On the variations of sea surface $pCO_2$ in the northern South China Sea: A remote sensing based neural network approach

Young-Heon Jo,¹ Minhan Dai,² Weidong Zhai,² Xiao-Hai Yan,¹,² and Shaoling Shang²

Received 8 November 2011; revised 2 April 2012; accepted 1 July 2012; published 18 August 2012.

Using a neural networking (NN) approach, we developed an algorithm primarily based upon sea surface temperature (SST) and chlorophyll (Chl$_a$) to estimate the partial pressure of carbon dioxide ($pCO_2$) at the sea surface in the northern South China Sea (NSCS). Randomly selected in situ data collected from May 2001, February and July 2004 cruises were used to develop and test the predictive capabilities of the NN based algorithm with four inputs (SST, Chl$_a$, longitudes and latitudes). The comparison revealed a high correlation coefficient of 0.98 with a root mean square error (RMSE) of 6.9 μatm. We subsequently applied our NN algorithm to satellite SST and Chl$_a$ measurements, with associated longitudes and latitudes, to obtain surface water $pCO_2$. The resulting monthly mean $pCO_2$ map derived from the satellite measurements agreed reasonably well with the in situ observations showing a generally homogeneous distribution in the offshore regions. The $pCO_2$ exerts a very dynamic feature in nearshore regions, especially in the coastal upwelling and estuarine plume regions. We identified three low $pCO_2$ zones (<330 μatm), two of which are influenced by coastal upwelling: off Hainan island in the western part of the NSCS; and off Guangdong province in the eastern part of the NSCS. The path of the Pearl River plume on the shelf was another zone with low $pCO_2$. For the monthly mean $pCO_2$ variations estimated based on the MODIS-SST and -Chl$_a$ values, an RMSE of ~6 μatm may be attributable to the measurement errors associated with MODIS measurements. As a first order estimation, we used the same sampling periods of remote sensing and in situ measurements, and were able to estimate $pCO_2$ with an accuracy of 12.05 μatm for onshore regions and 13.0 μatm for offshore regions, but with combined uncertainties associated with the NN Testing algorithm and MODIS SST and Chl$_a$ measurements.


1. Introduction

Although coastal oceans are relatively small in terms of surface area, they potentially play a disproportionally important role in global carbon cycling [Gattuso et al., 1998; Walsh et al., 1981; Rabouille et al., 2001; Chen et al., 2007]. However, it remains particularly challenging to constrain the carbon fluxes in the coastal ocean due primarily to the large variability in both time and space [Borges et al., 2005; Cai et al., 2006; Chen and Borges, 2009; Dai et al., 2009]. Although the estimation of sea surface $pCO_2$ has primarily relied on ship board measurements [e.g., Takahashi et al., 2009], efforts have been made to adopt remote sensing measurements to estimate $pCO_2$ in light of the great potential of remote sensing techniques for large-scale ocean surface mapping with a reasonably good time resolution. Among others, a multiple linear regression approach has been applied [e.g., Bates et al., 1998; Lefèvre et al., 2002; Chen et al., 2007; Zhu et al., 2009].

In addition, Lefèvre et al. [2005] made an attempt to apply a neural network (NN) technique based on a self-organizing maps (SOM) approach to estimate surface $pCO_2$ in the Atlantic subpolar gyre. They use an NCEP/NCAR reanalysis sea surface temperature (SST) data set, together with position and time as inputs for the SOM algorithm in the study areas in order to assign $pCO_2$ values to combinations of these inputs. This SOM approach has shown the potential of high applicability for mapping global $pCO_2$ on the basis of two-dimensional gridded data. Later, Friedrich and Oschlies [2009a, 2009b] apply the SOM algorithm to data from ARGO floats and a Voluntary Observing Ship to estimate $pCO_2$ in the north Atlantic Ocean in addition to satellite SST and Chl$_a$ measurements. The root mean square errors (RMSEs) are 15.9 μatm from 15°N to 65°N [Friedrich and
Oschlies, 2009a] and 19.0 μatm basin-wide in the Atlantic Ocean [Friedrich and Oschlies, 2009b]. Telszewski et al. [2009] also estimate pCO₂ in the North Atlantic using a SOM algorithm. They used SeaWiFS-MODIS Chlₐ, NCEP/NCAR reanalysis SST and the Forecasting Ocean Assimilation Model (FOAM) mixed layer depth [Bell et al., 2000] and obtain an RMSE of 11.6 μatm. Recently, Hales et al. [2012] estimate coastal surface water pCO₂ in the central North American Pacific continental margin from remote-sensing data based on a SOM approach and a nonlinear semi-empirical model of surface water carbonate chemistry. Their model-predicted pCO₂ level agrees with the highly variable observations with a RMSE deviation of <20 μatm, and with a correlation coefficient (r) of 0.81.

