UNIVERSITY OF SOUTHAMPTON

Multi Layer Perceptron Neural Network Algorithms for Ocean Colour Applications in Coastal Waters

by

Davide D'Alimonte

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy

School Of Ocean and Earth Science
Southampton Oceanography Centre
United Kingdom

April 2004

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SCIENCE
OCEANOGRAPHY

Doctor of Philosophy

Multi Layer Perceptron Neural Network Algorithms for Ocean Colour Applications in Coastal Waters

by Davide D'Alimonte

This work focuses on the development, performance assessment and identification of the range of applicability of ocean colour algorithms for the retrieval of seawater constituents in optically complex coastal regions. Multi Layer Perceptron neural network algorithms were implemented using a set of experimental data collected in the northern Adriatic Sea (CoASTS data set). The following quantities were modelled as a function of the remote sensing reflectance: i) *Chlorophyll-a* concentration (*Chl-a*); ii) Absorption of the pigmented particulate matter at 443 nm (a_{ph} (443)); iii) Absorption spectra of the coloured dissolved organic matter (a_{CDOM}) and of the non pigmented particulate matter (a_{NPPM}). The range of applicability of two MLP algorithms developed for the retrieval of the *Chl-a* in coastal and open sea regions, were compared using SeaWiFS images of European Seas. A method for algorithms blending was also formulated and applied to case studies.

Contents

Symbols viii

Mathematical Symbols x

Glossary xi

Acknowledgements xii

1 Introduction 1

2 Theoretical background 7

2.1 The light field 7

2.1.1 Coordinate systems 7

2.1.2 The solid angle 9

2.1.3 Radiant energy 10

2.1.4 Radiant power and intensity 10

2.1.5 Irradiance 11

2.1.6 Scalar irradiance 12

2.1.7 Radiance 13

2.2 Inherent and Apparent Optical Properties 14

2.2.1 Inherent Optical Properties 14

2.2.1.1 Absorbance, scatterance and emittance 14

2.2.1.2 Absorption, attenuation and beam attenuation coefficients 16

2.2.1.3 Single scattering albedo 16

2.2.1.4 Volume scattering function 17

2.2.2 Inherent Optical Properties of seawater and its components . 18

2.2.2.1 Absorption by pure seawater and its components . 18

Absorption by pure seawater 18

Absorption by phytoplankton 19

Absorption by coloured dissolved organic matter . . . 20

Absorption by non non pigmented particulate matter 21

2.2.2.2 Back-scattering by pure seawater and its components 22

Back-scattering by pure seawater 22

Back-scattering by phytoplankton 22

Back-scattering by non pigmented particulate matter 23

2.2.3 Apparent Optical Properties 23
2.2.3.1 Irradiance reflectance 24
2.2.3.2 Remote Sensing Reflectance 24
2.2.3.3 Diffuse attenuation coefficient 25
2.3 Empirical and semi-analytical ocean colour algorithms 26
2.3.1 Band-ratio empirical algorithms 26
2.3.2 Semi-analytical algorithms 28

3 The CoASTS data **30**
3.1 CoASTS 30
3.2 The CoASTS data processing system 31
3.2.1 Data calibration and formatting toolbox 32
3.2.2 Data filtering and processing toolbox 34
3.2.2.1 Definition of the extrapolation intervals 35
3.2.2.2 Correction with respect to the above water irradiance 36
3.2.2.3 Data filtering 36
3.2.2.4 Correction for perturbation effects 37
3.2.2.5 Sub-surface values 38
3.2.2.6 Products 39
3.3 The CoASTS data 39
3.3.1 Remote sensing reflectance 41
3.3.2 Chlorophyll-*a* concentration 42
3.3.3 Pigmented and non pigmented particle absorption coefficients 43
3.3.4 Coloured dissolved organic matter absorption coefficient . . 45

4 Multi Layer Perceptron and Novelty Detection **47**
4.1 The Multi Layer Perceptron (MLP) neural network 47
4.1.1 The MLP structure 48
4.1.2 The error function 50
4.1.3 Interpretation of the network outputs 52
4.1.4 Optimization algorithms 54
4.1.5 MLP generalization capability and performance assessment . 56
4.1.5.1 Model selection 56
4.1.5.2 Weigh decay factor 59
4.1.5.3 Cross-validation 60
4.2 The *novelty detection* 61

5 Retrieval of the Optically Significant Seawater Components **63**
5.1 Phytoplankton determination in an optically complex coastal region using a Multi Layer Perceptron 65
5.2 Determination of CDOM and NPPM Absorption Coefficient Spectra from Coastal Water Remote Sensing Reflectances 85

6 Range of applicability of the empirical ocean colour algorithms 105

6.1 Use of the novelty detection technique to identify the range of applicability of empirical ocean colour algorithms 107

7 Summary and conclusions **130**

7.1 Summary of the major results 130

7.1.1 Modelling of the phytoplankton component 131

7.1.2 Coloured dissolved organic matter and non pigmented particulate matter absorption coefficiens 131

7.1.3 Algorithms applicability 132

7.2 Accomplishment of the study objectives 134

7.3 Future work 135

7.3.1 The empirical versus the analytical approach 135

7.3.2 Algorithms assessment with data from other oceanographic areas 136

7.3.3 Reduction of the dimensionality of the input space 137

7.3.4 Modelling of the algorithms accuracy 138

7.3.5 Modelling of the noise affecting input data 138

7.3.6 Range of applicability and algorithm blending 139

References **140**

List of Figures

1.1 Schematic representation of the influence of the optically active sea-water components on the radiance leaving the sea surface 2

2.1 Coordinate systems 8
2.2 The solid angle 9
2.3 Light spectrum 11
2.4 Scalar irradiance sensor 12
2.5 Radiance sensor 13
2.6 Light-medium interaction 15
2.7 Spectral plot of the pure water absorption coefficient 19
2.8 Spectral plot of the *chlorophyll-a* specific absorption coefficient . . . 20

3.1 The Acqua Alta Oceanographic Tower 31
3.2 Software toolbox for the data calibration and formatting 33
3.3 Software toolbox for the processing of optical profile data 34
3.4 Example of optical profile data visualization 35
3.5 The WiSPER system and the above water sensor for irradiance measurements 40
3.6 Spectral plot of the *in situ* remote sensing reflectance 41
3.7 Distribution of the *chlorophyll-a* concentration within CoASTS data 43
3.8 Spectral plot of the *in situ* pigmented particulate matter absorption coefficient 43
3.9 Spectral plot of the *in situ* non pigmented particulate matter absorption coefficient 44
3.10 Spectral plot of the *in situ* coloured dissolved organic matter absorption coefficient 45

4.1 Example of Multi Layer Perceptron neural network 48
4.2 The hyperbolic tangent activation function 49
4.3 Interpretation of the networks outputs 54
4.4 Schematic representation of the *curse of dimensionality* 57
4.5 Schematic illustration of the cross-validation. 60

5.1 CoASTS data set partitioning in Case 1 and Case 2 waters 69
5.2 Schematic representation of data partitioning for cross-validation . . 72
5.3 Dependence of the MLP performance on the algorithm architecture for different sets of input variables 74

5.4 Schematic representation of the MLP adopted for the retrieval of *Chl-a* and $a_{ph}(443)$ 76
5.5 Computed versus measured *chlorophyll-a* concentration 81
5.6 Computed versus measured pigmented particulate matter absorption at 443 nm 83
5.7 Sample absorption spectra of coloured dissolved organic matter and of non pigmented particulate matter and corresponding exponential regressions 91
5.8 Scatter plot of computed regressional parameters versus the corresponding values derived from the interpolation of experimental data 96

6.1 Distribution of the *Chl-a* for the SeaWiFS and the CoASTS data sets 111
6.2 Distribution of the R_{rs} at 490 nm for the SeaWiFS and the CoASTS data sets 112
6.3 Schematic illustration of the partitioning of the data set for use in cross-validation 116
6.4 Scatter plot of computed versus *in situ* measured *chlorophyll-a* concentration for the SeaBAM and the CoASTS algorithms 117
6.5 Three-dimensional representation of the range of applicability of the SeaBAM and CoASTS algorithms 119
6.6 Central Mediterranean Sea. Range of applicability of the SeaBAM and CoASTS algorithm 123
6.7 North Sea. Range of applicability of the SeaBAM and CoASTS algorithm 125
6.8 Region of applicability of the SeaBAM and CoASTS algorithms for the Baltic Sea 127

