 from the Environmental Protection Department, South China Sea Ocean Data Base (http://www.ocdb.csdb.

### Table 1

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Period</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST</td>
<td>Sea surface temperature</td>
<td>9 years during 2003-2011</td>
<td><a href="http://wqrc.epd.gov.hk">http://wqrc.epd.gov.hk</a></td>
</tr>
<tr>
<td>TN</td>
<td>Total nitrogen</td>
<td>Three years during 2009– 2011, 8 good images used (20100115, 2010321, 20100630, 2010918, 20110308, 20110425, 20110827, 20111128)</td>
<td><a href="http://www.cresda.com">http://www.cresda.com</a></td>
</tr>
<tr>
<td>TSP</td>
<td>Total suspended particles</td>
<td>9 years during 2003-2011</td>
<td><a href="http://ccd.a.gov.cn">http://ccd.a.gov.cn</a></td>
</tr>
<tr>
<td>SSS</td>
<td>Sea surface salinity</td>
<td>9 years during 2003-2011</td>
<td><a href="http://ccd.a.gov.cn">http://ccd.a.gov.cn</a></td>
</tr>
<tr>
<td>TIN</td>
<td>Total inorganic nitrogen</td>
<td>9 years during 2003-2011</td>
<td><a href="http://ccd.a.gov.cn">http://ccd.a.gov.cn</a></td>
</tr>
<tr>
<td>Chl-a</td>
<td>Chlorophyll-a</td>
<td>9 years during 2003-2011</td>
<td><a href="http://ccd.a.gov.cn">http://ccd.a.gov.cn</a></td>
</tr>
<tr>
<td>DO</td>
<td>Dissolved oxygen</td>
<td>9 years during 2003-2011</td>
<td><a href="http://ccd.a.gov.cn">http://ccd.a.gov.cn</a></td>
</tr>
<tr>
<td>TVS</td>
<td>Total volatile solids</td>
<td>9 years during 2003-2011</td>
<td><a href="http://ccd.a.gov.cn">http://ccd.a.gov.cn</a></td>
</tr>
<tr>
<td>pH</td>
<td>pH</td>
<td>9 years during 2003-2011</td>
<td><a href="http://ccd.a.gov.cn">http://ccd.a.gov.cn</a></td>
</tr>
</tbody>
</table>

* The HJ-1 satellite was successfully launched in China on September 6, 2008.
2.2.2. Remotely sensed data

Remotely sensed data, namely HJ-1 CCD images, were acquired from the China Center for Resources Satellite Data and Application (http://www.cresda.com). The HJ-1 satellite is equipped with a charge-coupled device (CCD) camera and hyperspectral imager (HSI) or infrared camera (IRS) and was successfully launched in China on September 6, 2008 to monitor the environment and natural disasters. The HJ-1 CCD image has four spectral bands (Band 1: 0.43–0.52 μm, Band 2: 0.52–0.60 μm, Band 3: 0.63–0.69 μm, Band 4: 0.76–0.90 μm) with 30 m × 30 m spatial resolution. The two identical CCD cameras in the HJ-1-A satellite and HJ-1-B satellite can capture images with a ground swath width of 700 km and revisit cycle of two days. The preprocessing of HJ-1 CCD imagery includes atmospheric and geometric corrections and resampling (120 × 120 m²). CCD images with large amounts of clouds in the study area were excluded. For this study, CCD images were collected from 2009 to 2011. Eight clear CCD images were used in the model. The specific information is displayed in Table 1.

2.3. Methodology

To accurately monitor SSS in coastal waters, the method for assessing the spatial distribution of SSS can be implemented using the following steps (see Fig. 2):

1. The Pearson correlation analysis is used to derive the parameters sensitivity to SSS.
2. The RF algorithm is applied to establish the relationship between sea environmental variables and SSS.
3. Large-scale sea environmental variables are calculated, the pH and temperature are spatially interpolated using ordinary kriging, and Chl-a and TIN are derived from CCD images.
4. The spatial distribution of SSS is mapped based on the established RF salinity model.
3. Model development