[4] In this study, we applied another NN method, a feedforward backpropagation algorithm (FFBP), to examine the pCO₂ distribution in a large subtropical marginal sea with high spatial heterogeneity, the South China Sea (SCS). We contend that the FFBP algorithm was advantageous in the prediction of pCO₂ values from one-dimensional individual random ground truth measurements, and also enabled us to obtain a pCO₂ map using two-dimensional satellite measurements.

2. Study Area, Data Collection and Methods

2.1. Study Area

[5] This study was focused on the northern SCS (NSCS). Climatic variations in the atmosphere and in the upper ocean of the NSCS are primarily dominated by the Asian Monsoon [Hu et al., 2000]. During the southwest monsoon period, seasonal upwelling distributes large amounts of nutrients along the coast [Gan et al., 2009a, 2009b; Han et al., 2012]. Approximately 3.8 × 10¹¹ m³ year⁻¹ of runoff pours into the NSCS, approximately 90% of which is from the Pearl and Hanjiang Rivers [Han, 1998]. The Pearl River is a major river located in southern China, and the discharge typically results in a river plume extending up to a few hundred kilometers off the estuary mouth [Cao et al., 2011; Dai et al., 2008; Gan et al., 2010; Han et al., 2012]. Both the processes of seasonal upwelling and freshwater plumes are typically intensified in summer, when the rain-bearing southwest monsoon prevails and both contribute to a significant amount of new nutrients in the surface water, which typically stimulates primary production leading to mirrored changes in surface pCO₂ and dissolved oxygen (DO) [e.g., Dai et al., 2008; Cao et al., 2011; Han et al., 2012].

[6] Based on the seasonal air-sea CO₂ flux variation in the NSCS, Zhai et al. [2005, 2007] report that the NSCS is generally a CO₂ source to the atmosphere in spring, summer and autumn. Their study also shows that while pCO₂ nearshore is very dynamic and likely associated with the regional hydrodynamic and biogeochemical settings, the seasonal variation of pCO₂ is influenced to a large degree by SST on the outer shelf and slope.

2.2. In Situ Data Collection

[7] The data for this study were collected on three cruises onboard R/V Yanping II in May–June 2001, February and July 2004. Detailed sampling information can be found in several references [Zhai et al., 2005, 2007, 2009; Dai et al., 2008]. Figures 1a–1d show SST, Chlₐ, sea surface salinity (SSS) and pCO₂ along the cruise trajectories in May 2001. The trajectory of the 2001 cruise mainly covered a transect from the mouth of the Pearl River estuary (114° E, 22° N) to the southwest of the Dongsha Islands (115° 48′E, 20° 10′N). Data from this cruise have been analyzed for air-sea CO₂ fluxes offshore [Zhai et al., 2005], and for the carbonate system dynamics in the estuarine plume [Dai et al., 2008]. Using the 2001 cruise data, Zhu et al. [2009] show that the SST is higher offshore than nearshore. Chlₐ was very high as a result of the Pearl River plume, and the pCO₂ was similar to the SST distribution (Figures 1a and 1d). Note that Chlₐ in Figure 1b from the 2001 cruise was converted from fluorescence measurements based on Zhang et al. [2006], i.e., Chlₐ = 7.840 × fluorescence−0.56, for 0.05 mg/m³ < Chlₐ < 5.14 mg/m³. Figures 1e–1h and 1i–1l show SST, Chlₐ, SSS and pCO₂ along selected cruise trajectories in February and July 2004, respectively. Data from the July 2004 cruise have also been analyzed for the relationship between pCO₂ and DO [Zhai et al., 2009]. This July 2004 cruise showed similar features to that of the 2001 cruise. For example, SST was higher off the coast of the NSCS than nearshore. In addition, Chlₐ was relatively high and SSS relatively low as a result of the Pearl River plume. The February 2004 cruise was conducted in late winter and both the sampling configuration and data processing were similar to the July 2004 cruise [Zhai et al., 2009].

[8] The basic statistical properties of the data we used are summarized in Table 1. The maximum SST in July 2004 was 3.29°C and 8.87°C warmer than that in May 2001 and February 2004 due to seasonal heat advection through currents. Consequently, the minimum pCO₂ in July 2004 was 13.3 μatm higher than that of May 2001. However, while the maximum SST of July 2004 was 8.87°C warmer than that in February 2004 due to seasonal heat advection to the atmosphere through air-sea interface, the minimum pCO₂ in July 2004 was 27.6 μatm lower than that in February 2004. We should also note that in our analysis we excluded the data points within the Pearl River Estuary where the observations of pCO₂ revealed a complicated distribution pattern due to different physical and biogeochemical forces, as examined in previous studies [Dai et al., 2008].