List of Tables

3.1 Variability of the *in situ* measured remote sensing reflectance 42
3.2 Variability of the *in situ* measured pigmented particulate matter absorption coefficients 44
3.3 Variability of the *in situ* measured non pigmented particulate matter absorption coefficients 45
3.4 Variability of the *in situ* measured coloured dissolved organic matter absorption coefficients 46

5.1 Summary of the algorithms accuracy in retrieving the *chlorophyll-a* concentration and the absorption of the pigmented particulate matter at 443 nm 79
5.2 Summary of the accuracy in retrieving the coloured dissolved organic matter absorption coefficients 97
5.3 Summary of the accuracy in retrieving the non pigmented particulate matter absorption coefficients 97
5.4 Summary of the accuracy in retrieving the *slope* regressional coefficients 99

Symbols

θ	Zenith angle
ϕ	Azimuth angle
Ω	Solid angle
λ	Wavelength
$\mathbf{E}$	Electric field
$\mathbf{B}$	Magnetic field
$\mathbf{S}$	Poynting vector
U	Radiant energy
Φ	Radiant power
I	Radiant intensity
E	Irradiance
E_u	Upward irradiance
E_d	Downward irradiance
E_s	Above water downward irradiance
E_i	Diffuse sky irradiance
E_0	Scalar irradiance
L	Radiance
L_u	Upward radiance
L_w	water-leaving radiance
A	Absorbance
B	Scatterance
T	Transmittance
a	Absorption coefficient
b	Attenuation coefficient
c	Beam attenuation coefficient
ω_0	Single scattering albedo
β	Volume scattering function
b_b	Back-scattering coefficient
a_{ph}	Pigmented particulate matter absorption

a_{NPPM}	Non pigmented particulate matter absorption
a_{CDOM}	Coloured dissolved organic matter absorption
R	Irradiance reflectance
R_{rs}	Remote sensing reflectance
f_0	Extra-atmospheric sun irradiance
L_{wn}	Normalized water-leaving radiance
k_d	Diffuse attenuation coefficient

Mathematical Symbols

$\equiv$	is defined as
$\sim$	is on the order of
$g(\cdot)$	Activation function
$\mathcal{L}$	Likelihood function
$\mathcal{E}$	Error function
Δx	The change in x
$\sum_{n=0}^{N} x_n$	Sum of all quantities x_n from $n = 0$ to $n = N$
$\prod_{n=0}^{N} x_n$	Product of all quantities x_n from $n = 0$ to $n = N$
$N \to \infty$	N approaches to infinite
$\frac{df}{dx}$	Derivative of f with respect to x
$\frac{\partial f}{\partial x}$	Partial derivative of f with respect to x
$\int_{x_0}^{x} f(x')dx'$	Integral of f with respect to x' between x_0 and x
$\mathbf{i}$	The vector $\mathbf{i}$
$\mathbf{i} \cdot \mathbf{j}$	Internal product of $\mathbf{i}$ and $\mathbf{j}$
$\mathbf{i} \times \mathbf{j}$	External product of $\mathbf{i}$ and $\mathbf{j}$
pdf	Probability density function
$p(x)$	Density probability of x
$p(t, x)$	Joint density probability of t and x
$p(t\|x)$	Conditioned density probability of t on x
$\langle t\|x \rangle$	Conditional average of t on x
$P(x)$	Probability of x

Glossary

AAOT	Acqua Alta Oceanographic Tower
AOP	Apparent Optical Properties
ARD	Automatic Relevance Determination
CDOM	Coloured Dissolved Organic Matter
CoASTS	Coastal Atmosphere and Sea Time Series
COLORS	COastal region LOng-term measurements for colour Remote Sensing
COASTlOOK	COAstal Surveillance Through Observation of Ocean Colour
GMM	Guassian Mixture Model
GUI	Graphical User Interface
HPLC	High Performance Liquid Chromatography
ICA	Independent Component Analisys
IOP	Inherent Optical Properties
MBR	Maximum Band Ratio
ME	Mixture of Experts
MLP	Multi Layer Perceptron
MVD	Multichannel Visible Detector System
NN	Neural Network
NPPM	Non Pigmented Particulate Matter
SeaBAM	SeaWiFS Bio-Optical Algorithm Mini-Workshop
SCG	Scaled Conjugate Gradient
SeaPRISM	SeaWiFS Photometer Revision for Incident Surface Measurements
SeaWiFS	Sea-viewing Wide Field-of-view Sensors
TSM	Total Suspended Matter
WiSPER	Wire Stabilized Profiling Environmental Radiometer

Acknowledgements

I wish to express my gratitude to all the people who have helped me in carrying on this work. The first is Dr. Giuseppe Zibordi for the inestimable scientific support and suggestions. I'm particularly grateful to my advisors of the University of Southampton, Dr. Simon Boxall, Prof. Ian Robinson and Dr. Arian Tatnall. A sincere thank to Jean Françoise Berthon, Frédéric Mélin, Nicolas Hoepffner, Andrea Baraldi and to all the people I had the pleasure to meet at the JRC. Finally, a thank also to Stan Hooker for his encouraging "Davide, keep going the effort!".

To Ferdi

Chapter 1

Introduction

Seawater constituents affect the light back-scattered from the ocean according to their type and concentration. This is the basis for ocean colour application, which aims at retrieving optically significant constituents of seawater from water-leaving radiance. The use of sensors located on a remote platform (aircraft or satellite) and measuring the radiance leaving the sea (Figure 1.1) is a unique way to collect information on a large scale over the Earth's surface.

Optically significant water components are conventionally grouped, according to their different optical characters and their specific bio-geo-chemical origin, into: i) pigmented (biological) matter or phytoplankton, ii) non pigmented particulate matter or detritus and iii) coloured dissolved organic matter or yellow substance. The pigmented matter is the basis of the marine food web. Moreover, the fact that the ocean carbon fixation depends on the phytoplankton component underlines the importance of the estimation of this quantity. Several processes can lead to the detritus. Micro-organism shells, such as those of Coccolitophore, are an example of the non pigmented matter in open ocean. In coastal waters, either waves or tidal action can raise particles from the sea bottom. Transport of particles by rivers and wind represents another important source. The coloured dissolved organic matter can be generated either by the degradation of the phytoplankton or other organic substances. Its origin can be marine or terrestrial, via

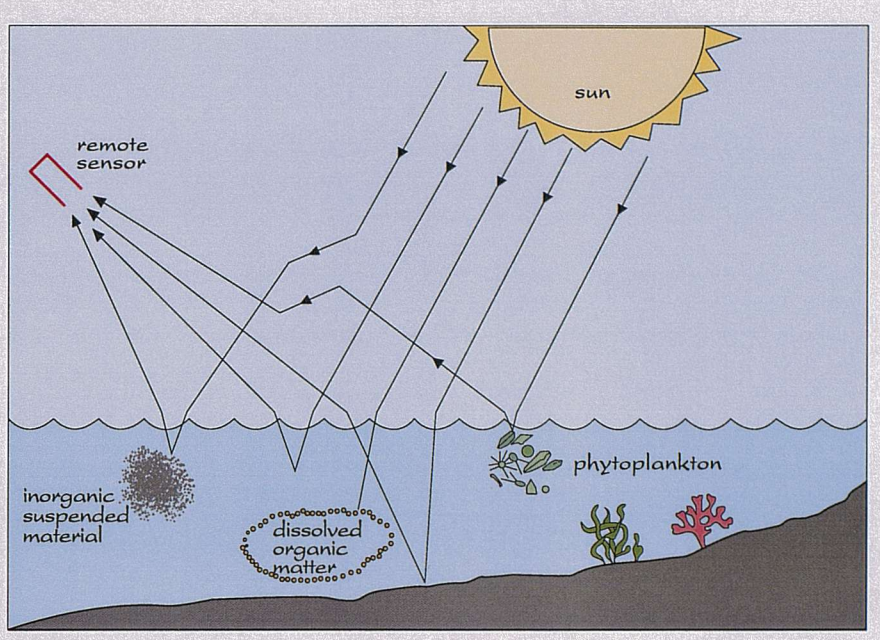


FIGURE 1.1: Schematic representation of the influence of the optically active seawater components on the radiance leaving the sea surface. Adapted from Sathyendranath (2000).

rivers, and in the latter case the coloured dissolved organic matter can represent an important tracer of the dynamics in the river deltas.