3.1. Input variables selection

One of the most important steps in the model development process is the determination of an appropriate set of input variables. Multiple regression analysis, principal component analysis (PCA), and Pearson correlation analysis are usually used to perform multivariate analysis (Capelli et al., 2000; Szefer et al., 2002). In this study, the Pearson correlation analysis was selected to identify which variables (SST, TN, TSP, TIN, Chl-a, DO, TVS and pH) control or significantly affect SSS in coastal waters. In this study, a value of $p < 0.05$ was considered to indicate a significant difference in the statistical analysis. The sea physicochemical parameters were derived by Pearson correlation analysis (Table 2). A high Pearson’s correlation coefficient indicated that the analyzed variables were significantly correlated with each other. As observed in Table 2, DO and pH were positively correlated to SSS, while other sea physicochemical parameters (e.g., TIN, SST, Chl-a, TSP, TN, and TVS) were negatively correlated with SSS. Obviously, five parameters, namely TIN, SST, pH, Chl-a and TN, had strong correlations with SSS, and the absolute value of Pearson’s correlation coefficient was greater than 0.50 for most of the year during 2003–2011. TN was excluded from the input variables because it was strongly related to TIN. In this study, therefore, the four parameters of TIN, Chl-a, SST and pH were taken as input variables for the SSS estimation.

3.2. Large-area variable acquisition

To predict the spatial distribution of SSS on a large scale, it is necessary to obtain a large-area input variable. In this study, the large-area variable acquisition is classified into two different schemes (Fig. 2). For SST and pH, spatial interpolation is used to acquire the large-area input variables. While for TIN and Chl-a, remote sensing retrieval was adopted to obtain large-area input variables.

3.2.1. Spatial interpolation

For SST and pH, the simplest methods of extrapolation and interpolation from a single point to an area have been used to obtain accurate data values at every desired point of the continuum. Ordinary kriging is very convenient as an estimation technique because of its simplicity and reliability. The theory of ordinary kriging is derived from that of regionalized variables (Oliver and Webster, 1990). In this study, the SST and pH were interpolated for each survey onto a regular grid of $120 \times 120$ m$^2$ resolution by ordinary kriging to keep consistency with the resampled CCD data spatial resolution. The spatial interpolation was conducted using the Geostatistical Analyst in ArcGIS tool. The true estimation accuracy of ordinary kriging methods was evaluated by error mean (EM) and root mean square (RMS) (Li, 2010). These values are computed by

$$EM = \frac{\sum [z(x, y) - z^*(x, y)]}{n}$$

$$RMS = \sqrt{\frac{\sum [z(x, y) - z^*(x, y)]^2}{n}}$$

where $n$ is the number of observations in the validation subset, $z(x, y)$ and $z^*(x, y)$ are observations and predictions in the sampling coordinate pair $z(x, y)$. Lower EM and RMS values indicate better estimation accuracy.

3.2.2. Remote sensing retrieval

Marine environmental parameters such as Chl-a, TIN, total suspended material (TSM) and CDOM are key parameters for monitoring water quality in the marine environment. A few studies suggest that large-scale spatial and temporal information on these parameters can be obtained by means of satellite remote sensing (Wang and Xu, 2008; Tilstone et al., 2011). Therefore, in this study, large-scale Chl-a and TIN as input variables were derived from optical remote sensing. Some researchers have demonstrated that single-band reflectance or the visible wavelength ratio, which are derived from different optical remote sensing data such as MODIS, SeaWiFS, MultiSpectral Scanner (MSS), and Thematic Mapper (TM), can be used for the derivation of Chl-a and TIN using statistical models (Silio-Calzada et al., 2008; Zhang et al., 2011). According to a previous study in combination with our experimental data, the retrieval models of Chl-a and TIN were constructed using the two spectral parameters derived from the CCD data (Fig. 3). The equation was computed as follows.

First, retrieval models of Chl-a

$$y = 0.47x^{1.285}$$

with a correlation coefficient $R=0.7016$, and where $y$ is the Chl-a concentration, $x$ is the spectral parameter, and $R_i$ is the reflectance of the $i$th band in CCD data.

Second, retrieval models of TIN are as follows:

$$y = 0.21x^2 - 0.21x + 0.36$$

and $x = \ln \left| \frac{R_4 - R_1}{R_4 + R_1} \right|$ with a correlation coefficient $R=0.6073$, and where $y$ is the TIN concentration, $x$ is the spectral parameter, and $R_i$ is the reflectance of the $i$th band in CCD data.