[9] Since the range and distribution of input data are important to obtain accurate outputs from an NN approach, we made histograms of our observation data (Figure 2). The distribution histograms are superimposed with a normal density curve determined by the mean and standard deviation (SD) for the SST, Chlₐ, SSS and pCO₂ data. The kurtosis (a measure of whether the data are peaked or flat relative to a normal distribution when kurtosis is zero) and the skewness (a measure of symmetry) as well as the means and range of the observed data are summarized in Table 2.

[10] Most of our observations show high kurtosis, which tends to have a distinct peak near the mean, but which declines rather rapidly, and has heavy tails, especially for Chlₐ measurements (Figures 2b, 2f and 2j). Data sets with relatively low kurtosis tend to have a flat top near the mean rather than a sharp peak (e.g., SST in Figure 2i). At the same time, all of the SST, SSS and pCO₂ data showed a positive skewness while the Chlₐ data had a negative skewness, indicating that the frequency distributions of SST, pCO₂,
SSS and Chl\textsubscript{a} are not symmetric but are skewed toward lower Chl\textsubscript{a}, but higher SST, SSS and pCO\textsubscript{2}.

### 2.3. Satellite Remote Sensing Data

[11] The primary sources of satellite SST and Chl\textsubscript{a} data for this study came from MODIS (Moderate Resolution Imaging Spectroradiometer, http://oceancolor.gsfc.nasa.gov/ftp.html). MODIS detects a wider range of electromagnetic energy and takes measurements at three spatial resolutions. We used monthly mean MODIS-SST and -Chl\textsubscript{a} 4 km spatial resolutions for the NN to derive pCO\textsubscript{2}.

### 2.4. The pCO\textsubscript{2} Estimation Based on an NN Using an FFBP Algorithm

[12] A feedforward NN design involves a series of interconnected nodes that are divided into three basic layers: input, hidden, and output layers (Figure 3). The NN design has two phases of processing: (1) two propagations of signal patterns and (2) updating weights and biases. The first phase includes forward propagation of training patterns’ inputs and backward propagation of output activations through NN processes. The other phase is updating weights and biases to reduce prediction error. For similar neural network designs for specific applications, one can refer to Stogryn et al. [1994], Keiner and Yan [1998], Jones et al. [1999], and Tanaka et al. [2004]. The input layer receives data and passes them to the hidden layer, which is a system of layered nodes. Each node in a layer is called a neuron. These neurons consist of a linear summation function and a nonlinear activation function which enable the network to process the data. The different combinations of inputs are discussed later in this paper.

| Table 1. Maximum, Minimum, and Standard Deviation Values in May 2001, February and July 2004, and a Combination of These Years |
|-----------------|-----------------|-----------------|-----------------|
|                 | SST (°C)        | Chl\textsubscript{a} (mg/m\textsuperscript{3}) | SSS (psu)       | pCO\textsubscript{2} (µatm) |
| May 2001        |                 |                 |                 |                     |
| Maximum         | 28.68           | 13.02           | 34.12           | 437.4               |
| Minimum         | 25.0            | 0.03            | 30.05           | 246.3               |
| Standard Deviation | 0.75        | 3.2             | 0.84            | 40.4                |
| February 2004   |                 |                 |                 |                     |
| Maximum         | 23.1            | 8.8             | 34.7            | 402                 |
| Minimum         | 11.6            | 0.6             | 30.4            | 286                 |
| Standard Deviation | 5.7            | 1.7             | 0.87            | 32.5                |
| July 2004       |                 |                 |                 |                     |
| Maximum         | 31.97           | 13.4            | 34.49           | 438                 |
| Minimum         | 25.02           | 0.41            | 29.44           | 259.6               |
| Standard Deviation | 1.15     | 1.16            | 0.51            | 21.5                |
| 2001 and 2004   |                 |                 |                 |                     |
| Maximum         | 31.97           | 13.4            | 34.7            | 438                 |
| Minimum         | 11.6            | 0.03            | 29.4            | 246.3               |
| Standard Deviation | 5.7            | 1.8             | 0.74            | 32.5                |

Figure 1. SST, Chl\textsubscript{a}, SSS and pCO\textsubscript{2} along cruise tracks in the northern South China Sea for (a–d) May 2001, (e–h) February 2004 and (i–l) and July 2004. The basic statistics of the measurements are summarized in Table 1.
Figure 2. Histogram with superimposed normal density curve (red) for observations of SST, Chl\textsubscript{a}, SSS and pCO\textsubscript{2} from the (a–d) May 2001 cruise, (e–h) February 2004 cruise, (i–l) July 2004 cruise and (m–p) combined data sets from all cruises. The kurtosis and skewness of measurements are summarized in Table 2.
inputs to each neuron are first routed to the summation function and can be represented as in the following equation:

\[ y_j = \sum_{i=1}^{p} W_{ij} X_i + b, \]  