Seawater can be separated into two major classes as a function of the relationships linking the optically active constituents (Morel and Prieur, 1977; Gordon et al., 1983). Case 1 waters, mainly present in the open ocean, are those where phytoplankton and its derivatives are the only water components responsible for the modification of the light field. More in general, Case 1 waters are represented by those conditions where phytoplankton and other water constituents, whereas produced from different sources, present a high level of covariance. Differently, in Case 2 waters the separate optically active constituents can vary more independently. Typical Case 2 waters are found in coastal zones and estuaries, where local sources, such as river runoff or meteorological events, can bring substances from the inland, strongly modifying the covariance between components typical of Case 1 waters.

The separation of water types into Case 1 and Case 2 is of particular importance in the ocean colour problem. In fact, in Case 1 waters, relatively simple empirical relationships (O'Reilly et al., 1998) can be used for the retrieval of water constituents from the water-

leaving radiance. In contrast, algorithms for Case 2 waters have to handle an increased level of complexity. This derives from the more independent relative variability of the optically active components and the spectral combination of their specific influence on the water leaving radiance. Whereas a large portion of seawater is Case 1 water, the effort in improving the bio-optical modelling in Case 2 waters is motivated by the relevance that coastal areas have on various human activities.

As stated by Sathyendranath (2000), "The geographic distribution of Case 2 waters is variable: waters of a given locality may drift between Case 1 and Case 2 conditions, depending on environmental forcing. **At the most elementary level, we need an algorithm to distinguish between Case 1 and Case 2 waters. This would allow masking of regions where Case 1 algorithms may not hold, and classifying these waters as belonging to more complex Case 2. At a more advanced level, we need algorithms that can derive more information about Case 2 waters**". The former highlighted elements have become the objectives of the present work.

This study makes use of experimental data collected in an optically complex coastal region. Here, *coastal region* is used to indicate those waters which are permanently or occasionally Case 2 waters as result of bottom resuspension, coastal erosion, river inputs, or by simple anthropogenic impact. The author participated in field campaigns to acquire these data and was strongly involved in the development of the software for their analysis. The *in situ* data set is used to develop and to assess the performance of empirical ocean colour algorithms.

Ocean colour algorithms can be grouped in two main categories: *analytical* and *empirical*. Analytical algorithms retrieve optically active seawater components by inverting a forward theoretical model of the radiative transfer process in the water medium. Besides the validity of the forward theoretical model, the effectiveness of this approach depends on i) any simplification adopted to make the forward model computationally more tractable and ii) any parametrizations based on experimental data (and in such a

case the model is also called *semi-analytical*) adopted to account for properties that depend on specific environmental conditions. By contrast, empirical algorithms only rely on experimental data to represent the statistical properties of the data generating process without embodying any physical representation. Whereas the analytical approach is the ideal solution, it requires more steps to be implemented (i.e., the identification of the forward model, its parameters and the method to solve the inverse problem) and this can affect the final accuracy in retrieving optically significant seawater components. On the other hand, advances of the recent years in the pattern recognition research field have shown that the Neural Network (NN) are particularly effective among the empirical algorithms, despite their weakness in providing an understanding of the underlying physical process, in representing complex non-linear relationship between multidimensional input and output space.

The aim of the present study is to evaluate the applicability of NN techniques to the ocean colour research field. Multi Layer Perceptron (MLP) NN algorithms are developed in the present study to retrieve i) the *chlorophyll-a* concentration (*Chl-a*), i) the absorption of the pigmented particulate matter at 443 nm a_{ph} (443)) and iii) the spectra of the absorption coefficients of the coloured dissolved organic matter (CDOM) and of the non pigmented particulate matter (NPPM) from remote sensing reflectance (R_{rs}). The accuracy of the algorithms is assessed using both *in situ* remote sensing reflectances as well as coincident SeaWiFS data. The study is restricted to the problem of the algorithm development and without investigating the critical aspect of the atmospheric correction in coastal regions (the SeaWiFS data used in this study were processed with the scheme presented by Sturm and Zibordi (2002) and assessed by Bulgarelli and Zibordi (2002) and Mélín et al. (2003)).

Several authors (Doerffer and Fischer, 1994; Shiller and Doerffer, 1997; Buckton et al., 1999; Gross et al., 2000; Cipollini et al., 2001; Dransfeld, 2003) have already attempted to use MLP or other neural network algorithms to retrieve seawater bio-optical components. However, they restricted the evaluation of the algorithms performance to simulated

data. Thus, an important insight and advance from the present work is to verify the effectiveness of the use of a limited set of experimental data for the MLP implementation. In order to better exploit the *in situ* data, specific techniques are applied both to optimize the MLP performance and its generalization capability (*complexity optimization*), as well as to evaluate its performance (*cross-validation*).

The second objective of the study is the assessment of the range of applicability of the algorithms. This is an important issue, because algorithms for Case 2 are likely to have only a regional validity that can not be simply defined on a geographical basis (due to the dynamics of seawater type and distribution). The technique here proposed for the definition of the range of applicability of the algorithms, is based on the statistical representativeness of the data used for the algorithm development, determined through the *novelty detection* method (Bishop, 1994). This method assumes that the algorithm is likely to produce reliable results if applied to data well represented in the training data set. On the contrary, the algorithm products obtained from data poorly represented in the training data set (i.e., *novel data*) are likely to be less accurate.

The study on the algorithm applicability makes use of a second data set representative of open ocean conditions, in addition to the coastal data set. The Case 1 water data set is exploited to implement a new algorithm for *chlorophyll-a* retrieval, whose applicability is compared with that of the coastal water algorithm. The range of applicability of the two algorithms is analyzed on the basis of SeaWiFS remote sensing data representative of different oceanographic regions (i.e., the Central Mediterranean Sea, the North Sea and the Baltic Sea). To complete the study, a new approach is proposed for blending products deriving from different algorithms.

This work includes:

- Quantities describing the light field and the seawater optical properties, as well as the principles of the empirical and analytical algorithms, in Chapter 2.

- Details on the coastal data set and on the processing scheme for the evaluation of R_{rs} from optical profile data, in Chapter 3.

- Fundamentals on the MLP algorithms and the novelty detection technique, in Chapter 4.

- Details on the implementation and performance assessment of the proposed MLP algorithms for the retrieval of the optically active seawater components, in Chapter 5. The results of this analysis are presented through two papers submitted to *IEEE Transactions in Geoscience and Remote Sensing.* The first work presents a new methodology for the retrieval of the absorption coefficient spectra of the coloured dissolved organic matter and non pigmented particulate matter from remote sensing reflectances. The second work (accepted for publication) investigates the accuracy of various empirical algorithms in retrieving both the *chlorophyll-a* concentration as well as the absorption of the non pigmented particulate matter at 443 nm.

- The analysis of the range of applicability of the algorithms for the *Chl-a* retrieval in coastal and open ocean conditions for different European Seas, in Chapter 6. The results of the analysis are presented through a paper, submitted and accepted for publication by *IEEE Transactions in Geoscience and Remote Sensing.*

- Conclusion and summary, in Chapter 7.

Chapter 2

Theoretical background

Light, interacting with the the optically active seawater constituents is modified due to absorption and scattering processes. This chapter presents the quantities describing the light field (Section 2.1) and the optical properties of the seawater (Section 2.2). The band-ratio empirical approach and the semi-analytical approach are here also presented (Section 2.3).

2.1 The light field

This section presents the fundamental radiometric properties as well as the principles of sensor technology.

2.1.1 Coordinate systems

The *orthogonal coordinate system* (Figure 2.1(a)) will be hereafter used to define the quantities describing the light field. The origin O of the coordinate system is a reference point at the sea surface. The directions of the unit basis vectors are chosen according to specific needs. The unit vector $\mathbf{i}$, corresponding to the X-axis, belongs to the sea-surface plane. Its direction can be determined by the intersection of the plane containing the light field source and the horizontal plane (Kirk, 1994), or by the wind flowing direction

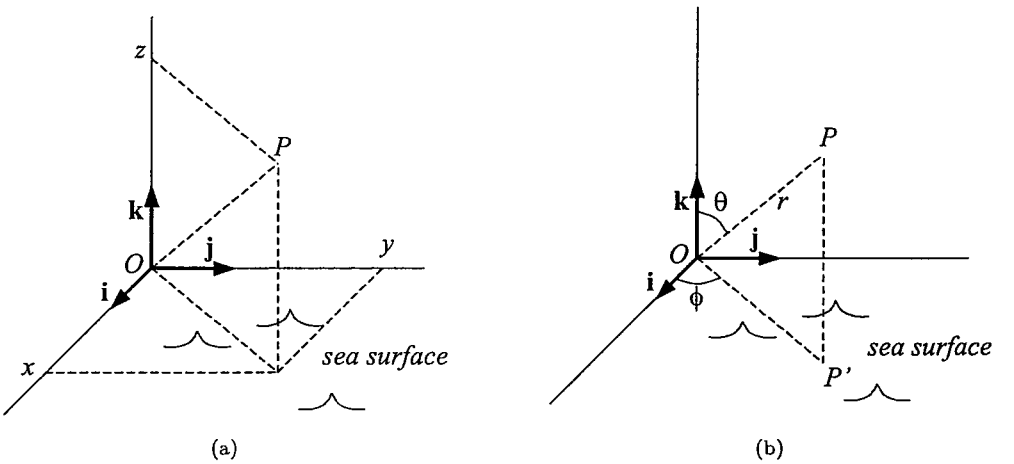


FIGURE 2.1: Schematic representation of a) the orthogonal and b) the polar coordinate system.

