Chapter 5

RETRIEVAL OF OCEANIC CONSTITUENTS IN CASE II WATERS

5.1. Introduction

Algorithms for operational retrieval of chlorophyll concentration in Case I waters from satellite ocean colour data are now available. These algorithms often fail in Case II waters. The following facts make the retrieval of oceanic constituents in Case II waters more difficult than in Case I waters:

(1). There are more constituents present in Case II waters. Ocean colour at any wavelength of interest can not be related directly to any one single constituent.

(2). Some of the IOPs of the various constituents which influence ocean colour are similar. For example, the absorption spectra of both CDOM and non-chlorophyllous particles can be modelled by a similar exponential function; the absorption coefficients of phytoplankton and CDOM decrease both from about 440 nm to 550 nm. These phenomena may result in similar ocean colour for different combinations of the constituents concentrations. In these cases, ambiguities may result when retrieving the constituents concentrations from ocean colour.

(3). The compositions of exogenous particulate matter and exogenous CDOM in Case II waters vary strongly with time and region. This means that the IOPs of exogenous particulate matter and exogenous CDOM may greatly vary from time to time and region to region.

(4). There is strong scattering of SPM. If the concentrations of SPM is high, it dominates the water-leaving radiance and may overshadow the contributions of other constituents to the measured signal.

In the following, a method is proposed for the retrieval of the oceanic constituent concentration in Case II waters, based on Artificial Neural Network (ANN) techniques. Input to the presented method is the spectral hemispherical reflectance just below the sea surface. A synthetic data set from radiative transfer simulations is used for the training of an ANN. As mentioned in Chapter 1, a prerequisite for this is that the Inherent Optical Properties (IOPs) of the water constituents required as input to the RT simulations are well representing the conditions to which the trained ANN is later on applied.

As shown in Chapter 4, using a newly developed model of the back scattering probability for marine particles together with other bio-optical models developed from the COASTLOOC data set, the simulated hemispherical reflectance just below the sea surface agrees well with the

corresponding *in-situ* measurements. The IOPs models used herein are therefore deemed to satisfactorily represent the situation encountered in European coastal waters during the COASTLOOC campaigns. This justifies to use the synthetic data set as training data in this study.

5.2. Data Sets

There are two different kind of data used in this study: *in-situ* measurement and RT simulations. In both cases, the data sets relate hemispherical reflectance to the three oceanic parameters: pigment concentration, total suspended particulate matter concentration and the absorption coefficient of CDOM at 443 nm. Based on their individual role in this study, the three different data sets are referred to as:

1. *training data:* a synthetic data set obtained from RT simulations used to train the different ANNs,
2. *validation data:* one part of the *in-situ* measurement data that have been used to develop the model of $\widetilde{b}_b(\lambda)$. This data is used to evaluate the performance of each individual ANN and such to identify the most appropriate one with respect to the retrieval of specific oceanic constituents,
3. *test data* : the other part of the *in-situ* measurement data. This data is used to assess in how far the ANN-based oceanic constituent retrieval scheme is applicable to independent data.

These three data sets are described in more detail in the following.

5.2.1. Training Data

The synthetic data set used to train the ANNs for oceanic constituent retrieval was created using the computer code MOMO [*Fell and Fischer*, 2001]. The IOP models of the oceanic constituents are described in Chapter 4. Based on these models, the IOPs of sea water as required for the RTC can be obtained for given concentrations of oceanic constituents.

It is well known that MLPs, the type of ANN used in this study, has good performance for interpolation, but should not be used for extrapolation. Therefore, it must be made sure that the ranges of the parameters in the training data cover the actual ranges observed in the marine environment of interest. In this study, the ranges of the three components are listed in Table 5.1.

Table 5.1. Ranges of oceanic constituent concentrations

Variable	Concentration Unit	Min	Max
CHL	mg/m^3	0.05	50
SPM	g/m^3	0.05	100
$a_y(443)$	m^{-1}	0.005	1.0

When simulating the radiative transfer in Case II waters, it is often assumed that the concentration of each of the constituents is independent from the other constituents. However, as can be seen from Figure 5.1 displaying COASTLOOC measurements, there is a certain degree of covariance between SPM and CHL, CDOM and CHL, as well as CDOM and SPM also in Case II waters. For a given concentration of one constituent, the other two constituents vary no more than 2 orders of magnitude. This fact may be used to greatly reduce the number of radiative transfer calculations used for the development of Case II algorithms. It allows to identify combinations of the oceanic constituents which are very unlikely to occur in the natural environment and therefore need not to be modelled. Again from the COASTLOOC data, upper and lower boundaries were defined for each constituent concentration in relation to the other constituents (sea also Figure 5.1):

SPM against CHL:

$$spm\_l(chl) = 0.12 \times CHL^{0.77} \tag{5.1}$$

$$spm\_h(chl) = 5.89 \times CHL^{0.77} \tag{5.2}$$

CDOM against CHL

$$cdom\_lc(chl) = 0.0079 \times CHL^{0.75} \tag{5.3}$$

$$cdom\_hc(chl) = 0.25 \times CHL^{0.75} \tag{5.4}$$

CDOM against SPM

$$cdom\_ls(spm) = 0.0063 \times SPM^{0.9} \tag{5.5}$$

$$cdom\_hs(spm) = 0.44 \times SPM^{0.9} \tag{5.6}$$

The combinations of oceanic constituent concentrations used for radiative transfer simulations were selected according to the following steps. Here, a logarithmic distribution of the three oceanic constituent concentrations was applied, so that each order of magnitude is represented with a similar number of cases:

(1). CHL is randomly selected within the range 0.05 and 50 mg /m^3.

(2). For the selected CHL, SPM is randomly selected between the lower and upper boundaries defined by Equations (5.1) and (5.2).

(3). For the selected SPM, CDOM at 443 nm is selected between the lower and upper boundaries defined by Equations (5.3) and (5.4), as well as (5.5) and (5.6).

(4). Steps (1)~(3) are repeated until a sufficient number of combinations of the oceanic constituents has been generated.

The objective of this approach is to potentially reduce the likelihood for ambiguous solutions for the retrieval of the three constituents by reducing the ranges of the constituent distributions.