3.3. Model establishment

In this study, RF is used to establish a SSS retrieval model. RF is a popular and very efficient algorithm introduced by Breiman (2001), based on model aggregation ideas, for both classification and regression problems (Arun and Langmead, 2006; Diaz-Uriarte and de Andres, 2006; Grimm et al., 2008; Gislon et al., 2006; Guo et al., 2011). Recently, RF has mostly been applied for regression (Genuer et al., 2010; Vincenzi et al., 2011; Mutanga et al., 2012; Adusumilli et al., 2013). Within the training procedure, the RF algorithm produces multiple CART-like trees (Breiman, 2001), each of which is based on a bootstrap sample of the original training data. Numerous trees are generated within the algorithm and finally aggregated to give a single estimation, which is the average of the individual tree outputs (Fig. 2).

The random forest regression is a non-parametric regression approach. It is assumed that a random vector $\theta_k$ is generated for the $k$th tree in this procedure, independent of the past random vectors $\theta_1, ..., \theta_{k-1}$ but with the same distribution. Additionally, a regression tree is grown using the training set and $\theta_k$, resulting in a...
set of \( k \) trees \([h_1(X), h_2(X) \ldots h_k(X)]\), where \( h_k(X) = h(X, \theta_k) \), \( X = \{x_1, x_2, \ldots, x_p\} \) is a \( p \)-dimension input vector that forms a forest. The ensemble produces \( k \) outputs corresponding to each tree \( \hat{Y}_i = h_1(X), \hat{Y}_i = h_2(X), \ldots, \hat{Y}_i = h_k(X) \), where \( \hat{Y}_i \) is the \( k \)th tree output. To obtain the final output, an average of all tree predictions is calculated.

The procedure of the RF algorithm in this study is as follows:

1. Identify the input and output for developing random forest regression model. Given an input–output dataset \( \{(X_i, Y_i), (X_2, Y_2), \ldots, (X_n, Y_n)\} \), where \( X_i, i = 1, \ldots, n \) is an input vector containing SST, pH, Chl-a and TIN, \( Y_i \) is the total variance of the observed output and \( n \) represents the total number of out of bag samples. \( \text{Var}(Y_i) \) is the total variance of the observed output.

2. Draw a bootstrap sample from the available dataset. The RF model is typically made up of hundreds of decision trees. Here, each decision tree is built from a bootstrap sample of the original data set \( n = 200 \). Generally, approximately two-thirds of the samples will be included in a bootstrap sample and one-third will be left out, which can also be called the out-of-bag (OOB) samples (Freund and Schapire, 1996).

3. Adjust the essential parameters in the RF algorithm. RF in regression problems depends on three user-defined parameters: the number of trees (ntree) in the forest, the minimum number of data points in each terminal node (nodesize) and the number of features tried at each node (mtry). In this study, nodesize=4, and differing values of ntree and mtry were tested.

4. Evaluate the estimation error and accuracy. Interpretation in RF is facilitated by two important evaluation parameters. The first is mean square error (MSE). The optimum ntree in the forest is dependent on MSE (Adusumilli et al., 2013). The change in OOB error for each randomly permutated predictor gives an indication of the importance of that particular predictor. The second is the variance explained in the RF model \( R^2_{RF} \) (Acharjee et al., 2011). The variance explained \( R^2_{RF} \) from RF is a value that is relevant for the estimation of new independent samples. The greater the increase in the value of node purity, which dictates how the data are split at each node, the greater is the importance of that particular variable (Breiman, 2001).

The equations for MSE and \( R^2_{RF} \) are given as

\[
\text{MSE} \approx \text{MSE}^{\text{OOB}} = n^{-1} \sum_{i=1}^{n} (\hat{Y}(X_i) - Y_i)^2
\]

\[
R^2_{RF} = 1 - \frac{\text{MSE}^{\text{OOB}}}{\text{Var}(Y)}
\]

where \( \text{MSE}^{\text{OOB}} \) estimates determine how efficient the random forest prediction would be when it is exposed to unknown samples. \( R^2_{RF} \) is the variance explained in the RF model. \( \hat{Y}(X_i) \) is the predicted output corresponding to a given input sample. \( Y_i \) is the observed output and \( n \) represents the number of out of bag samples. \( \text{Var}(Y_i) \) is the total variance of the observed output.