where \( X_i \) are the inputs entering the node from the previous layer; \( W_{ij} \) are the weights associated with each input/node connection \( i \); \( b \) is the bias node; and \( p \) is the number of inputs or nodes. The \( m \) and \( n \) represent the number of neurons in the first and the second hidden layers, respectively. Thus, the first and the second index for the nodes in the hidden layers (H) stand for number of the nodes and number of hidden layer (\( m \) for number of node in the first layer and \( n \) for number of). [13] The specific number of nodes to predict \( pCO_2 \) is discussed in section 3.1 with Figures 5–8. The result of this function is passed to the nonlinear activation function (the sigmoidal function in this study’s network). The output, \( y_j \), is mapped to a value between \((-1, 1)\). After the initial data pass through the neurons of the hidden and output layers, a value for the desired geophysical parameter is produced as shown in equation (2)

\[ Output = ag \left[ \sum_{k=1}^{j} W_{k} Z_{k} + \beta \right], \]  

where \( a \) is a scaling factor; \( g[ ] \) is the nonlinear activation function; \( W_{k} \) and \( Z_{k} \) are the weights and inputs between the neuron layers; \( \beta \) is the bias associated with the output layer; and \( j \) is the number of neurons in the previous hidden layer. This output is then compared to the target data, from which the network error is determined. The error is then backpropagated through the network in order to adjust the weights and biases associated with each neuron in the network layers. The repeated feedforward of the network inputs in conjunction with the backpropagation progressively minimizes the network error.

[14] The training process was conducted using Matlab Neural Network Toolbox [Beale et al., 2010]. The minimization of the network was performed by the Levenberg-Marquardt approximation method. For each network structure, the training procedure was run 100 times, and for each training procedure the weights and biases were randomly initialized using a technique following Nguyen and Widrow [2005], which was developed primarily for function approximation [Krasnopolsky et al., 1995].

[15] In addition to predicting \( pCO_2 \) with remote sensing measurements, a feedforward algorithm has been created to estimate winds [Stogryn et al., 1994; Krasnopolsky et al., 1995], ocean and atmospheric variables [Krasnopolsky et al., 1999, 2000; Buckton et al., 1999; Roberts et al., 2010], air

Table 2. The Mean, Range, Kurtosis and Skewness of Data Collected in May 2001, February and July 2004 and a Combination of These Years

<table>
<thead>
<tr>
<th></th>
<th>SST (°C)</th>
<th>Chl\textsubscript{a} (mg/m\textsuperscript{3})</th>
<th>SSS (psu)</th>
<th>( pCO_2 ) (µatm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 2001</td>
<td>Mean</td>
<td>27.8</td>
<td>2.59</td>
<td>33.4</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>3.7</td>
<td>13</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>5.8</td>
<td>5.2</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>−1.5</td>
<td>1.9</td>
<td>−2.6</td>
</tr>
<tr>
<td>February</td>
<td>Mean</td>
<td>18.5</td>
<td>2.3</td>
<td>33.7</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>11.5</td>
<td>8.3</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>2.4</td>
<td>4.1</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>−0.3</td>
<td>1.4</td>
<td>−0.98</td>
</tr>
<tr>
<td>July 2004</td>
<td>Mean</td>
<td>29.5</td>
<td>0.98</td>
<td>34.0</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>6.9</td>
<td>13</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>3.7</td>
<td>34.8</td>
<td>14.6</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>−1.0</td>
<td>4.6</td>
<td>−2.8</td>
</tr>
<tr>
<td>2001 and</td>
<td>Mean</td>
<td>24.6</td>
<td>1.7</td>
<td>33.8</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>20.3</td>
<td>13.8</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>1.7</td>
<td>11.4</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>−0.5</td>
<td>2.6</td>
<td>−1.8</td>
</tr>
</tbody>
</table>

Figure 3. Diagram for neural network design. The primary inputs are SST and Chl\textsubscript{a}, but different combinations of inputs (shown as “Others”) were used. Other inputs are (1) locations as a function of longitudes and latitudes, (2) salinity and (3) locations and salinity.
Figure 4. (a) Correlation and (b) RMSE values between in situ and NN Testing $p\text{CO}_2$ for different combinations of neurons in the first and second layers. The inputs are SST, Chl$_a$, longitudes and latitudes. The maximum correlation and minimum RMSE between in situ and NN Testing $p\text{CO}_2$ were obtained at 351 on the $x$ axis (shown by the blue line), which consisted of 20 nodes for the first layer and 17 nodes for the second layer.

Figure 5. Comparisons for both cruises with two inputs (SST and Chl$_a$) of (a) the values of in situ and of predicted $p\text{CO}_2$ (blue, in situ value; red, NN Training; black, NN Validation; green, NN Testing); (b) the correlation and RMSE values between in situ and predicted $p\text{CO}_2$, and sample numbers (colors as above); and (c) the RMSE values versus the $p\text{CO}_2$ numbers (blue, in situ $p\text{CO}_2$; green, predicted $p\text{CO}_2$; black, RMSE).
temperature [Jones et al., 1999], sea surface Chl\(_a\) and sediments [Keiner and Yan, 1998; Keiner and Brown, 1999], and water properties [Tanaka et al., 2004].