Based on the above strategy and using the IOP models defined in Chapter 4, simulations of the hemispherical reflectance just below the sea surface were made for:

- 1000 combinations of three oceanic constituents,
- 8 wavelengths: 411, 443, 490, 509, 559, 619, 665, 705 nm,

- 17 solar zenith angles between 0° and 87°.

  Besides, the same assumptions and simplifications were made as specified in Section 3.2.

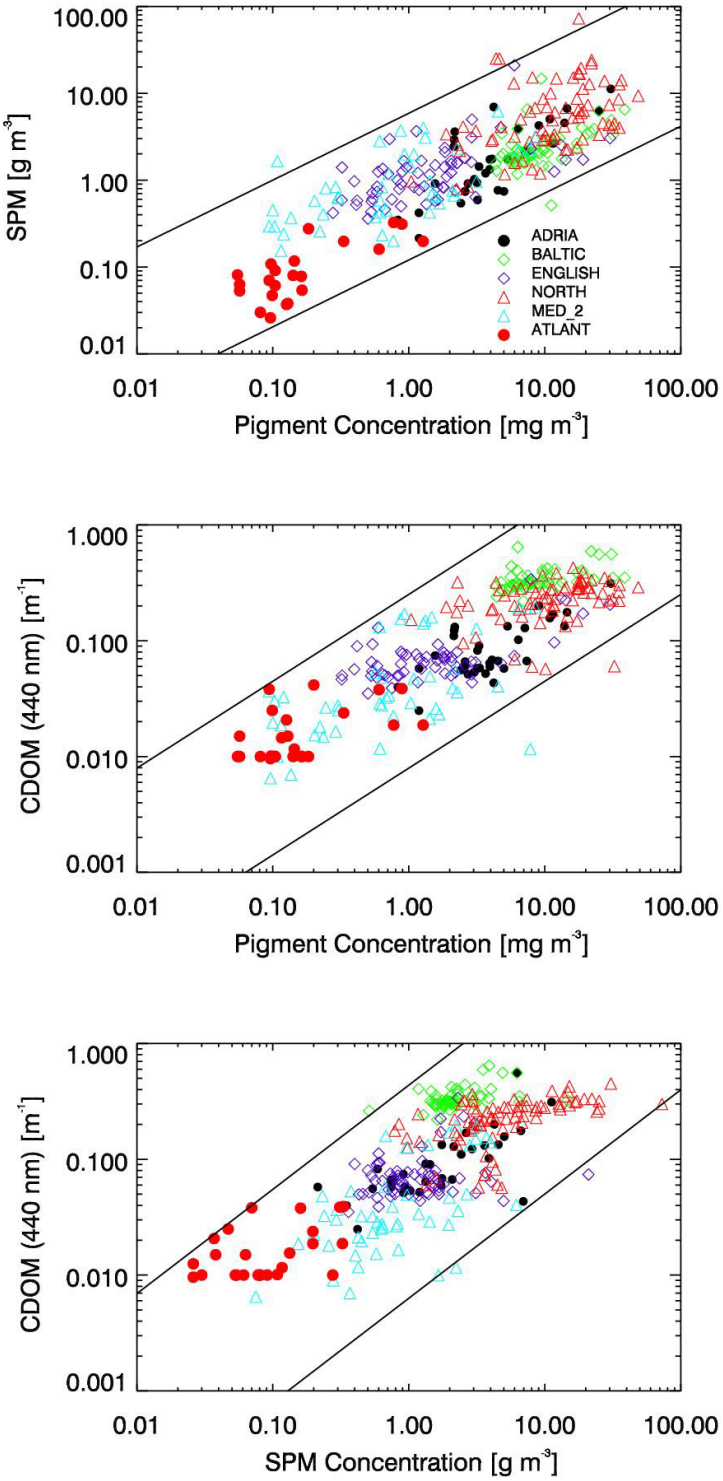

Figure 5.1. Scatter plots of SPM against CHL (top), CDOM absorption against CHL (middle), and CDOM absorption against SPM (bottom). All data set collected during the COASTLOOC cruises.

5.2.2. *In-situ* Measurement Data Sets

The *in-situ* measurement data used in this study come from the following data sets: COASTLOOC (Coastal Surveillance Through Observation of Ocean Colour) [*Babin*, 2000], and PMNS (Particulate Matter North Sea) [*Shimwell et al.,* 1995]. Detailed information on the COASTLOOC data set has been given in Chapter 4.

During the PMNS programme [*Shimwell et al*., 1995], IOPs, AOPs and water quality parameters were measured in the southern North Sea during five cruises. This region is commonly described as the Rhine region of fresh water influence. The waters in this area are characterised as Case II waters.

Table 5.2 provides more information on the *in-situ* measurement data sets chosen for this study. As shown in Table 5.2 (a) and (b), the selected *in-situ* measurement data sets are divided into two parts: *validation data* and *test data*. The *validation data* (which have been used to develop the back scattering model for marine particles in Chapter 4) consist of seven subsets of COASTLOOC data. The *test data* consist of the PMNS data and three subsets of the COASTLOOC data. The radiometric parameter specified in the COASTLOOC and PMNS data sets is the hemispherical reflectance just below the sea surface. All *validation data* set were obtained in Case II waters. The *test data* were partly obtained in CASE I waters, partly in Case II waters. Only chlorophyll concentration is available in the subset COAST_9. For the other subsets, the concentrations of all three oceanic constituents are available.

5.2.3. Data Processing

The *in-situ* data used in this study stem from different sources and differ in a number of details. The following processing procedures were applied in order to generate a consistent data set.

(1). Conversion of reflectance at 532 nm to 509 nm

As shown in Table 5.2, some subsets contain data which are not available at some wavelengths. For subsets COAST_1, COAST_3, COAST_5, COAST_8 and COAST_9 there is no reflectance available at 509 nm, but at 532 nm. Subsets COAST_2, COAST_4, COAST_6 and COAST_7 contain reflectance measurements at both 509 nm and 532 nm. A statistical regression between the reflectances at these two wavelengths yields the following equation:

$$R(509)=-0.00523+0.779[R(532)]^{0.922} \quad (5.7)$$

$$(r^2=0.985, N=180)$$

Hence, R(509) and R(532) are highly correlated. Therefore, Equation (5.7) was applied to derive the reflectances at 509 nm in subsets COAST_1, COAST_3, COAST_5, COAST_8, and COAST_9.