5. Calculate the final output. Each tree acts as a regression function on its own, and the final output is taken as the average of the individual tree outputs.

\[
\mathbf{Y} = a_{\mathbf{y}}(\hat{Y}_1 = h_1(X))
\]

where \( \mathbf{Y} \) is the final output, and \( a_{\mathbf{y}}(\cdot) \) is the average of the individual tree outputs.

After finishing the training procedure of the RF algorithm to establish the relationship between the above four input parameters and SSS, we used an independent dataset \( n = 98 \) to validate the predictive performance of the random forest regression model.

3.4. Model evaluation parameters

Apart from these two important evaluation parameters (MSE and \( R^2_{RF} \)) in the RF model, we additionally reported the correlation coefficient \( R \), the root mean square error (RMSE), the absolute percent error (APE), and the mean absolute errors (MAE). The four parameters were computed by

\[
R = \frac{\sum_{i=1}^{n} (Y_{ai} - \bar{Y}) \sum_{i=1}^{n} (Y_{mi} - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (Y_{ai} - \bar{Y})^2 \sum_{i=1}^{n} (Y_{mi} - \bar{Y})^2}}
\]

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (Y_{ai} - Y_{mi})^2}{(n - 1)}}
\]

\[
\text{APE} = 1/n \sum_{i=1}^{n} \left| Y_{ai} - Y_{mi} \right| \times 100\%
\]

\[
\text{MAE} = 1/n \sum_{i=1}^{n} \left| Y_{ai} - Y_{mi} \right|
\]
where $y_{ai}$ and $y_{mi}$ are the predicted value and the measured value, and $\bar{y}_a$ and $\bar{y}_m$ are the average predicted value and the average measured value, respectively. The value $n$ is the sample number. APE, RMSE and MAE indicate estimation errors. Lower values of APE, RMSE and MAE indicate fewer estimation errors. The $R$ value indicates that there is an existing linear relationship between the measured and predicted values, and it is based on the goodness-of-fit in normal ordinary least square (OLS) regression.

4. Results

4.1. Model establishment

4.1.1. Parameter optimization

The RF estimation performance was optimized obtaining high values of $R_{RF}^2$ and low MSE. To obtain good estimation performance, we applied the RF algorithm with nodesize = 4, $ntree = 2000, 5000, 10,000, 15,000,$ and $20,000$ and $mtry = 2, 4, 8,$

Fig. 4. Dynamic results of random forest algorithm for estimating SSS during the training and validation process.
16 and 32. By trying different numbers of mtry and ntree in the RF algorithm, the optimum model was obtained according to the greatest \( R^2_{\text{pred}} \) and relatively lower MSE, when \( \text{mtry} = 32 \) and \( \text{ntree} = 2000 \). In this study, we find that \( R^2_{\text{pred}} \) is insensitive to ntree and at high ntree values also insensitive to mtry. That is, the total range and the differences in the \( R^2_{\text{pred}} \) and MSE between training settings of mtry and ntree were relatively small, namely the \( R^2_{\text{pred}} \) is approximately 0.85, and the MSE is approximately 2.85 with differences in the third decimal. This indicated that the RF algorithm was usually slightly stable in the models’ performance, when the model parameters were different in estimating sea biochemistry. It also agreed well with our previous study (Liu et al., 2013).

4.1.2. Performance evaluation

Based on the above analysis, the RF algorithm with ntree=2000, nodesize=4, and mtry value of 32 was selected as the optimum model to estimate SSS. In the training process of the RF model, SSS served as the output variable, and four parameters were taken as input variables, that is, SST, pH, Chl-a and TIN. The RF estimation performance is shown in Fig. 4(a, c, e) for the training in which the \( R^2_{\text{pred}} \) and MSE change with increasing ntree. Namely, the value of \( R^2_{\text{pred}} \) increased with increasing ntree, and MSE decreased with increasing ntree. In addition, \( R^2_{\text{pred}} \) and MSE remained stable when ntree was approximately 100. The \( R \) value between the measured and predicted value was 0.92. RMSE, APE and MAE were 1.68, 4.20% and 1.03, respectively (Fig. 4e). This result led to the conclusion that the RF model performed well in the estimation of SSS.