3. Results and Discussion

3.1. Estimation of \(pCO_2\) Using the NN

[16] The total number of measurements were 1082 for May 2001, 6033 for February 2004, 6790 for July 2004. All together they sum to 13,926 measurements. 70% of the total observations of SST and Chl\(_a\) were used to train the NN, defined as the "NN Training", which was subsequently compared with the in situ measurements. The other 30% were used for validating and testing with in situ \(pCO_2\) observations and these were called "NN Validation" and "NN Testing," respectively. The validation data set was used to stop training the NN once generalization began, and the testing data set was used to examine network performance. In order to predict certain parameter(s), the NN Testing results could be used to analyze the accuracy of the NN predictive power. Note that we also tested three different partitioning combinations by dividing 50%/70%/80% for NN Training and 25%/15%/10% for NN Validation and NN Testing, respectively. It was revealed that there were no significant differences in terms of NN performance.

[17] Figure 3 shows a diagram for our NN design, and we examined the sensitivity of different combinations of inputs to improve NN performance. The SST and Chl\(_a\) data were the primary inputs, and additional inputs included SSS, longitudes and latitudes. Figure 4 shows how we determined the numbers of neurons in the first and second layers. We used the numbers of neurons when the NN Testing showed the highest correlation coefficient and the least RMSE between predicted and in situ \(pCO_2\). Beyond those combinations of neurons, shown with a vertical line, the NN performances showed the same correlation and RMSE (Figure 4). We obtained 14 and 12 neurons for the first and second layers, which corresponded to 364 on the x axis as shown by the green line. Likewise, we examined the different inputs to analyze how we could obtain better predictabilities.

[18] Figures 5a, 6a, 7a and 8a show the comparisons between in situ \(pCO_2\) and \(pCO_2\) estimated from different performances (i.e., NN Training, NN Validation, and NN Testing) with two inputs (SST and Chl\(_a\)), three inputs (SST, Chl\(_a\), and SSS), four inputs (SST, Chl\(_a\), longitudes, and latitudes) and five inputs (SST, Chl\(_a\), SSS, longitudes and latitudes), respectively. Since many \(pCO_2\) values are highly correlated and thus are overlapped, we also presented the comparisons between in situ and predicted \(pCO_2\) in response to sample number as in Figures 5b, 6b, 7b and 8b with the correlation, the RMSE and the sample numbers for each experiment. The correlation coefficient and RMSE for the two inputs were 0.91 and 13.5 \(\mu\text{atm}\) for the combined cruises. Those for three inputs were 0.96 and 8.9 \(\mu\text{atm}\), those for four inputs were 0.98 and 6.9 \(\mu\text{atm}\), and those for five inputs were 0.98 and 6.8 \(\mu\text{atm}\). In addition to the combined
longitudes and latitudes (Figure 7), we examined the contribution of each longitude and latitude to the NN predictivity. Separately, each contributed to the same degree, but gave a higher RMSE (as much as 0.5 μatm) than the combined latitudes and longitudes. The correlation coefficient for both cases was the same (0.97).

Also shown in Figures 5c, 6c, 7c and 8c are the numbers of in situ and predicted pCO₂ values versus RMSE with different inputs. Most of the pCO₂ measurements (approximately 4000) are around 358.9 μatm, which is close to the mean values shown in Table 2. The distribution of predicted pCO₂ values is similar to the in situ pCO₂ values from low to high pCO₂ values. The RMSEs in response to the numbers of pCO₂ measurements were constant with relatively small ranges from 280 to 410 μatm. However, the RMSE was larger when pCO₂ exceeds 410 μatm.

As we added more inputs, the NN performance improved. The criteria regarding the numbers of neurons are discussed in section 3.1. The neurons for each experiment were 12 and 11 (Figure 5), 14 and 13 (Figure 6), 14 and 26 (Figure 7) and 32 and 36 (Figure 8), for the first and second layers. In order to illustrate how well they agreed with each other, we added two dotted lines (shown as P⁺ and P⁻) to indicate 20 μatm higher or less than the slope (shown with a solid black line), respectively. As the RMSE decreased, the pCO₂ values showed better agreement. One can see that there are a few points outside P⁺ and P⁻ in Figure 8. Although Figure 8 (with five inputs) showed the best result, we used the four inputs as in Figure 7 from satellite measurements to obtain pCO₂ maps (see section 3.2) because there are no available SSS satellite observations for coastal regions yet.