Table 5.2 (a) The characteristics of *in-situ* data sets

	Data set	N	R	Locations	Water type	Available concentration		
						CHL	SPM	$a_y(443)$
Validation data	COAST_1	35	R(0-)	Adriatic Sea	2	yes	yes	yes
	COAST_2	57	R(0-)	Baltic Sea	2	yes	yes	yes
	COAST_3	17	R(0-)	English channel	2	yes	yes	yes
	COAST_4	51	R(0-)	English channel	2	yes	yes	yes
	COAST_5	10	R(0-)	North sea	2	yes	yes	yes
	COAST_6	64	R(0-)	North sea	2	yes	yes	yes
	COAST_7	9	R(0-)	North Sea	2	yes	yes	yes
Test data	COAST_8	48	R(0-)	Mediterranean Sea	2	yes	yes	yes
	COAST_9	34	R(0-)	Mediterranean Sea	1	yes	no	no
	COAST_10	28	R(0-)	Atlantic Ocean	1	yes	yes	yes
	PMNS	131	R(0-)	North Sea	2	yes	yes	yes

Table 5.2 (b) The characteristics of *in-situ* data sets

	Data set	Wavelength of reflectance								
Validation data	COAST_1	411	443	490	---	532	559	619	665	705
	COAST_2	411	443	490	509	532	559	619	665	705
	COAST_3	411	443	490	---	532	559	619	665	705
	COAST_4	411	443	490	509	532	559	619	665	705
	COAST_5	411	443	490	---	532	559	619	665	705
	COAST_6	411	443	490	509	532	559	619	665	705
	COAST_7	411	443	490	509	532	556	---	665	705
Test data	COAST_8	411	443	490	---	532	559	619	665	705
	COAST_9	411	443	490	---	532	559	619	665	705
	COAST_10	411	443	490	509	532	556	---	665	705
	PMNS	412	443	490	513	---	559	622	665	701

(2). Conversion of reflectance at 665 nm to 619 nm

For subsets COAST_7 and COAST_10, the reflectance at 619 nm is not available. A statistical regression between the reflectances at 619 nm and 665 nm in subsets of COAST_1, COAST_2, COAST_3 , COAST_4, COAST_5, COAST_6, COAST_8 and COAST_9 has been obtained, expressed as:

$$\log[R(619)]=0.1106+0.957\log[R(665)] \quad (5.8)$$

(r^2= 0.991, N=315)

R(619) and R(665) are also highly correlated. Thus, Eq. (5.8) was applied to derive the reflectance at 619 nm in subsets COAST_7, COAST_10. Since the correlation between R(619) and R(665) (r^2= 0.963, N=315) is weaker than that between log[R(619)] and log[R(665)], the log scale was here used instead of the linear scale.

(3). Conversion of the absorption coefficient of CDOM at 380 nm to 443 nm

The PMNS data set indicates the absorption coefficient of CDOM at 380 nm instead of 443 nm. The following equation was used to convert the absorption coefficient of CDOM from wavelength 380 nm to 443 nm [*Bricaud et al*., 1981]:

$$a_y(443) = a_y(380)e^{-S_y(\lambda_1-380)}, \quad (5.9)$$

where S_y=0.0176, and λ_1=443 nm.

(4). Conversion of the chlorophyll-a concentration to pigment concentration

The PMNS data set comprises the concentration of chlorophyll-a instead of the pigment concentration. The following relationship was used to convert from chlorophyll-a concentration to pigment concentration [*O'Reilly et al*., 1998]:

$$[pigment] = 1.34 \times [chl\ \mathrm{a}]^{0.983} \quad (5.10)$$

5.3. Retrieval of the Oceanic Constituents with ANN

5.3.1. Artificial Neural Network

In this study, the Multi-Layer-Perceptron (MLP) is used to approximate the relationship between ocean colour and the concentrations of the oceanic constituents in Case II waters. The theory of MLP is described in detail in Section 2.4. It consists of three layers: input layer, one hidden layer and output layer. A bias parameter is added both to the input layer and to the hidden layer.

In this study, three different MLPs are used to retrieve pigment, SPM and CDOM separately, rather than using one single MLP to retrieve the three constituents at one time. Thus, in the output layer, there is one neuron which corresponds to one of the three constituents concentrations. Although more effort is required for the training process, it is of advantage to have one trained MLP for each constituent since it allows to individually optimise constituent retrieval (see Section 5.3.3).

To determine which and how many spectral bands or band ratios are best suited for the retrieval of CHL, SPM and CDOM, 13 combinations of input data (listed in Tables 5.3, 5.4 and 5.5, respectively, for the retrieval of CHL, SPM and CDOM) were tested. Of the 13 listed cases, the first four are combinations of absolute reflectance values at different wavelengths. The other

nine cases are combinations of reflectance ratios. The dimensionality of the input data determines the number of neurons in the input layer.

The optimal number of neurons in the hidden layer depends on various factors. The determination of the number of neurons in the hidden layer is described below.

5.3.2. ANN Training

The performance of each of the algorithms based on the inputs listed in Tables 5.3, 5.4 and 5.5 depends mainly on the following two parameters: neurons in hidden layer and noise level added to the training data set. The same procedures as described in Chapter 3 were used to determine the optimal number of hidden neurons and appropriate noise level. To find the optimal number of hidden neurons, the performance of ANNs with 6, 12, 20, and 30 hidden neurons were tested. To determine the appropriate noise level, 10 %, 20 %, 30 % and 40 % noise was added to the synthetic training data used as input.

A synthetic data set for ANN training has been generated from the RT simulations outlined in Section 5.2. The synthetic data set is composed of 1000 hemispherical reflectance spectra, corresponding to 1000 combinations of three constituent concentration values.

For each of the 13 input combinations listed in Tables 5.3, 5.4 and 5.5, there are five training data sets which correspond to five different noise levels of 0%, 10%, 20%, 30%, 40%, respectively, and four ANNs with 6, 12, 20, or 30 neurons in the hidden layer. These five training data sets were used to train the four different ANNs, thus a total of 5×4 = 20 trained ANNs were obtained, which are the candidates of the corresponding algorithm.