(Mobley, 1994). The unit vector $\mathbf{k}$, corresponding to the Z-axis, can be oriented towards the sky zenith (as in the present description) or in the opposite direction (nadir). The Y-axis, identified by $\mathbf{y}$, is defined by the cross product $\mathbf{k} \times \mathbf{i}$. A generic point P is then identified by the position vector $\mathbf{p} = x \cdot \mathbf{i} + y \cdot \mathbf{j} + z \cdot \mathbf{k}$ (being x, y and z the projection of P on the basis vectors). Some times there is the need to distinguish vector direction from vector position. The former will be hereafter identified by the versor

$$\xi = \frac{x \cdot \mathbf{i} + y \cdot \mathbf{j} + z \cdot \mathbf{k}}{\sqrt{x^2 + y^2 + z^2}}. \tag{2.1}$$

Thus, ξ has then the same direction of the vector $\mathbf{p}$ but unit length.

In some cases, the properties of the light field can be efficiently described through the *polar coordinates system* (Figure 2.1(b)). A generic point P is identified by its distance r from origin O, and by two angles, the *zenith* θ and the *azimuth* ϕ. The zenith angle corresponds to the angle between the OP and $\mathbf{k}$. The azimuth is the angle between the OP' and $\mathbf{i}$, being P' the projection of the point P onto the plane identified by $\mathbf{i}$ and $\mathbf{j}$.

Orthogonal coordinates are expressed as a function of the polar coordinates by the following relationship:

$$x = r \cdot \sin\theta \cdot \cos(\phi) \tag{2.2a}$$

$$y = r \cdot \sin\theta \cdot \sin(\phi) \tag{2.2b}$$

$$z = r \cdot \sin\theta. \tag{2.2c}$$

2.1.2 The solid angle

The *solid angle*, also used to define some radiometric quantities, can be easily introduced in the polar coordinate system (Figure 2.2). Consider a generic point $P(r, \theta, \phi)$. The infinitesimal area, corresponding to the azimuth and zenith increments ($d\phi$ and $d\theta$, respectively), is $dA = r^2 \sin\theta d\theta d\phi$. The infinitesimal solid angle, $d\Omega$, is the ratio between the area dA and the square of its distance r from the origin O

$$d\Omega = \sin\theta d\theta d\phi \qquad [\text{sr}]. \tag{2.3}$$

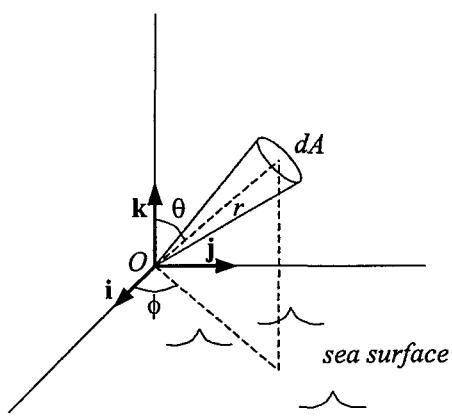


FIGURE 2.2: Schematic representation of the solid angle.

2.1.3 Radiant energy

The *radiant energy* U, associated to the electromagnetic field flowing across a surface of area ΔA for a time ΔT, can be expressed as

$$U = \int_{\Delta A} \int_{\Delta t} \frac{1}{\mu_0} \mathbf{E}(\mathbf{p}, t, \lambda) \times \mathbf{B}(\mathbf{p}, t, \lambda) \cdot \mathbf{n} dA dt, \qquad [\mathrm{J}] \tag{2.4}$$

with $\mathbf{E}$ and $\mathbf{B}$ being the electric and magnetic field, $\mathbf{n}$ the direction perpendicular to the surface ΔA in the point $\mathbf{p}$, t the time and μ_0[1] the permeability of the free space. The quantity

$$\mathbf{S}(\mathbf{p}, t, \lambda) = \frac{1}{\mu_0} \mathbf{E}(\mathbf{p}, t, \lambda) \times \mathbf{B}(\mathbf{p}, t, \lambda) \tag{2.5}$$

is known as *Poynting vector* (see, for example, Halliday et al., 2001).

The radiant energy can be defined for a specific wavelength domain according to the case considered. For example, in the visible region (Figure 2.3), it ranges between 400 nm and 700 nm. If we are considering the radiant energy reaching the *photo detector* (see Section 2.1.5) of an operational spectral instrument, this range can be of the order of few manometers.

The properties of the light field are described in marine optics by means of quantities derived from the radiant energy. These derived quantities are defined in order to take into account variations of the radiant energy with respect to other fundamental quantities like the time (t), the solid angle (Ω), the area of the surface collecting the light field (A) and the wavelength (λ). The explicit dependence from these quantities may be hereafter omitted for brevity of notation.

2.1.4 Radiant power and intensity

The *spectral radiant power*, Φ, is the radiant energy per unit of time and wavelength

$$\Phi = \frac{\partial^2 U}{\partial t \partial \lambda} \qquad [\mathrm{J} \cdot \mathrm{s}^{-1} \cdot \mathrm{nm}^{-1}], \tag{2.6}$$

[1] $\mu_0 = 4\pi \cdot 10^{-7}\ T \cdot m \cdot A^{-1}$

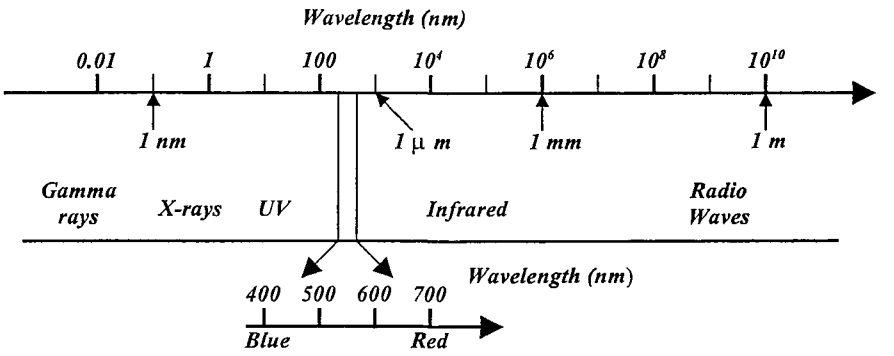


FIGURE 2.3: Schematic representation of the electromagnetic field domains.

The spectral radiant power is useful to characterize the optical behavior of the medium. In fact, variations of the propagation direction of the incident light, or its transformation into other forms of energy, can be directly associated with changes of the spectral radiant power (see Section 2.2.1).

The *spectral radiant intensity*, I, is the spectral radiant power per unit of solid angle

$$I = \frac{\partial^3 U}{\partial t \partial \Omega \partial \lambda} \qquad [\mathrm{J \cdot s^{-1} \cdot sr^{-1} \cdot nm^{-1}}]. \tag{2.7}$$

2.1.5 Irradiance

The *spectral irradiance*, E, is the spectral radiant power per unit of area collected by a plane detection surface

$$E = \frac{\partial^3 U}{\partial t \partial A \partial \lambda} \qquad [\mathrm{J \cdot s^{-1} \cdot m^{-2} \cdot nm^{-1}}]. \tag{2.8}$$

There are two principal directions for irradiance measurements. When the detector is oriented in the $\mathbf{k}$ direction (i.e., towards the sky zenith), the spectral downward irradiance, E_d, is measured. The opposite orientation corresponds to spectral upward irradiance, E_u.