For the four-input experiment (Figure 7), we analyzed two groups of pCO₂ estimations. One group (group I) was confined to a range of pCO₂ between -20 μatm (< P⁺ < 20 μatm), which accounted for 97% of the total data points. The other group (Group II with 3% of the total data points) was confined by the upper (P⁺) and the lower bounds (P⁻), that is, above 20 μatm (P⁺ > 20 μatm) and below -20 μatm (P⁻ < -20 μatm), respectively. While Group I could be considered as a well-trained/validated pCO₂ group within the normal data distribution with the majority of the pCO₂ data (Figures 7a and 7c), Group II could be considered as a relatively poorly trained/validated group in the abnormal high and low data distribution with a small number of pCO₂ data points (Figures 7a and 7c). In order to analyze what made the two groups different, we examined the distribution of each input data set statistically by means of the kurtosis and skewness of the pCO₂ values derived from training and validating the NN. Group I showed that most of the kurtosis and skewness were closer to those of the original input data.
(Table 2) than those for Group II. Because of this, the maximum, minimum, mean and range were closer to those of the original inputs than those for Group II.

[22] In addition, the uncertainties in the predicted \( p_{CO_2} \) values resulting from the high and low \( p_{CO_2} \) values in the data (Figures 7a and 8a) were estimated based on Figures 7b and 8b. As Figure 7b shows, the RMSEs for \( NN\ Training \) and \( NN\ Testing \) are 5.9 and 6.9 \( \mu \text{atm} \), respectively. Likewise, Figure 8b shows that the RMSEs for \( NN\ Training \) and \( NN\ Testing \) were 5.6 and 6.8 \( \mu \text{atm} \), respectively. Thus, the differences between \( NN\ Training \) and \( NN\ Testing \) were about 1 \( \mu \text{atm} \), which is only 0.28% compared to the mean \( p_{CO_2} \) (358.9 \( \mu \text{atm} \) in Table 2). This small discrepancy suggests that the NN would perform well to predict \( p_{CO_2} \) in this study.

[23] In order to examine the predictivity of the NN algorithm developed based on the data collected in different seasons, we compared the RMSE of each prediction. The correlation coefficients and RMSEs based on the warm seasons data only, winter data only and the combined data sets were 0.97 with 5.9, 0.96 with 4.1, 0.98 with 6.9 \( \mu \text{atm} \) (Figure 7), respectively. If the ranges of inputs (SST and Chl\(_a\)) indeed represent their annual maximum and minimum (Table 2), the predictivity of the NN for other seasons should be achievable. In other words, as long as SST and Chl\(_a\) vary in the range of 11.6–31.97°C and 0.03–13.4 mg/m\(^3\), respectively, our NN is able to predict \( p_{CO_2} \) with a RMSE of ~6.9 \( \mu \text{atm} \).

### 3.2. Mapping \( p_{CO_2} \) Distribution Using Remote Sensing

[24] We applied the NN algorithm based upon four inputs to satellite measurements of SST, Chl\(_a\) and the associated longitudes and latitudes in order to obtain \( p_{CO_2} \) values of high spatial resolution. The satellite derived SST and Chl\(_a\) values are shown in Figure 9, which represents the case in July 2004. It should be pointed out that many missing Chl\(_a\) data due to cloud contamination was linearly interpolated for use in the NN inputs. Although we used data (May 2001, February and July 2004) for \( NN\ Training \), \( NN\ Validation \) and \( NN\ Testing \) (Figures 5–8) and employed the remote sensing data to predict \( p_{CO_2} \) values in the NSCS, we used \( p_{CO_2} \) values collected in July 2004 for further comparisons and error analysis. The general patterns of SST and Chl\(_a\) agree well with many prior researches and in situ measurements [Wu and Li, 2003; Huang et al., 2008; Gan et al., 2009a, 2009b; Jing et al., 2009], suggesting that our remote sensing SST and Chl\(_a\) data should be correct.

[25] We obtained the \( p_{CO_2} \) values based upon the satellite measurements of SST and Chl\(_a\) together with the associated longitudes and latitudes, and using our validated NN algorithm for the July 2004 case, as shown in Figure 10 with the locations of field measurements superimposed. Similarly to
the in situ measurements (Figure 1), low \( pCO_2 \) was located along the mouth of the Pearl River, along the river plume path and at upwelling zones in the nearshore NSCS.

Figure 10a shows the point-to-point comparisons between the in situ and NN derived \( pCO_2 \). It is clear that whenever the comparison could be drawn, such as along the coastal and offshore zones, discrepancies were relatively small, suggesting that the general features between the two data sets were in overall agreement. However, in the transitional zone to the north of Hainan Island, relatively large discrepancies were observed. The overall RMSE for the point-to-point comparisons in Figure 10 was 24.14 \( \mu \text{atm} \).

We must point out that any direct comparison between the in situ and the satellite derived \( pCO_2 \) was difficult because the two data sets represented different time scales with the predicted values given as monthly averages and the in situ \( pCO_2 \) as instantaneous measurements. Limiting factors also included the fact that our in situ \( pCO_2 \) measurements very often were spatially sparse, and therefore the RMSE values of 24.14 \( \mu \text{atm} \) may not have been meaningful in terms of NN predictivity. We noted that, when we had \( pCO_2 \) values at the same spatial and temporal resolution, we obtained an accuracy of 6.9 \( \mu \text{atm} \) (Figure 7).