5.3.3. Determining ANN Architecture and Noise Adding for Optimal Oceanic Constituent Retrieval

In order to find the ANN best suited for a specific oceanic constituent retrieval from all the candidate ANNs constructed as described in section 5.3.2, the ANN forecasts were compared to the '*validation data*' by two error measures: root mean square error (RMSE) (defined in Section 3.3) and the square of the Pearson's correlation coefficient r^2. The optimum number of hidden neurons and the appropriate noise level with regard to the above two error measures are also given in Tables 5.3, 5.4 and 5.5 for each of the retrieval algorithms of three oceanic constituents. From the results shown in Tables 5.3, 5.4 and 5.5, the following conclusions can be drawn:

(1). Regarding the retrieval of pigment concentrations, the best results (lowest RMSE and highest r^2) were obtained using the seven reflectance ratios as input (case No. 6 in Table 5.3). In this case, the r^2 value is 0.729, and the RMSE value is 0.274. It is better than that of case No. 1 (RMSE=0.285, and r^2=0.707) which has highest performance of the cases using absolute reflectance value as input.

Table 5.3. Performance with regard to pigment retrieval of spectral band combinations used input to ANNs

Case No.	Input	Neurons in Hidden layer	Noise level (%)	Training data (N=1000)		Validation data (N=205)	
				RMSE	r^2	RMSE	r^2
1	R411, R443, R490, R510, R559, R619, R665, R705	12	30	0.240	0.926	0.285	0.707
2	R411, R443, R490, R510, R559, R619, R665, R705, θS	6	40	0.268	0.905	0.288	0.702
3	R411, R443, R490, R510, R559, R665	6	30	0.280	0.898	0.292	0.705
4	R412, R443, R490, R510, R559, R665, θS	6	30	0.276	0.901	0.321	0.705
5	R443/R411, R490/R411, R510/R411, R559/R411, R619/R411, R665/R411, R705/R411	6	40	0.302	0.879	0.290	0.715
6	R411/R443, R490/R443, R510/R443, R559/R443, R619/R443, R665/R443, R705/R443	12	40	0.290	0.889	0.274	0.729
7	R411/R490, R443/R490, R510/R490, R559/R490, R619/R490, R665/R490, R705/R490	6	40	0.272	0.904	0.292	0.703
8	R411/R559, R443/R559, R490/R559, R510/R559, R619/R559, R665/R559, R705/R559	12	40	0.286	0.891	0.306	0.675
9	R411/R665, R443/R665, R490/R665, R510/R665, R559/R665, R619/R665, R705/R665	6	40	0.285	0.898	0.297	0.712
10	R411/R443, R490/R443, R510/R443, R559/R443, R665/R443	30	40	0.325	0.860	0.280	0.703
11	R411/R559, R443/R559, R490/R559, R510/R559, R665/R559	30	40	0.327	0.858	0.300	0.671
12	R411/R665, R443/R665, R490/R665, R510/R665, R559/R665	6	40	0.319	0.865	0.312	0.678
13	R443/R559, R490/559, R510/R559	6	40	0.396	0.793	0.321	0.605

Table 5.4. Performance with regard to SPM retrieval of spectral band combinations used input to ANNs

Case No.	Input	Neurons in Hidden layer	Noise level (%)	Training data (N=1000)		Validation data (N=218)	
				RMSE	r^2	RMSE	r^2
1	R411, R443, R490, R510, R559, R619, R665, R705	6	0	0.0576	0.994	0.201	0.772
2	R411, R443, R490, R510, R559, R619, R665, R705, θS	6	0	0.0299	0.998	0.212	0.750
3	R411, R443, R490, R510, R559, R665	12	10	0.0726	0.991	0.215	0.756
4	R411, R443, R490, R510, R559, R665, θS	6	0	0.0268	0.999	0.223	0.739
5	R443/R411, R490/R411, R510/R411, R559/R411, R619/R411, R665/R411, R705/R411	6	20	0.239	0.903	0.242	0.654
6	R411/R559, R443/R559, R490/R559, R510/R559, R619/R559, R665/R559, R705/R559	12	20	0.221	0.916	0.244	0.660
7	R411/R665, R443/R665, R490/R665, R510/R665, R559/R665, R619/R665, R705/R665	6	20	0.222	0.916	0.225	0.693
8	R443/R411, R490/R411, R510/R411, R559/R411, R665/R411,	12	20	0.247	0.895	0.232	0.657
9	R411/R559, R443/R559, R490/R559, R510/R510 R665/R559,	6	20	0.233	0.906	0.221	0.693
10	R411/R665, R443/R665, R490/R665, R510/R665, R559/R665	6	20	0.228	0.911	0.228	0.672
11	R510/R665, R559/R665, R619/R665, R705/R665	12	20	0.228	0.911	0.242	0.675
12	R559/R510, R619/R510, R665/R510, R705/R510	12	20	0.230	0.909	0.233	0.666
13	R490/R619, R510/619, R559/R619	12	20	0.245	0.897	0.244	0.630

Table 5.5. Performance with regard to CDOM retrieval of spectral band combinations used input to ANNs

Case No.	Input	Neurons in Hidden layer	Noise level (%)	Training data (N=1000)		Validation data (N=214)	
				RMSE	r^2	RMSE	r^2
1	R411, R443, R490, R510, R559, R619, R665, R705	20	40	0.153	0.941	0.198	0.814
2	R411, R443, R490, R510, R559, R619, R665, R705, θS	30	40	0.154	0.940	0.200	0.818
3	R411, R443, R490, R510, R559, R665	30	40	0.160	0.936	0.183	0.830
4	R412, R443, R490, R510, R559, R665, θS	30	40	0.163	0.933	0.189	0.835
5	R443/R411, R490/R411, R510/R411, R559/R411, R619/R411, R665/R411, R705/R411	20	40	0.227	0.870	0.166	0.787
6	R411/R490, R443/R490, R510/R490, R559/R490, R619/R490, R665/R490, R705/R490	30	40	0.220	0.878	0.189	0.778
7	R411/R559, R443/R559, R490/R559, R510/R559, R619/R559, R665/R559, R705/R559	12	40	0.206	0.893	0.188	0.746
8	R411/R665, R443/R665, R490/R665, R510/R665, R559/R665, R619/R665, R705/R665	12	40	0.218	0.880	0.197	0.784
9	R443/R411, R490/R411, R510/R411, R559/R411, R665/R411,	12	40	0.206	0.892	0.167	0.807
10	R411/R559, R443/R559, R490/R559, R510/R559, R665/R559	6	40	0.217	0.881	0.169	0.786
11	R411/R665, R443/R665, R490/R665, R510/R665, R559/R665	20	40	0.222	0.875	0.201	0.774
12	R443/R411, R490/R411, R510/R411, R559/R411, R619/R411	30	40	0.208	0.891	0.165	0.809
13	R443/R411, R490/R411, R510/R411, R559/R411	6	40	0.225	0.872	0.161	0.790