Figure 2.4(a) schematically represents an instrument for measuring the spectral downward irradiance E_d. The *diffuser* corresponds to the part of the instrument collecting

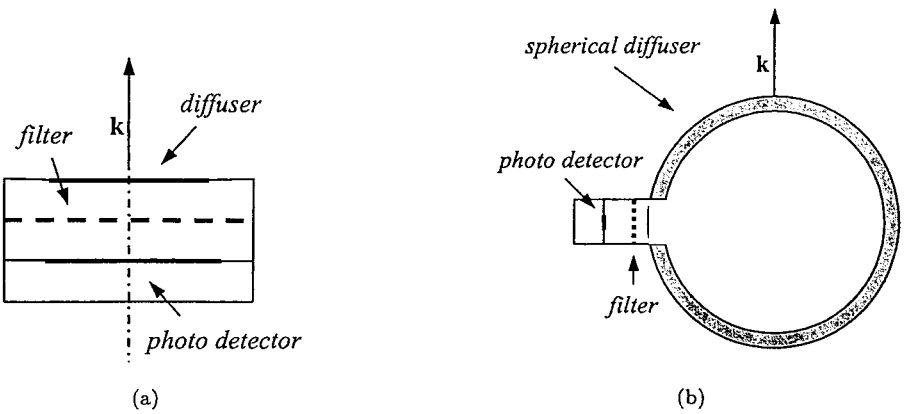


FIGURE 2.4: Schematic representation of instruments for a) irradiance and b) scalar irradiance measurements (adapted from Mobley, 1994).

the environmental light, while the *filter* lets only a very small wavelength window be visible to the *photo detector* (few manometers, but it can vary from an instrument to another according to specific needs).

Consider a light field corresponding to a Poynting vector (Equation 2.5) that does not vary within the detector diffuser surface. In this case the radiant energy incident on the collecting plane surface, depends on the cosine of the angle between the detector's principal axis and the light field propagation direction. This dependence is indicated as the *cosine response* of the irradiance detector.

The quantity corresponding to the spectral irradiance, but referred to the radiant energy source, is called spectral *exitance*. It represents the emitted radiant energy per unit of time, wavelength and surface area.

2.1.6 Scalar irradiance

The *spectral scalar irradiance*, E_{0d}, is the spectral radiant power per unit of area collected by a spherical detection surface (Figure 2.4(b)). Specific measurements of scalar irradiance can be made considering the radiant energy related to a hemisphere only. In this case, the hemisphere of interest can be defined considering a plane perpendicular to $\mathbf{k}$ and separating the space into two parts. The upward scalar irradiance, E_{0u}, takes

into account the radiant energy of the part for which $\mathbf{k} \cdot \mathbf{h} < 0$, being $\mathbf{h}$ a vector with the origin in the plane and belonging to the part considered. The downward scalar irradiance, E_{0d}, is referred to the other part.

2.1.7 Radiance

The *spectral radiance*, L, is the spectral radiant power per unit of area and solid angle collected by a plane detection surface (Figure 2.5)

$$L = \frac{\partial^4 U}{\partial t \partial A \partial \Omega \partial \lambda} \qquad [\mathrm{J \cdot s^{-1} \cdot m^{-2} \cdot sr^{-1} \cdot nm^{-1}}]. \tag{2.9}$$

The spectral radiance gives a detailed description of the light field, and all other quantities previously introduced can be derived from the spectral radiance field. For example, the spectral radiant flux is described in terms of the spectral radiance in Equation 2.10a. Subsequent equations are those for the spectral intensity (Equation 2.10b), upward and downward irradiance (Equation 2.10d and 2.10c) and upward and downward scalar irradiance (Equation 2.10f and 2.10e)

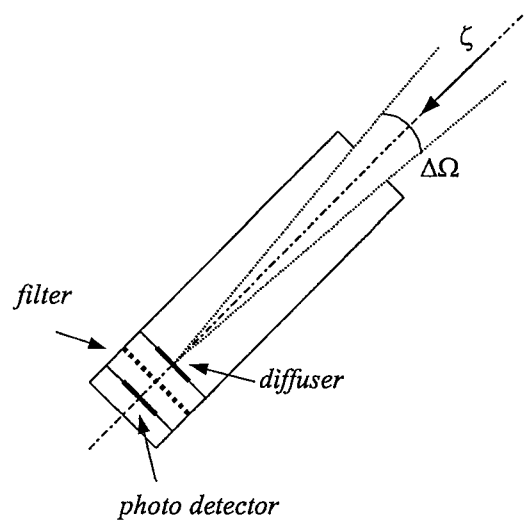


FIGURE 2.5: Schematic representation of an instrument for the radiance field measurement (adapted from Mobley, 1994).

$$\Phi = \int_{\phi=0}^{2\Pi} \int_{\theta=0}^{\Pi/2} \int_{\Delta A} L \cos\theta \sin\theta d\theta d\phi dA \qquad [\mathrm{J \cdot s^{-1} \cdot nm^{-1}}] \tag{2.10a}$$

$$I = \int_{\Delta A} L \cos\theta \sin\theta dA \qquad [\mathrm{J \cdot s^{-1} \cdot sr^{-1} \cdot nm^{-1}}] \tag{2.10b}$$

$$E_d = \int_{\phi=0}^{2\Pi} \int_{\theta=0}^{\Pi/2} L \cos\theta \sin\theta d\theta d\phi \qquad [\mathrm{J \cdot s^{-1} \cdot m^{-2} \cdot nm^{-1}}] \tag{2.10c}$$

$$E_u = \int_{\phi=0}^{2\Pi} \int_{\theta=\Pi/2}^{\Pi} L|\cos\theta| \sin\theta d\theta d\phi \qquad [\mathrm{J \cdot s^{-1} \cdot m^{-2} \cdot nm^{-1}}] \tag{2.10d}$$

$$E_{0d} = \int_{\phi=0}^{2\Pi} \int_{\theta=0}^{\Pi/2} L \sin\theta d\theta d\phi \qquad [\mathrm{J \cdot s^{-1} \cdot m^{-2} \cdot nm^{-1}}] \tag{2.10e}$$

$$E_{0u} = \int_{\phi=0}^{2\Pi} \int_{\theta=\Pi/2}^{\Pi} L \sin\theta d\theta d\phi \qquad [\mathrm{J \cdot s^{-1} \cdot m^{-2} \cdot nm^{-1}}]. \tag{2.10f}$$

2.2 Inherent and Apparent Optical Properties

2.2.1 Inherent Optical Properties

Light, interacting with the molecules of a medium, can be only absorbed or scattered. *Inherent Optical Properties*, IOPs, give a quantitative description of these processes. IOPs depend only on the intrinsic characteristic of the medium, and not on the spatial or temporal distribution of the incident light field.

2.2.1.1 Absorbance, scatterance and emittance

Figure 2.6 shows a collimated monochromatic light beam, propagating in the direction $\boldsymbol{\xi}$, perpendicularly incident on the surface ΔA of a portion of medium of thickens Δr and volume ΔV. The spectral radiant power of the incident light is indicated by Φ_i. As

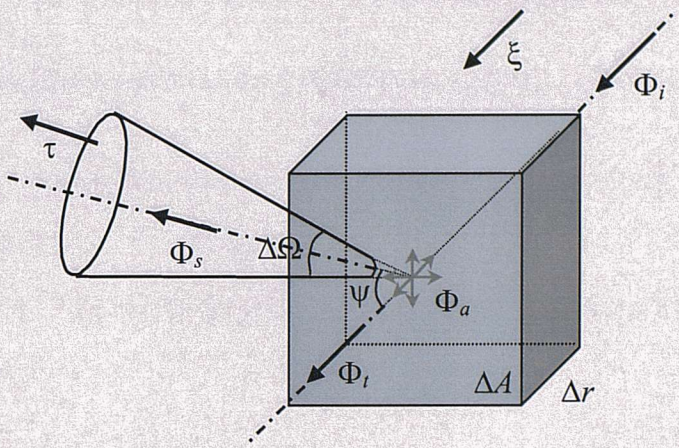


FIGURE 2.6: Schematic representation of the interaction of an incident collimated monochromatic light beam with a portion of the medium

a result of the interaction of the light field with the molecules of the medium, part of the incident light beam is transmitted, Φ_t, without changing its propagation direction. Another part is absorbed (i.e., transformed into other form of energy). Φ_a indicates the spectral power of this absorbing process. Finally, Φ_s indicates the non absorbed incident radiant power propagating in a direction different than $\boldsymbol{\xi}$. Scattering process are here assumed not to affect the wavelength of the incident light. For the principle of energy conservation