4.1.3. Variable importance

Input variable importance revealed that different input variables have different influences on the model (Fig. 5). The value of importance (%) is referred that each input variable importance accounts for the total input variable importance. From Fig. 5, the most important input variable for SSS estimation was SST, closely followed by TIN and then Chl-a and pH. This finding also agreed with our previous results where the maximum absolute value of \( R \) between input variables and salinity was SST, followed by TIN as well as Chl-a and pH (Liu et al., 2013). This result suggested that RF can serve as one of the methods for selecting the input variables of model construction. According to variable importance evaluation, SST and nitrogen have a significant important effect on the distribution of SSS, which agreed with the results of previous studies (Davies, 2004; Duan et al., 2012; Rushdi, 2012).

Fig. 5. Variable importance of different input variables in SSS estimation model based on random forest algorithm.

4.2. Model validation

To examine the credibility and stability of the established RF salinity model, the model was verified using 98 independent samples. SST, TIN, Chl-a and pH were taken as input variables. Fig. 4(b, d, f) shows the RF performance in predicting SSS in the process of validation. The value of \( R^2_{\text{pred}} \) increased with increasing ntree, and MSE decreased with increasing ntree. In addition, \( R^2_{\text{pred}} \) and MSE remained stable when ntree was approximately 100. It is further inferred that \( R^2_{\text{pred}} \) is insensitive to ntree. The constructed model provided better performance between the measured and model-predicted values of SSS for the validation sets with \( R \) values of 0.86. More detailed performance parameters for the RF model for estimating SSS are also shown in Fig. 4(f). According to the above four parameters for assessing the performance of the RF model, a proper model should have low RMSE, MAE and APE, and the value of \( R \) should be close to 1. While predicting SSS for the validation sets, lower RMSE, MAE and APE were achieved in the RF model with MSE = 1.64, APE = 3.61%, and MAE = 1.02. Based on the above analysis, with SST, TIN, Chl-a and pH as input variables constructed RF model has the capacity to estimate SSS.

4.3. Model application

4.3.1. Large-area variable acquisition

The spatial interpolation and remote sensing retrieval were used to obtain the large-area input variable. In the study area, SST and pH were obtained using ordinary kriging based on 76 measured datasets. Comparing the different semi-variogram models, the spherical models had relatively less predication error. Thus, the spherical models were applied to obtain the pH and SST on a large scale. Detailed statistics for spatial estimations for pH and SST in the study area are presented in Table 3. From Table 3, the mean pH in summer and autumn was greater than that of spring and winter, while the largest mean SST in the four seasons was found in summer, as expected. For estimation accuracy, pH and SST in the four seasons had low EM and RMS values. Regardless of the season, EM was below −0.001 and RMS lower than 0.12 for pH, whereas EM was below 0.01 and RMS lower than 0.6 for SST. This result indicated that the estimation accuracy of pH and SST by kriging interpolation methods was high.

Based on the above analysis, TIN and Chl-a in the large area were derived from CCD data according to the Eqs. \( (3) \) and \( (4) \). Detailed statistics for TIN and Chl-a in the study area are shown in Table 3. As observed in Table 3, the largest mean Chl-a value of the four seasons peaked in summer. While the relatively higher values of the mean TIN value occurred in spring, the mean TIN has a similar value during the other three seasons.

4.3.2. Large-area SSS estimation

The above constructed RF model based on in situ data was applied to calculate SSS in spring (March–May), summer (June–August), autumn (September–November) and winter (December–February) in the study area. In this study, four cloudless images were selected during a typical date for the four seasons: 21 March, 30 June, 18 September and 15 January 2010, which correspond to spring, summer, autumn and winter, respectively. The spatial distribution of SSS is shown in Fig. 6. For clarity, Fig. 6 displays the respective mean values of SSS in the four different seasons. The SSS map in spring, autumn and winter reveals an offshore gradient: SSS values increased with increasing distance from the coast, namely lower SSS was identified near the Hong Kong land area, including nearby Lantau Island, Kowloon and Hong Kong Island; however, relatively higher SSS levels were observed around the South China Sea. SSS near Hong Kong’s land area was lower
because of an increase in river runoff. The results agreed with the previous study where low SSS always appears in river run-off regions and semi-enclosed seas (Qi and Wei, 2012). The spatial distribution features of SSS in different seasons may be attributed to freshwater runoff from rivers and meteorological conditions (i.e., temperature).