(2). Regarding the retrieval of SPM concentrations, the performance of the ANNs using absolute reflectance values as input is significantly higher than that of ANNs using the reflectance ratios. The best results (lowest RMSE and highest r^2) were obtained using the eight reflectances as input (case No. 1 in Table 5.4). In this case, the r^2 value is 0.772, and the RMSE value is 0.201.

Besides, the performance of case No. 1 and case No. 2 using eight reflectances as input is slightly better than that of case No. 3 and case No. 4 using six reflectances as input. This means that providing additional information at wavelengths of 619 nm and 709 nm can improve the accuracy of the SPM concentration retrieval.

(3). Regarding the retrieval of CDOM absorption, there is no significant difference of the ANNs performance between using absolute reflectance values and the reflectance ratios as input. The lowest RMSE was obtained in case No. 13 (in Table 5.5) using the four reflectance ratios as input (RMSE=0.161, and r^2=0.790), while the highest r^2 was obtained in case No. 4 using six reflectances plus solar zenith angle as input (RMSE=0.189, and r^2=0.835). By a comprehensive consideration, case No. 4 was taken as the best case for the retrieval of CDOM absorption.

Besides, the performance of case No. 3 and case No. 4 using six reflectances as input is slightly better than that of case No. 1 and case No. 2 using eight reflectances as input. This means that providing additional information at wavelengths of 620 nm and 709 nm does not improve the accuracy of the CDOM absorption retrieval, but in contrast reduces it. The reason for this is that the reflectances at 620 nm and 709 nm are highly correlated with the reflectance at 665 nm, and the information of the reflectance at 665 nm is sufficient for the retrieval of CDOM. Using noisy information at these wavelengths (620 nm and 709 nm) will therefore increase the error of the retrieval CDOM absorption coefficient.

(4). It has been shown in Chapter 3 that the performance of pigment retrieval algorithms based on ratio input are much better than that of algorithms based on absolute reflectance input in Case I waters. The reason is that spectrally correlated noise is partly cancelled out through division of the reflectances at two wavelengths. From the results of this chapter, however, the algorithm best suited for the retrieval of pigment concentration uses the reflectance ratios as input, while the algorithms best suited for the retrieval of SPM and CDOM use the absolute reflectance values as input. The observed behaviour may be explained in the following way. The absorption spectra of CDOM as well as SPM are characterised as an exponential function. Thus, the contribution of CDOM absorption or SPM absorption to the reflectance for different channels acts as a similar way. Therefore, through the division of the reflectances at two wavelengths, on the one hand, the spectrally correlated noise can be partly cancelled out, on the other hand, some useful information for the retrieval of CDOM or SPM may also be partly removed. However, the absorption spectra of the pigment are

characterised by significant difference over the whole spectral domain. Therefore, the contribution of the pigment absorption to the reflectance may not be significantly reduced through division of the reflectances at two wavelengths, while only the spectrally correlated noise can be partly removed.

(5). For each of the retrieval of the three constituents, the noise levels added to training data set are significantly different for optimal retrieval of three oceanic constituents. To sum up, the optimal retrieval of SPM requires little noise added to training data set. However, the optimal retrieval of CHL and CDOM requires more noise added to the training data set. The reason behind the observed behaviour may be as follows. The concentration of SPM is much more sensitive to the variation of reflectance spectra than that of the other two constituents, because of its strong scattering. Therefore, the number of the training cases required for retrieval of SPM is relative small. While pigment concentration and absorption coefficient of CDOM are less sensitive to the variation of reflectance spectra, because of influence of the strong scattering of SPM, as well as their influence from each other. Therefore, the number of the training cases required for the retrieval of pigment and CDOM is relative large. In principle, a sufficiently large set of training cases is necessary to get a good generalisation. If there are no enough number of training cases, a good generalisation can also be obtained by adding an appropriate noise level to the training data set. This technique has been commonly used to reduce the effort of the creation of training data and ANN training process. Generally speaking, the less training cases are available, the more noise is needed. In this study, 1000 training cases for retrievals of all three constituents were used. Therefore, the noise level adding to training data set for the optimal retrieval of pigment and CDOM should be larger than that for the optimal retrieval of SPM.

5.4. Evaluating the Performance of the ANN-based Oceanic Constituents Retrieval Algorithms

5.4.1. Assessing the Performance of the Trained ANN

The potential of the selected optimal ANNs (listed in Table 5.6) for each of three constituents from real measurements is assessed in three steps by applying it a) to the synthetic data ("*training data*") used for the ANN training, b) to "*validation data*" used to determine the ANN architecture and noise level to be added to the input data, and c) to "*test data*" which have not been used for the ANN development.

(1). Performance with respect to the training data

The ANN forecasts of each individual oceanic constituent concentration based on the simulated hemispherical reflectance are compared to the corresponding oceanic constituent

concentrations used as input for the RT simulations. As shown in Table 5.6, as well as in Figures 5.2 (A), 5.3 (A) and 5.4 (A), respectively, for CHL, SPM, and CDOM, the inversion is successful with regard to the synthetic training data set.