$$\Phi_i = \Phi_t + \Phi_a + \Phi_s. \tag{2.11}$$

The optical properties of the medium have to be independent from the incident spectral radiant power. The *spectral absorbance* (A, Equation 2.12a), the *spectral scatterance* (B, Equation 2.12b) and the *spectral transmittance* (T, Equation 2.12c) are then defined considering the ratio of the spectral radiant power resulting from the interaction with the medium and the incident one

$$A = \frac{\Phi_a}{\Phi_i} \tag{2.12a}$$

$$B = \frac{\Phi_s}{\Phi_i} \tag{2.12b}$$

$$T = \frac{\Phi_t}{\Phi_i}. \tag{2.12c}$$

2.2.1.2 Absorption, attenuation and beam attenuation coefficients

As defined the absorbance, scatterance and transmittance depend on the thickness of the medium considered, and not only on its optical properties. In order to eliminate this dependence, the *spectral absorption coefficient* (a, Equation 2.13a) and the *spectral scattering coefficient* (b, Equation 2.13b) are introduced

$$a = \frac{dA}{dr} \qquad [\mathrm{m}^{-1}] \tag{2.13a}$$

$$b = \frac{dB}{dr} \qquad [\mathrm{m}^{-1}]. \tag{2.13b}$$

Both absorption and scattering processes subtract spectral radiant power form the initial propagation direction. The *spectral beam attenuation coefficient*

$$c = a + b \qquad [\mathrm{m}^{-1}] \tag{2.13c}$$

takes into account the conjunct effect of these two processes

2.2.1.3 Single scattering albedo

The afore introduced IOPs can be interpreted as the quantification of the fraction of radiant power that will be subject to the corresponding process per unit of length of the medium thickness. Considering the quantum nature of the light, a probabilistic interpretation is also possible. In the hypothesis of a single photon travelling in the direction $\boldsymbol{\xi}$, the probability per unit of length that it is absorbed by the medium corresponds to the spectral absorption coefficient. Analogous interpretation is valid for the other quantities.

The theorem of the conjunct probability for two events, $\mathcal{S}$ and $\mathcal{I}$, states that

$$P(\mathcal{S},\mathcal{I}) = P(\mathcal{S}|\mathcal{I})P(\mathcal{I}), \tag{2.14}$$

being $P(\mathcal{S},\mathcal{I})$ the probability that both $\mathcal{S}$ and $\mathcal{I}$ happens, $P(\mathcal{I})$ the probability of the event $\mathcal{I}$, and $P(\mathcal{S}|\mathcal{I})$ the probability that $\mathcal{S}$ happens if $\mathcal{I}$ happens. Let $\mathcal{S}$ be the scattering event and $\mathcal{I}$ the interaction event. $P(\mathcal{S},\mathcal{I})$ is then corresponding to the spectral scattering coefficient, while $P(\mathcal{I})$ (i.e., probability of having an interaction) is the sum of the absorption and scattering coefficients. Finally, $P(\mathcal{S}|\mathcal{I})$ is the probability that, if there is an interaction, it is not an absorption process. Thus, $P(\mathcal{S}|\mathcal{I})$ represents the photon probability of survival or *single spectral scattering albedo*, ω_0, and can be expressed, according to the Equation 2.14, as

$$\omega_0 = \frac{b}{c}. \tag{2.15}$$

2.2.1.4 Volume scattering function

The spectral scattering coefficient takes into account the amount of the incident spectral radiant power that has not been absorbed by the medium. In order to define its spatial distribution, indicate the scattering direction with $\boldsymbol{\tau}$, and the angle between $\boldsymbol{\tau}$ and $\boldsymbol{\xi}$ with ψ (see Figure 2.6). The ratio between the scattered spectral intensity I and the corresponding incident spectral irradiance E, per unit of volume of the medium, is the *spectral volume scattering function* $\beta(\psi)$

$$\beta(\psi) = \frac{d}{dV}\left(\frac{I(\psi)}{E}\right) \qquad [\mathrm{m}^{-1} \cdot \mathrm{sr}^{-1}]. \tag{2.16}$$

The spectral volume scattering function fully describes the optical properties of the medium, but to be determined it requires an experimental effort.

In the hypothesis of scattering process characterized by axial symmetry around the direction $\boldsymbol{\xi}$, the *back-scattering coefficient*, b_b, is defined as

$$b_b = \int_{\pi/2}^{\pi} \beta(\psi) \sin\psi d\psi \qquad [\mathrm{m}^{-1}]. \tag{2.17}$$

Similarly, the *forward scattering coefficient*, b_f, is

$$b_f = \int_0^{\pi/2} \beta(\psi) \sin\psi d\psi \qquad [\mathrm{m}^{-1}]. \tag{2.18}$$

2.2.2 Inherent Optical Properties of seawater and its components

The possibility to retrieve the phytoplankton, the coloured dissolved organic matter and the non pigmented particulate matter from the radiance leaving the sea surface depends on i) the different inherent optical properties that characterise the various water components and ii) the fact that these properties also differ from those of the pure seawater. A specific description of the variability of these properties within the data set applied in this study is provided in Chapter 3. The aim of this section is to describe the inherent optical properties of the pure seawater and its optically significant seawater components, highlighting elements that are relevant for the development of bio-optical models and ocean colour algorithms.

2.2.2.1 Absorption by pure seawater and its components

The experimental determination of the absorption coefficient of the pure seawater, as well as that of its optically significant components, represents a complex task and some of the major limiting factors are hereafter summarized. First of all, the pure seawater is only weakly absorbing in the blue region of the light spectra, and thus very sensitive instruments are required. Then, the scattering effect has to be accounted for in order to avoid estimating the absorption coefficient. In addition, the need to derive the absorption coefficients through "concentrated" samples resulting from seawater filtration (due to the low concentrations naturally occurring in-situ) increases the complexity of the measurement protocol and the chance of error.

Absorption by pure seawater Various investigators focused on the determination of the absorption properties of the pure seawater (Morel and Prieur, 1977; Smith and

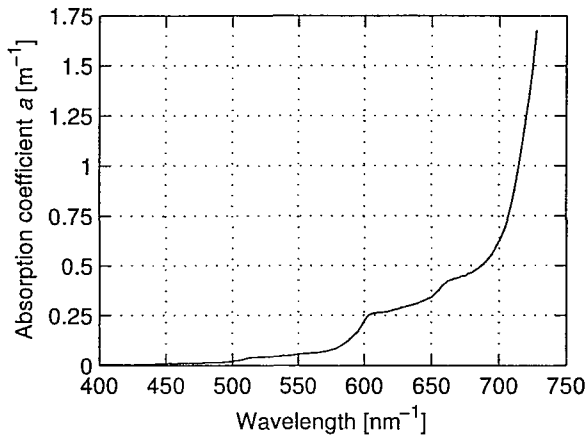


FIGURE 2.7: Spectral plot of the pure water absorption coefficient, as determined by Pope and Fry (1997).

Baker, 1981; Pope and Fry, 1997). The results provided by Pope and Fry (1997) are those commonly applied in the optical oceanography studies of the recent years. As represented in Figure 2.7, the seawater absorption strongly increases with the wavelength upon 500 nm. Since the absorption properties of the pure seawater do not almost depend on the level of salinity and the temperature of the seawater, its effect on the water-leaving radiance can be considered as a common constant additive background with respect to the absorption effects of the other optically significant components.