In the study area, the SSS values exhibit a peak in spring and winter (Fig. 6a, d). The distribution of SSS did not significantly differ in spring and winter, with the seasonal averaged SSS value for most of the part concentrating between 32 and 33 psu, whereas the minimum values of SSS were recorded in autumn with the seasonal averaged SSS value normally below 31 psu (Fig. 6c). The spatial variability of SSS in the different seasons suggests that the physical and chemical composition of the sea has a significant effect on SSS. From Table 3, the minimum mean values of Chl-a and SST occur in the spring and winter. Furthermore, the data demonstrate that SSS has an opposite tendency with Chl-a and SST in spatial variability.

### 4.3.3. Large-area SSS validation

To examine the credibility and stability of the RF model, the model was verified using 76 datasets, derived from monitoring stations. Fig. 7 shows the plots between the measured and model-predicted values of SSS in the four seasons. The DB zone has a
Great estimate error in predicting SSS because there are relatively lower measured SSS values than in the other zones. This stems from land runoff from continental China. From Fig. 7, the model-predicted values agreed well with the measured values in nine zones. In other words, the constructed RF had good performance in the model application in all zones apart from the DB zone. More detailed estimate errors for the RF model for estimating SSS in 2010 are calculated in Table 4. As observed in Table 4, low RMSE, MAE and APE in all water control zones except the DB zone were achieved in the RF model for estimating SSS. To illustrate the ability of the RF model to be used widely, the DB zone was excluded from the total area. RMSE, MAE and APE in the total area are also shown in Table 4. Compared with SSS in summer and autumn, SSS in spring and winter had a lower estimation error with smaller RMSE, APE and MAE. This finding may be associated with different driving factors of SSS seasonal variation in different seasons. In summer and autumn, river runoff and rainfall largely affect the seasonal variation of SSS (Guo et al., 2013). Generally, a lower estimate error was obtained in the whole area (except DB zone), specifically the RMSE value ranged from 0.2 to 1.9 and the MAE between 0.2 and 1.5, while the APE ranged from 0.4% to 5%. This result implied that the RF model was effective for estimating SSS on a large scale.

5. Discussion

5.1. Comparison of different models

To investigate whether the RF model estimations of SSS have an advantage over other models, the other three widely used statistical models were also developed to estimate SSS using the same input parameters. First, multiple linear regression (MLR) modeling was considered because MLR is one of the most widely used methodologies for expressing how a response variable depends on several independent variables. Second, a classification and regression tree (CART) was chosen because CART has similar algorithms with RF for predicting continuous variables and they are both tree-based data mining techniques. CART is a recursive partitioning method; it builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). The classic CART algorithm was proposed by Breiman (1996) (Freund and Schapire, 1996). RF consists of an ensemble of randomized CART (Breiman, 2001). Therefore, RF is an improvement based on the CART algorithm. Third, an artificial neural network (ANN) model was selected because ANN can determine the input–output relationships within a complicated system on the basis of the interconnections among the experimental data. ANNs have been used for a variety of applications, including estimation, classification, system modeling, signal processing, and noise filtering (Tumbo et al., 2002; Noh et al., 2006; Ghorbani et al., 2010). Back propagation (BP) neural networks are popular neural network architectures in ANN models (Foody, 1995; Pachepsky et al., 1996); thus, a BP neural-network model was established to estimate SSS.

Table 4

Statistics for estimation errors of SSS area based on monitoring stations in the study in 2010.