Table 5.6. Performance of the selected optimal ANNs for the retrievals of three oceanic constituents

Consti-tuent	Input	Neurons in Hidden layer	Noise level (%)	Training data N=1000		Validation data N=205		Test data N=163	
				RMSE	r^2	RMSE	r^2	RMSE	r^2
CHL	R411/R443, R490/R443, R510/R443, R559/R443, R619/R443, R665/R443, R705/R443	12	40	0.290	0.889	0.274	0.729	0.339	0.860
SPM	R411, R443, R490, R510, R559, R619, R665, R705	6	0	0.0576	0.994	0.201	0.772	0.338	0.910
CDOM	R411, R443, R490, R510, R559, R665, θS	30	40	0.163	0.933	0.189	0.835	0.279	0.769

(2). Performance with respect to the *validation data*

In a second step, the ANNs for the retrieval of the different constituent retrievals are applied to the *'validation data'* consisting of *in-situ* measurements. The retrieval results for pigment, SPM and CDOM are also listed in Table 5.6. The results are depicted in Figures 5.2 (B), 5.3 (B) and 5.4 (B) for CHL, SPM and CDOM, respectively. The concentrations of the three constituents derived from the '*validation data'* agree well with the corresponding *in-situ* measurements. This is not surprising, since the *'validation data'* have a) been used to derive the back scattering model for marine particles used for the RT simulations (see Chapter 4), and b) been selected to identify the most appropriate ANN architecture and noise level.

(3). Performance with respect to the *test data*

In a third step, the ANNs for the retrieval of the different constituents are applied to the second set of *in-situ* measurement data set ('*test data'*). The retrieval results for pigment, SPM and CDOM are also listed in Table 5.6. The results are depicted in Figure 5.2 (C), 5.3 (C) and 5.4 (C) for CHL, SPM and CDOM, respectively. Satisfactory performance is observed even though the '*test data*' are totally independent from the '*validation data*' and have not been used in any respect for the development of the ANN.

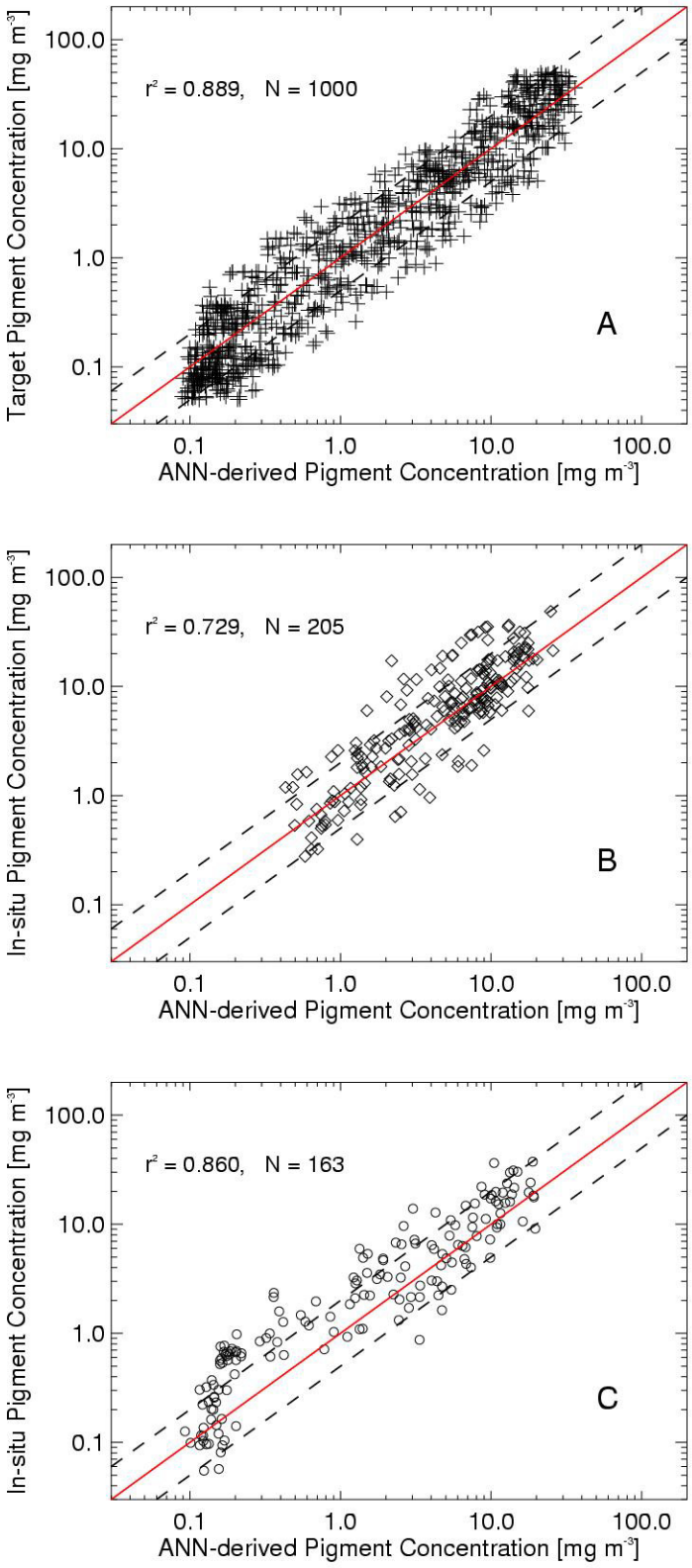

Figure 5.2. Scatter plot showing the performance of the ANN-based pigment retrieval algorithms for the synthetic training data set (A), validation data set (B), and test data (C). The dashed lines indicate the factor 2 error margin.