Absorption by phytoplankton The phytoplankton absorption spectrum is characterised by two main peaks, one in the blue at 490 nm and one in the red at 665 nm (see Figure 2.8). This absorption spectrum is due to various pigments, among which the *chlorophyll-a* is the main important and is used as an index to represent the phytoplankton itself. Photosynthetic pigments are distributed non randomly among the cell, but present localized packages called *chloroplasts* (Kirk, 1994). Various factors, such as the availability of nutrients, the physiological state of the cell and the illumination condition, can determine the number, the localization and the size of the chloroplasts. From the optical point of view, this affects the efficiency of the cell in absorbing the light given the same amount of *chlorophyll-a* (i.e., *pigment packaging effect*), as depicted in Figure 2.8. Moreover, different pigments are characterized by a different absorbing efficiency

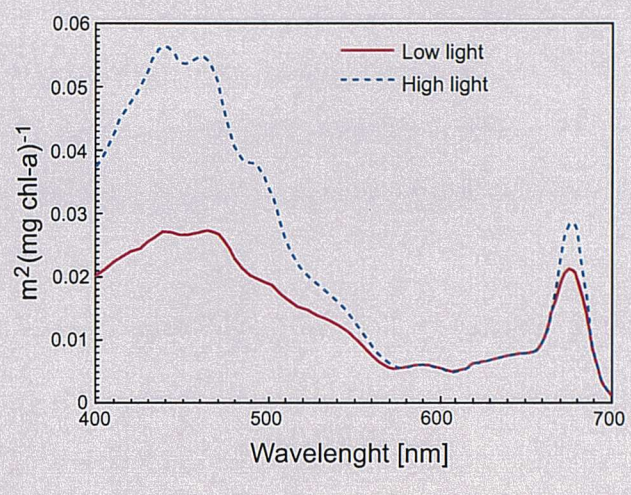


FIGURE 2.8: Example of the variability of the phytoplankton specific absorption coefficient due to the adaptation to different conditions of illumination. The two main absorption peaks characteristic of the phytoplankton absorption coefficient spectra, one in the blue at 490 nm and one in the red at 665 nm, can be observed regardless of the absolute value of the specific absorption coefficient. Adapted from Sathyendranath (2000).

and can appear in different relative proportions, contributing in this way to increase the phytoplankton absorbing variability. It has to be noticed that the natural variability of the *specific absorption* (i.e., the absorption normalized by the *chlorophyll-a* concentration) represents a limit to the accuracy in retrieving the phytoplankton component from the water-leaving radiance.

Bricaud et al. (1995) investigated the spectral absorption due to the phytoplankton component. The proposed parametrisation is here presented with the notation reported in the original work

$$a^*_{ph}(\lambda) = A(\lambda)\langle chla\rangle^{-B(\lambda)}, \tag{2.19}$$

where $a^*_{ph}(\lambda)$ is the specific spectral absorption coefficient, $\langle chla\rangle$ is the *chlorophyll-a* concentration and the model parameters, $A(\lambda)$ and $B(\lambda)$, are defined through a least square fit with experimental data.

Absorption by coloured dissolved organic matter The coloured dissolved organic matter is represented by a group of dissolved organic substances, consisting of humic and fulvic acids (Kirk, 1994). Optically, it is characterized by an absorption spectrum that monotonically decreases with the wavelength, and for this reason the coloured

dissolved organic matter is also called "yellow substance". The CDOM originates from the decomposition of organic material, such as phytoplankton, by the action of microbes. During this process, the photosynthetic pigments rapidly loose their original absorption properties.

The spectral dependence of the coloured dissolved organic matter absorption can be expressed as

$$a_{CDOM}(\lambda) = a_{CDOM}(\lambda_0) \exp\left[S(\lambda - \lambda_0)\right], \tag{2.20}$$

where λ_0 is the *reference wavelength*. The *slope* parameters, S, is determined fitting the exponential function to experimental data. Bricaud et al. (1981a) reported values of S ranging between 0.014 and 0.019 nm^{-1}. However, a higher variability of S has also been observed (for example Højerslev (1998) reported values of S between 0.008 and 0.042 nm^{-1}). It is here highlighted that one of the objective of the present work is to verify the possibility to retrieve the S parameter from the water-leaving radiance instead of the absorption coefficient at some specific wavelength (see Chapter 5).

Absorption by non non pigmented particulate matter The non pigmented particulate matter comprehends those particles not included in the phytoplankton component. Its composition can be highly heterogeneous, varying form finely ground quartz sand resuspended from the sea bottom or carried by the wind to microorganism shells, such as those produced by the Coccolithophore (Kirk, 1994). Despite its different origins, the non pigmented particulate matter presents an exponential dependence of the absorption coefficient on the wavelength analogous to that of the CDOM, but characterized by a different *slope* coefficient; values of S within the 0.006 and 0.014 nm^{-1} are reported (Roesler, 1989). Another difference with respect to the spectral absorption coefficient of the CDOM is that the NPPM can show a non negligible absorption also in the blue region of the light. For this reason the NPPM absorption coefficient is parametrised as

$$a_{NPPM}(\lambda) = a_{NPPM}(\lambda_0) \exp\left[S(\lambda - \lambda_0)\right] + cb, \tag{2.21}$$

where the *background* parameter, *cb*, has added with respect to Equation 2.20 to account for the absorbing effects at the higher wavelengths.

2.2.2.2 Back-scattering by pure seawater and its components

The scattering properties of the water and its optically significant constituents are difficult to accurately determine since the light can be scattered in different directions with a probability that may vary of various orders of magnitude. An exhaustive analysis of the processes that determine the scattering effect is out of the scope of this work (for a detailed description, see Mobley, 1994). Instead, here we mainly focus to describe the back-scattering properties of the pure seawater and its constituents. This is motivated by the fact that the back-scattering coefficient can be directly linked to the radiance leaving the sea-surface through a reflectance model, such as that presented in Section 2.2.3.2. Notice also that the scattering effects of the coloured dissolved organic matter are negligible, and therefore are not hereafter mentioned.

Back-scattering by pure seawater The scattering process by the water is due to the fluctuation "in the number of molecules in a volume ΔV, where ΔV is small with respect to the wavelength of the light but large compared to atomic scales" (Mobley, 1994). On the basis of the Einstein-Smoluchoski theory, the scattering by the pure water depends on the wavelength as λ^{-4}. The presence of various ions in the pure seawater leads to an increased scattering and to a dependence on the wavelength that can be expressed as λ^{-n}, with $n = 4.32$ (Morel, 1974). The seawater volume scattering function is symmetric with respect to the 90° angle (i.e., with respect to the propagation direction of the incident light), and thus the back-scattering represent the 50% of the total scattering.

Back-scattering by phytoplankton The phytoplankton presents a very low back-scattering coefficient (Bricaud et al., 1981b). Such small values would often not justify

the observed back-scattering measurements taken in natural Case 1 waters and the additional contribution of the back-scattering due to the microscopic organism covarying with the phytoplankton better explains the observed back-scattering (Morel, 1988; Gordon et al., 1988).

Back-scattering by non pigmented particulate matter Various particles, mainly present in the Case 2 waters, can contribute significantly to back scatter the light. As for the pure seawater, the general law that expresses the spectral dependence of the back-scattering due to the non pigmented particulate matter is λ^{-n} (Morel, 1973), with n varying from 0 in turbid coastal water to 2 in oligotrophic waters (Sathyendranath, 1989). Such variability is strongly linked to the specific properties of the non pigmented particulate matter, which can vary both between different areas for the specific regional geochemical properties, as well as in the same region due to the advection by the wind or rivers of particulate matter from remote sources. Despite the analogous power law, the scattering phase function of the non pigmented suspended sediments is quite different from that of the seawater. In fact, inorganic particles are characterised by a highly peaked forward scattering.

2.2.3 Apparent Optical Properties

The *Apparent Optical Properties*, AOPs, depend on both the environmental illumination and observation conditions and not just on the properties of the medium. However, in order to characterize the water properties through the AOPs, the latter have to be as independent as possible from environmental changes. Because of this, quantities as gradients or ratios are often used to define the AOPs.

2.2.3.1 Irradiance reflectance

The *spectral irradiance reflectance*, R, is the ratio of the upwelling and the downwelling spectral irradiance

$$R = \frac{E_u(z)}{E_d(z)}. \tag{2.22}$$

The spectral irradiance reflectance corresponds to the portion of the downward irradiance reflected at depth z, as the result of the interaction of the light field with the medium.

2.2.3.2 Remote Sensing Reflectance

The spectral *Remote Sensing Reflectance*, R_{rs}, is the ratio of the upwelling radiance and the downwelling irradiance, both "just above" the seawater surface ($z = 0^+$)

$$R_{rs}(\theta, \lambda) = \frac{L_u(0^+, \theta, \lambda)}{E_d(0^+, \lambda)} \qquad [\mathrm{sr}^{-1}]. \tag{2.23}$$

The above water spectral irradiance, $E_d(0^+)$, can be derived from the corresponding subsurface ($z = 0^-$) downwelling and upwelling spectral irradiance values through

$$E_d(0^+, \lambda) = \frac{E_d(0^-, \lambda) - 0.49E_u(0^-, \lambda)}{1 - \rho_s(\lambda, \theta = 0)} \qquad [\mathrm{J \cdot s^{-1} \cdot m^{-2} \cdot nm^{-1}}], \tag{2.24}$$

where ρ_s is the surface Fresnel reflectance (see Mobley, 1994). Obviously, the above water irradiance can also be directly measured positioning an irradiance sensor above the seawater surface (in this case it will be hereafter also indicated as E_s).