In order to examine the climatological pattern of \( pCO_2 \) distribution in the NSCS, we used monthly mean SST and Chl\(_a\) and their SD values in July of each of six years from 2002 to 2007 to derive the sea surface \( pCO_2 \), and the results are shown in Figure 11. The distribution of the satellite-derived \( pCO_2 \) generally resembled that of the SST, as is suggested by the analysis of the in situ data [Zhai et al., 2005; 2007]; and the mean and SD of \( pCO_2 \) offshore were generally homogeneous. However, in the coastal regions, three zones of low \( pCO_2 \) (<330 \( \mu \text{atm} \)) could be identified in the eastern part of the NSCS off the Guangdong coast (Zone A), near the Taiwan Strait (Zone C), and the NSCS off Hainan Island (Zone D). An interesting area was Zone B, considered to be a transitional area between onshore and offshore regions. Zones A and D with their low mean \( pCO_2 \) values (Figure 11) are also characterized by low SST and high Chl\(_a\) (Figures 9a and 9b). Zones C and D, with their high \( pCO_2 \) values (Figure 11) are characterized by relatively low SST and high Chl\(_a\) (Figures 9a and 9b), likely to have resulted from enhanced biological consumption of CO\(_2\) fueled by new nutrients supplied from the summer coastal upwelling. These two upwelling zones are well documented based on various field observations and numerical modeling [Wu and Li, 2003; Gan et al., 2009a, 2009b; Jing et al., 2009; Cao et al., 2011].
Zone C is identical to the summer Pearl River plume path [Gan et al., 2009b; Han et al., 2012], and should also be related to the enhanced primary productivity induced by the river plume.

3.3. Error Analysis for Satellite Derived $pCO_2$

In order to examine the site dependent uncertainties in NN prediction, we divided all in situ $pCO_2$ data into two regions: onshore and offshore. The specific areas for onshore and offshore are defined in the caption of Figure 12. The RMSE for overall $pCO_2$ predictivity was 6.9 atm (Figure 7) with four inputs (SST, Chl $a$, Longitudes and Latitudes), but the RMSEs for onshore and offshore were 7.7 and 4 atm, respectively (Figure 12). The correlation (R), RMSE and numbers of data points for each case are indicated in Figures 12c and 12d. The RMSE resulting from NN performances for offshore was smaller than that for onshore. The RMSE due to the accuracy of satellite measurements is discussed as follows.

In order to evaluate how well the NN algorithm performed with satellite measurements, we estimated the uncertainties of the $pCO_2$ values associated with the satellite measurements. We defined relative errors in $pCO_2$ using a chain rule, i.e.,

$$d(pCO_2) = \frac{\partial(pCO_2)}{\partial(SST)}d(SST) + \frac{\partial(pCO_2)}{\partial(Chl_a)}d(Chl_a).$$

According to Brown and Minnett [1999], the MODIS-SST data can be considered as accurate to ±0.25°C, and according to Carder et al. [2003], the MODIS-Chl $a$ data can be considered as accurate to 0.13 mg/m$^3$ for the case II water based on University of South Florida data. In order to determine the coefficient in equation (3), we estimated two coefficients for onshore (relatively rich Chl $a$ and low SST) and offshore (relatively poor Chl $a$ and warm SST) regions.

Since we can estimate the $pCO_2$ changes due to SST or Chl $a$ variability by constraining Chl $a$ and SST based on our NN (Figure 7), we computed each of the parameters, $\frac{\partial(pCO_2)}{\partial(SST)}$ and $\frac{\partial(pCO_2)}{\partial(Chl_a)}$.

Corresponding $pCO_2$ values in function of latitudes for (c) onshore and (d) offshore. While offshore is defined as areas by lower than 21°N and higher than 112°E in this study, onshore is defined as remaining areas after offshore regions. All the in situ $pCO_2$ data are shown in Figure 11.

Figure 11. The (a) mean and (b) standard deviation of predicted $pCO_2$ in July of 2002 to 2007. The four different regions are indicated as Zones A to D, which are discussed in the text.