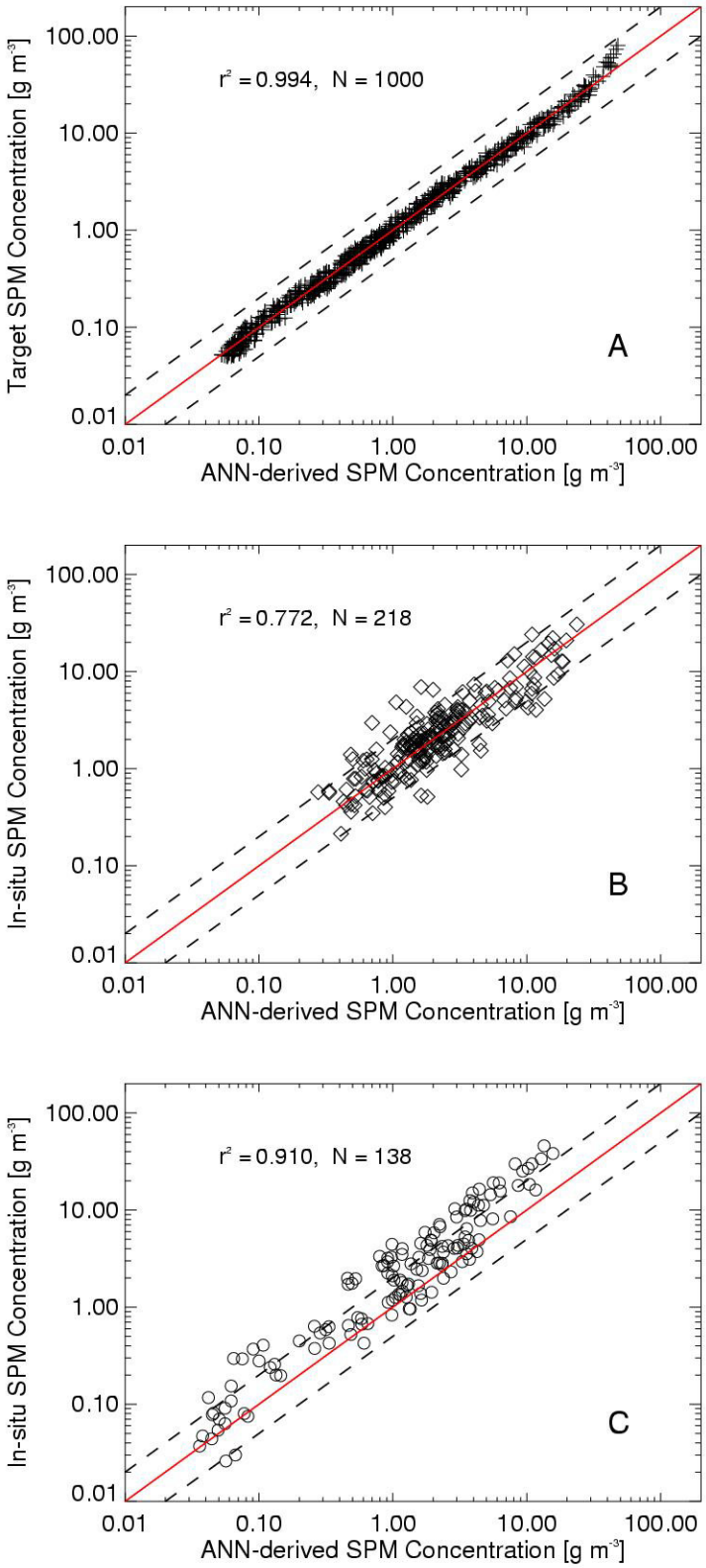

Figure 5.3. Scatter plot showing the performance of the ANN-based SPM retrieval algorithms for the synthetic training data set (A), validation data set (B), and test data (C). The dashed lines indicate the factor 2 error margin.

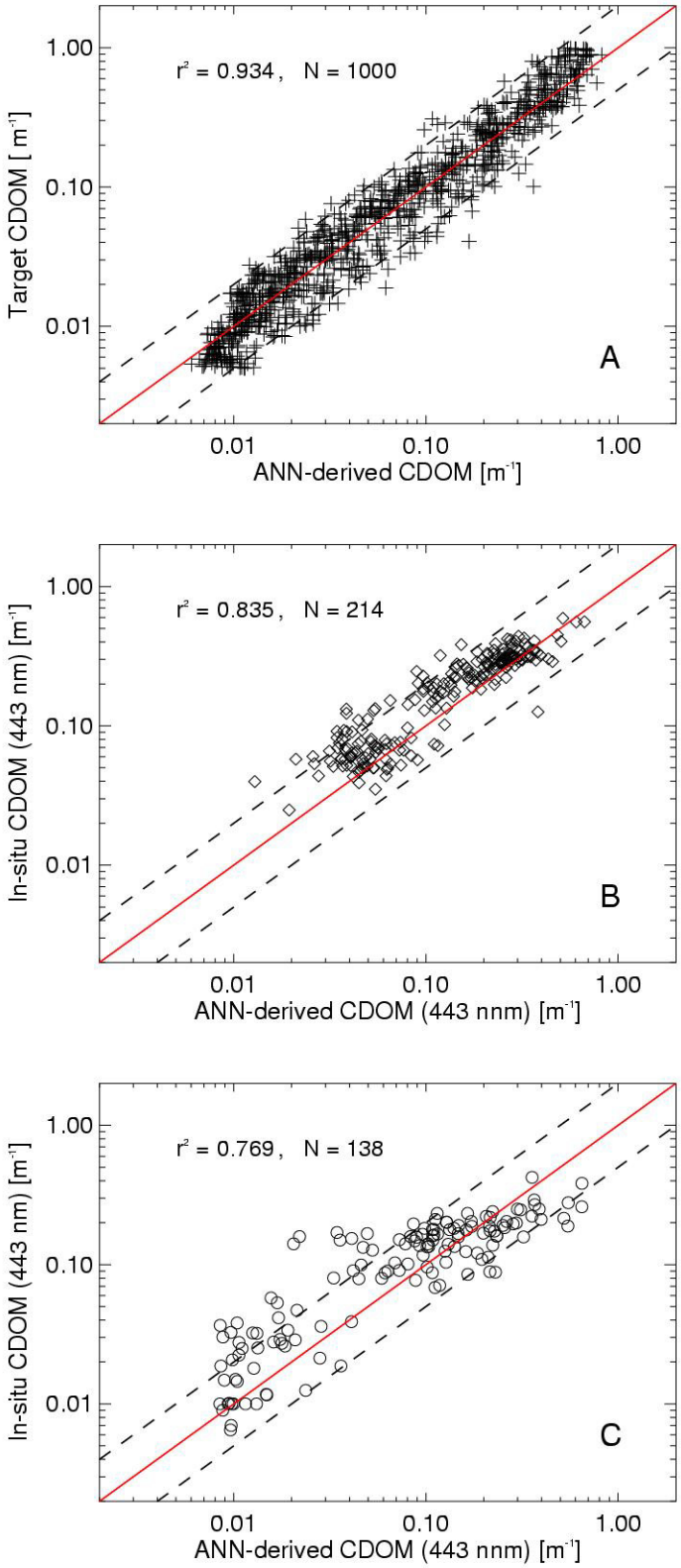

Figure 5.4. Scatter plot showing the performance of the ANN-based CDOM retrieval algorithms for the synthetic training data set (A), validation data set (B), and test data (C). The dashed lines indicate the factor 2 error margin.