<table>
<thead>
<tr>
<th>Season</th>
<th>parameters</th>
<th>TC</th>
<th>S</th>
<th>PS</th>
<th>JB</th>
<th>DB</th>
<th>MB</th>
<th>NW</th>
<th>WB</th>
<th>EB</th>
<th>VH</th>
<th>Total area a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring</td>
<td>RMSE</td>
<td>0.44</td>
<td>0.99</td>
<td>0.67</td>
<td>0.98</td>
<td>8.53</td>
<td>0.79</td>
<td>0.85</td>
<td>0.69</td>
<td>1.50</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>APE (%)</td>
<td>1.16</td>
<td>2.69</td>
<td>1.88</td>
<td>2.07</td>
<td>31.11</td>
<td>1.96</td>
<td>1.87</td>
<td>1.70</td>
<td>3.66</td>
<td>2.13</td>
<td>2.11</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>0.37</td>
<td>0.89</td>
<td>0.62</td>
<td>0.69</td>
<td>7.47</td>
<td>0.65</td>
<td>0.61</td>
<td>0.55</td>
<td>1.23</td>
<td>1.23</td>
<td>0.70</td>
</tr>
<tr>
<td>Summer</td>
<td>RMSE</td>
<td>0.91</td>
<td>1.98</td>
<td>0.57</td>
<td>1.60</td>
<td>15.67</td>
<td>3.27</td>
<td>1.26</td>
<td>0.82</td>
<td>1.32</td>
<td>1.28</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>APE (%)</td>
<td>2.09</td>
<td>6.01</td>
<td>1.46</td>
<td>3.38</td>
<td>99.51</td>
<td>10.15</td>
<td>3.18</td>
<td>2.12</td>
<td>4.53</td>
<td>3.50</td>
<td>4.91</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>0.68</td>
<td>1.58</td>
<td>0.47</td>
<td>1.10</td>
<td>13.87</td>
<td>2.86</td>
<td>1.00</td>
<td>0.65</td>
<td>1.49</td>
<td>1.06</td>
<td>1.41</td>
</tr>
<tr>
<td>Autumn</td>
<td>RMSE</td>
<td>0.80</td>
<td>1.10</td>
<td>0.73</td>
<td>1.06</td>
<td>12.40</td>
<td>2.75</td>
<td>0.40</td>
<td>0.49</td>
<td>0.83</td>
<td>1.55</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>APE (%)</td>
<td>2.07</td>
<td>2.92</td>
<td>1.74</td>
<td>2.44</td>
<td>72.79</td>
<td>7.39</td>
<td>1.44</td>
<td>1.24</td>
<td>2.07</td>
<td>4.45</td>
<td>3.49</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>0.58</td>
<td>0.82</td>
<td>0.54</td>
<td>0.75</td>
<td>10.61</td>
<td>2.23</td>
<td>0.35</td>
<td>0.35</td>
<td>0.64</td>
<td>1.28</td>
<td>1.03</td>
</tr>
<tr>
<td>Winter</td>
<td>RMSE</td>
<td>0.65</td>
<td>0.09</td>
<td>0.13</td>
<td>0.03</td>
<td>7.13</td>
<td>0.14</td>
<td>0.18</td>
<td>0.16</td>
<td>0.30</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>APE (%)</td>
<td>1.29</td>
<td>0.19</td>
<td>0.32</td>
<td>0.06</td>
<td>22.05</td>
<td>0.37</td>
<td>0.47</td>
<td>0.32</td>
<td>0.38</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>0.41</td>
<td>0.06</td>
<td>0.10</td>
<td>0.02</td>
<td>5.28</td>
<td>0.12</td>
<td>0.15</td>
<td>0.10</td>
<td>0.12</td>
<td>0.13</td>
<td></td>
</tr>
</tbody>
</table>

a The nine other water zones except DB zone were included in total area.
addition, different numbers of hidden layers and different numbers of neurons in each hidden layer were used to obtain the optimum model architectures. Therefore, MLR, CART, ANN (i.e., BP) and RF models were compared to estimate SSS.

For clarity, the training and validation set were verified using four different models. From Fig. 8, it is observed that the RF model performed better than the other three models, regardless of the training data and validation data. Furthermore, several performance evaluation parameters for all models used in estimating SSS are summarized (Table 5).

As said, according to the four parameters for assessing the performance of the models, the optimal model should have the lowest RMSE, APE and MAE, and the value of R should be close to 1. As observed in Table 5, the optimal model for estimating SSS was RF, with \( R > 0.85 \), RMSE < 1.7, APE < 4.5% and MAE < 1.10. Additionally, of all four models, the accuracy of the RF model estimations was always better than the other models for estimating SSS. This study reveals that the RF model can be used for predicting SSS having a broader ability in predicting it when compared to the other models.