Figure 12. Comparisons between in situ and predicted $pCO_2$ for (a) onshore and (b) offshore. Corresponding $pCO_2$ values in function of latitudes for (c) onshore and (d) offshore. While offshore is defined as areas by lower than 21°N and higher than 112°E in this study, onshore is defined as remaining areas after offshore regions. All the in situ $pCO_2$ data are shown in Figure 11.
and $\frac{\partial(pCO_2)}{\partial(Chl_a)}$ in equation (3). The relative errors obtained were approximately 4.4 $\mu$atm/°C for onshore regions and 22.8 $\mu$atm/°C for offshore regions for $\frac{\partial(pCO_2)}{\partial(SST)}$; and 25 $\mu$atm/1 mg/m³ for onshore regions and 10 $\mu$atm/1 mg/m³ for offshore regions for $\frac{\partial(pCO_2)}{\partial(Chl_a)}$. Thus, the relative errors were approximately 1.1 $\mu$atm for nearshore regions and 5.7 $\mu$atm for offshore regions due to the accuracy of the SST measurements, and 3.25 $\mu$atm for nearshore regions and 1.3 $\mu$atm for offshore regions due to the accuracy of Chl_a measurements.

[33] Therefore, the uncertainty in $pCO_2$ estimations ($dpCO_2$) using the inputs of the MODIS-SST and $-Chl_a$ measurements for the NN Testing was 4.35 $\mu$atm for onshore regions and 7 $\mu$atm for offshore regions. Accordingly, we assumed that if we used the same sampling periods for remote sensing and in situ measurements, we could estimate $pCO_2$ with an accuracy of 11.25 $\mu$atm (4.35 + 6.9 $\mu$atm) for onshore regions and 13.9 $\mu$atm (7 + 6.9 $\mu$atm) for offshore regions, resulting from the NN Testing algorithm error (6.9 $\mu$atm, Figure 7) and measurement errors for MODIS-SST and $-Chl_a$. The NN Testing algorithm error may include approximately 1 $\mu$atm of uncertainty resulting from the high and low $pCO_2$ values in NN Training as discussed in section 3.1.

4. Concluding Remarks

[34] It remains difficult to reliably assess variations of $pCO_2$ for the global oceans, including marginal seas, due primarily to the lack of sufficient spatial and temporal $pCO_2$ field measurements in these complex regions [Canadell et al., 2003]. Remote sensing with applicable algorithms can certainly be a potentially important approach complementary to ship-board observations. This is the first attempt to apply NN to an extremely dynamic coastal ocean and the algorithm that we developed using NN demonstrated high correlation coefficients between predicted $pCO_2$ and in situ observations. However, it is clearly mandatory to test the derived algorithm generated based on this study for other coastal seas. Nevertheless, based on an examination of the distribution of input data for our NN, we found that if the remote sensing data used to predict $pCO_2$ had similar statistical distributions to the in situ measurements, we could estimate $pCO_2$ with an accuracy of 6.9 $\mu$atm. However, it is worth noting that according to Friedrich and Oschlies [2009a, 2009b], our $pCO_2$ estimations with RMSE of 6.9 $\mu$atm with in situ data may be larger than that for larger areas with remote sensing measurements not only because of not only the uncertainties in remote sensing, but also due to limited in situ data for training $pCO_2$.

[35] Our derived $pCO_2$ map based on the well trained NN algorithm and the monthly mean MODIS-SST, $-Chl_a$, longitudes and latitudes showed general agreement with the in situ measurements, indicating a relatively homogeneous distribution of $pCO_2$ in the offshore area while dynamic changes occurred in the nearshore regions associated with coastal upwelling and river plumes.

[36] In addition to the uncertainty associated with the NN algorithm error, we estimated that the relative errors due to measurement errors in MODIS-SST and $-Chl_a$ were 4.35 $\mu$atm for onshore regions and 7.0 $\mu$atm for offshore regions. Furthermore, we could estimate $pCO_2$ with an accuracy of 12.05 $\mu$atm (accuracy of NN Testing, 7.7 $\mu$atm (Figure 12c) + satellite measurement errors, 4.35 $\mu$atm for nearshore and 13.0 $\mu$atm (accuracy of NN Testing, 4 $\mu$atm (Figure 12d) + satellite measurement errors, 7.0 $\mu$atm for offshore.

[37] The results of this study (and others) show that the NN approach to mapping $pCO_2$ can greatly expand the coverage of the generally sparse in situ observations. Although we only examined the summer sea surface $pCO_2$ predictability based on the NN algorithm, the NN can be potentially applicable to other seasons because the NN does not depend on the time, but the inputs. As long as the inputs are within the same ranges used for the NN training, the NN can predict $pCO_2$ with small uncertainties as discussed above. If the inputs for the NN are out of ranges in other seasons, one cannot derive $pCO_2$ for those regions. Therefore, challenges to a detailed understanding of the regional dynamics of both physics and biogeochemistry remain and more research is needed to validate whether or not our approach is applicable to other seasons or regions.

[38] Acknowledgments. This research was supported by the National Basic Research Development Program (973 Program) of China 2006CB412000 and by National Natural Science Foundation of China through grants 90711005 and 41121091. This research was also supported by NASA through its Physical Oceanography and EPSCoR programs and by the NOAA Sea Grant program. Professor John Hodgkiss is thanked for editorial assistance with English.

References