5.4.2. Comparison with Existing Retrieval Algorithms

In order to further evaluate the performance of the trained ANNs, it was compared to the PMNS empirical algorithms listed in Table 5.7(a), (b) and (c). The PMNS algorithms [*Shimwell et al.,* 1995] were developed for the retrieval of CHL, SPM and CDOM retrieval, and are based on data collected in the southern North Sea. For each of three constituents, there are two retrieval algorithms based on different band ratios. These algorithms were applied to the reflectance data from the '*validation data*' and '*test data'*. The performance of the algorithms for the three oceanic constituents is also given in Table 5.7(a), (b) and (c). The ANN has the highest r^2 and lowest RMSE of all compared algorithms for the three constituents. The relative success of the ANN-based retrieval is partly explained by the fact that the underlying IOP models represent the *in-situ* measurement data well. Further reason is that it uses more spectral information than do the empirical algorithms PMNS.

Table 5.7 (a). Performance of the ANN-based pigment retrieval algorithms as compared to the PMNS algorithms

Name of algorithm	Algorithm form	Validation data			Test data		
		RMSE	r^2	N	RMSE	r^2	N
PMNS_1	$3.4*[R(510)/R(560)]^{-3.65}$	0.369	0.575	205	0.505	0.753	163
PMNS_2	$22.3*[R(665)/R(705)]^{-2.85}$	0.441	0.351	205	0.713	0.630	163
ANN	ANN	0.274	0.729	205	0.339	0.860	163

Table 5.7 (b). Performance of the ANN-based SPM retrieval algorithms as compared to the PMNS algorithms

Name of algorithm	Algorithm form	Validation data			Test data		
		RMSE	r^2	N	RMSE	r^2	N
PMNS_1	$12.4*[R(412)/R(665)]^{-1.0}$	0.741	0.588	218	0.390	0.855	138
PMNS_2	$53.1*[R(560)/R(620)]^{-2.58}$	0.517	0.681	218	0.433	0.741	138
ANN	ANN	0.201	0.772	218	0.338	0.910	138

Table 5.7 (c). Performance of the ANN-based CDOM retrieval algorithms as compared to the PMNS algorithms

Name of algorithm	Algorithm form	Validation data			Test data		
		RMSE	r^2	N	RMSE	r^2	N
PMNS_1	$0.76*[R(490)/R(665)]^{-0.83}$	0.396	0.694	214	0.450	0.764	138
PMNS_2	$0.72*[R(520)/R(620)]^{-1.05}$	0.407	0.714	214	0.490	0.700	138
ANN	ANN	0.163	0.835	214	0.279	0.769	138

5.5. Conclusions

In this study, a methodology for the retrieval of three constituents from ocean colour in Case II waters have been derived. The retrieval method is derived by applying ANN techniques to a set of hemispherical reflectance spectra typical of Case II waters, which have been obtained from RT simulations.

Three ANN-based algorithms were obtained in this study for the retrievals of CHL, SPM and CDOM, respectively. Each individual ANN has three layers: one input layer, one hidden layer and one output layer. A bias parameter is added both to the input layer and to the hidden layer. The output layer consists of one neuron which corresponds to one of the three constituent concentrations. The number of neurons in the input layer and in the hidden layer which was determined in terms of the optimal retrieval of constituents is different for these algorithms. For the optimal retrieval of pigment, the ANN has seven neurons in the input layer corresponding to the seven reflectance ratios, and 12 hidden neurons. For the optimal retrieval of SPM, the ANN has eight neurons in the input layer corresponding to the eight reflectances, and 6 hidden neurons. For the optimal retrieval of CDOM, the ANN has seven neurons corresponding to six reflectances and solar zenith angle, and 30 hidden neurons.

Applying the three trained ANNs either to the *validation data*, or to the *test data* which have not been used to derive the back scattering probability model of marine particles used for the RT simulations, the results for the retrieval of all three constituents are satisfactory. For example, for retrieval of pigment concentration, applying the algorithm to the *validation data* gives a correlation between predicted and measured pigment concentrations of $r^2 = 0.729$ and RMSE = 0.274; applying it to the *test data*, results in $r^2 = 0.860$ and RMSE = 0.339. For the retrieval of SPM concentration, applying the algorithm to the *validation data* gives a correlation between predicted and measured pigment concentrations of $r^2 = 0.772$ and RMSE = 0.201; applying it to the *test data*, results in $r^2 = 0.910$ and RMSE = 0.338. For the retrieval of CDOM, applying the algorithm to the *validation data* gives a correlation between predicted and measured pigment concentrations of $r^2 = 0.835$ and RMSE = 0.189; applying it to the *test data*, results in $r^2 = 0.769$ and RMSE = 0.279.

The performance of the ANN-based retrieval scheme is generally better than that of the empirical algorithms PMNS. For example, for the retrieval of pigment concentration, applying the PMNS algorithm PMNS_1 to the *test data* set gives $r^2 = 0.753$, and RMSE = 0.505 as compared to $r^2 = 0.860$ and RMSE = 0.263 for the ANN-based algorithm. For the retrieval of SPM concentration, applying the PMNS algorithm PMNS_1 to the *test data* set gives $r^2 = 0.855$, and RMSE = 0.390 as compared to $r^2 = 0.910$ and RMSE = 0.338 for the ANN-based algorithm. For the retrieval of CDOM, applying the PMNS algorithm PMNS_1 to the *test data* set gives $r^2 = 0.764$, and RMSE = 0.450 as compared to $r^2 = 0.769$ and RMSE = 0.262 for the ANN-based algorithm.