5.2. Application of the model in other years

To illustrate the ability of the RF model to be used extensively, we apply it to another year. Likewise, four cloudless images were collected during typical dates in four season: 25 April, 27 August, 28 November and 1 January 2011, which correspond to spring, summer, autumn and winter, respectively. To compare the results from the different years, the methods of collecting and analyzing the data are the same as the above discussion. Fig. 9 shows the full temporal and spatial distribution of SSS in the study area. This figure clearly shows that the variability of SSS was spatially and seasonally large. Low SSS values were observed in summer, and high SSS values were found in spring, autumn, and winter. SSS values in the eastern area were persistently higher than those found in the western region in spring, while there is an opposite distribution pattern in the three other seasons. Compared with 2010, there is a relatively higher SSS value in 2011. Notably, the SSS value for most of the region was greater than 32 psu in spring, autumn, and winter. Furthermore, the drastic changes in SSS that occur in all seasons from one year to the next may be attributed to the following reasons: SSS in coastal waters was strongly influenced by meteorological forces (i.e., East Asia monsoon, heat waves), hydrological features (i.e., upwelling, oceanic fronts, or a mixing with the subsurface waters), and the general circulation in the study region (Jilan, 2004; Chen and Wang, 2006; Chen, 2008; Luo et al., 2011; Guo et al., 2013; Nan et al., 2014). However, the driving factors of the SSS inter-annual variation will need further clarification.
To investigate whether RF has good prediction accuracy, Fig. 10 shows the plots between the measured and model-predicted values of SSS based on 76 monitoring stations for 2011. From Fig. 10, the constructed RF obtained relatively low estimate errors in nine water control samples. More detailed estimate errors for the RF model for estimating SSS in 2011 are calculated in Table 6. According to Tables 4 and 6, 2011 performances in the whole area are worse than 2010. As observed in Table 6, low RMSE, MAE and APE in all water control zones apart from the DB zone were achieved in RF for predicting SSS. The DB zone has a great estimate error in predicting SSS because there is a lower measured SSS value than in the other zones. The low SSS value in the DB zone was strongly influenced by hydrological conditions, namely the large runoff of the Pearl River, and thus, it cannot be effectively detected by the RF algorithm. Apart from the DB zone, there is still a relatively greater estimate error (RMSE > 2 psu) in certain seasons of some zones, such as spring and summer in JB and summer and autumn in MB. Fortunately, a low estimate error was obtained in the whole area (except the DB zone), namely the RMSE value ranged from 0.8 to 1.8, MAE between 0.7 and 1.5, while APE ranged from 2% to 5%.

Based on the above analysis, the RF model for predicting SSS performed well in the two years, confirming that the RF model can be applied to other periods and is universally applicable. The primary reason is that RF can serve as a powerful tool for dealing with high dimensional non-linear relationships for the continuous predictors, and it also involves ocean environmental parameters as input variables, which facilitates the application of the RF model in different periods and thus increases the ability of the model to be used extensively.

6. Conclusion

In this study, an attempt was made to estimate SSS in coastal waters from the multi-physicochemical parameters as a proxy of salinity using RF algorithm. Pearson correlation analysis was used...
to extract the relationships between sea physicochemical parameters and SSS and the results indicated that SST, pH, Chl-a and TIN had a strong correlation with SSS. In addition, the most important factor for SSS estimation among them was SST, followed by TIN and then Chl-a and pH. This study has shown that the RF salinity model based on four input variables is useful for assessing the temporal–spatial distribution of SSS. After a thorough training and validation of the RF algorithm, the SSS was estimated for two years. The satisfactory performances of RF model establishment, model validation and model application were achieved with RMSE less than 2.0 psu, MAE below 1.5 psu and APE lower than 5%. Additionally, the RF robustness is also checked against three other models (MLR, CART, ANN) and it is showed that the RF salinity model had lower estimation errors. At the same time, the RF model can be successfully applied to other years. The RF model is thus reproducible and can be used extensively under different periods. This result suggests that the RF salinity model is capable of estimating the SSS of coastal waters. This demonstrated that the multi-physicochemical parameters in the ocean as a proxy of salinity have the potential to estimate SSS in coastal waters on a broad scale by integrating in situ and remotely sensed data.

In conclusion, we have presented a method for estimating SSS in coastal waters by capitalizing on the relationships of physicochemical parameters with SSS. Such method may offer insights into assessing other sea chemistry parameters, such as ocean nutrient fertility.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No. U0933005) and the Fundamental Research Funds for the Central Universities (No. 2-9-2012-18). The authors wish to thank the anonymous reviewers for their constructive comments that helped improve the scholarly quality of the paper.

References