The above water upward spectral radiance takes the name of *water-leaving radiance*, L_w. It depends on the corresponding subsurface value according to

$$L_w(\theta, \lambda) = \frac{1 - \rho_s(\theta, \lambda)}{n_w^2(\lambda)} L_u(0^-, \theta, \lambda) \qquad [\mathrm{J \cdot s^{-1} \cdot m^{-2} \cdot sr^{-1} \cdot nm^{-1}}], \tag{2.25}$$

where n_w is the refractive index of the seawater. The radiance transmission coefficient through the sea surface depends weakly on the environmental conditions and can be approximated by a constant value (Austin, 1980)

$$\frac{1-\rho_s(\theta,\lambda)}{n_w^2(\lambda)} \simeq 0.544. \tag{2.26}$$

In the hypothesis that λ does not change during the scattering process, the R_{rs} can be expressed (Gordon, 1976; Morel and Prieur, 1977) as function of a and b_b by

$$R_{rs} = f\frac{b_b}{a+b_b}, \tag{2.27}$$

being f dependent on the wavelength.

The *normalized water-leaving radiance*, L_{wn}, is the product of the R_{rs} for the spectral mean sun irradiance at the top of the atmosphere, $f_0(\lambda)$, corrected for the sun-earth distance, $d(\delta)$

$$L_{wn}(\theta,\lambda) = f_0(\lambda)d(\delta)R_{rs}(\theta,\lambda) \qquad [\mathrm{J}\cdot\mathrm{s}^{-1}\cdot\mathrm{m}^{-2}\cdot\mathrm{sr}^{-1}\cdot\mathrm{nm}^{-1}]. \tag{2.28}$$

The correction coefficient $d(\delta)$ takes into account the elliptical shape of the earth orbit

$$d(\delta) = \left[1 + 0.0167\cos\left(\frac{2\pi(\delta-3)}{365}\right)\right], \tag{2.29}$$

being δ the day of the year.

2.2.3.3 Diffuse attenuation coefficient

The *diffuse attenuation coefficient*, k_d $[m^{-1}]$, expresses the spectral downwelling irradiance subtracted for unit length of medium by the absorption and scattering processes

$$\frac{dE_d(z)}{dz} = k_d E_d(z). \tag{2.30}$$

Equation 2.30 describes the exponential dependence of the downward irradiance on the depth z. In this convention $\mathbf{k}$ is oriented toward the zenith, so z decreases toward the sea bottom. Thus, integrating the 2.30 between z_0 and z (with $z < z_0$) one obtains

$$E_d(z) = E_d(z_0)e^{k_d(z-z_0)}. \tag{2.31}$$

Equivalent diffuse attenuation coefficients can also be defined for other quantities such as the upwelling irradiance E_u or radiance L_u.

2.3 Empirical and semi-analytical ocean colour algorithms

This section overviews the elements basis of the empirical and semi-analytical algorithms (Sections 2.3.1 and 2.3.2, respectively). The neural network method applied in the presented study belongs to the class of the empirical algorithms. The NN effectiveness to represent any complex continuous mapping between multidimensional input (i.e., R_{rs} at different individual wavelengths) and output space is one of the main reasons for its use to develop Case 2 water ocean colour algorithms. A description of the neural network approach requires slightly more extensive statistical elements, and it is separately presented in Chapter 4.

2.3.1 Band-ratio empirical algorithms

Band-ratio empirical algorithms for the *Chl-a* retrieval are based on a direct regression of the ratio of reflectances at two wavelengths, or the maximum of the ratio of the reflectances at more wavelengths, to the *chlorophyll-a* concentration

$$Chl\text{-}a = f\left(a_0; a_1; \ldots; a_n; \frac{R_{rs}(\lambda_i)}{R_{rs}(\lambda_j)}\right), \tag{2.32}$$

where the algorithm coefficients $(a_0; a_1; \ldots; a_n)$ are determined through the minimization of the squared difference between the regressional function, f, and the experimental data;

λ_i can vary among 443, 490 and 510 nm, while λ_j is usually 555 nm.

Various authors investigated different regressional function (i.e., power, hyperbolic, cubic and multiple regression) and a review of these algorithms for the *chlorophyll-a* retrieval was presented by O'Reilly et al. (1998). In that work, and in its successive revision (O'Reilly et al., 2000), the authors showed that the most accurate results were obtained from the so called OC4 band-ratio empirical algorithm, developed for the processing of SeaWiFS data. This algorithm consists of a third order polynomial interpolation of the logarithm of the *chlorophyll-a* concentration

$$Chl\text{-}a = 10.0^{\left(a_0+\sum_{i=1}^{3} a_i R_{4S}^i\right)} + a_4, \tag{2.33}$$

where R_{4S} is $\max\left\{R_{555}^{\lambda}\right\}$, that is the maximum of the ratio of the R_{rs} at 443, 490 and 510 nm with respect to the R_{rs} at 555 nm (Maximum Band Ratio approach, MBR), and a_i, with $i = 1,\ldots,3$, are the model parameters to be defined on the basis of the experimental data. The MBR approach was adopted to use the R_{rs} ratio at the most suitable wavelengths for different *chlorophyll-a* concentration ranges. Specifically, the algorithm switches from the 443 nm channel in oligotrophic waters to the 510 nm channel in eutrophic waters. The OC4 algorithm (namely, the "V4" version, O'Reilly et al. (2000)) is routinely used for the processing of the SeaWiFS data. An evaluation of the performance of the OC4V4 algorithm with respect to other algorithms developed on the basis of the coastal data set applied in the present study is presented in Section 5.1.

Empirical band-ratio algorithms allow a straight forward implementation, very low computational cost and perform well in Case 1 conditions. On the other side the band-ratio approach performs less well in Case 2 waters due to the independent variability of the CDOM and NPPM with respect to the *Chl-a* (i.e., in Case 2 waters the same amount of *Chl-a* can correspond to different R_{rs} band-ratio values and this reduces the intrinsic algorithm accuracy with respect to Case 1 conditions). Actually, one of the main reasons to address the present study to the development of neural network algorithms is to try

to overcome some of the limitations deriving from the band-ratio approach by using the R_{rs} at individual wavelength as algorithm input.

It is important to furthermore highlight that a strong limitation of the empirical approach, especially for what concerns the neural networks, is that they do not embody a physical but a merely statistical representation of the relationship between optically active seawater components and water-leaving radiance.

2.3.2 Semi-analytical algorithms

Semi-analytical algorithms relate the inherent optically properties of the water constituents to the spectra of the water-leaving radiance through a reflectance model such as that presented in Equation 2.27 (Gordon, 1976; Morel and Prieur, 1977), and here rewritten for clarity

$$R_{rs} = f\frac{b_b}{a + b_b}, \tag{2.34}$$

where a and b_b are the total absorption and the total back scattering coefficients, respectively, while f is a proportional constant factor.

Both the total absorption and the total back-scattering coefficients of Equation 2.34 can be expressed through the sum of the absorption and scattering contribution of the pure seawater and those of its optically active components, namely

$$a = a_w + a_{CDOM} + a_{ph} + a_{NPPM} \tag{2.35a}$$

$$b_b = b_{b,w} + b_{b,ph} + b_{b,NPPM} \tag{2.35b}$$

where the subscription w indicates the seawater contribution. Examples of bio-optical models representing the optical properties of the water components are presented in Sections 2.2.2.2 and 2.2.2.1.

Since the R_{rs} is expressed in Equation 2.34 through a linear function upon the IOPs, the retrieval of seawater optical properties correspond to a matrix inversion problem, as described in Hoge and Lyon (1996, 1999). Differently, numerical optimization methods have in general to be used when the forward model is not a linear combination of the unknowns (see for example, Lee et al., 1999; Garver and Siegel, 1999). In such cases an additional limiting factor may be represented by the computational cost.

The semi-analytical formulation has the advantage to be based on a physical model. Nevertheless, it still may not represent an universal solution since the relationship between IOPs and water constituents can be driven by specific regional properties that contribute to prevent the possibility of developing an universal ocean colour algorithm.

Chapter 3

The CoASTS data

This study has been carried out on the basis of an experimental data set collected in the framework of the CoASTS (Coastal Atmosphere and Sea Time Series) program at the Acqua Alta Oceanographic Tower (AAOT) (Zibordi et al., 2002a). The CoASTS program and the characteristics of the sampling site are presented in Section 3.1. The software toolbox, specifically developed for the processing of the *in water* optical profile measurements, is presented in Section 3.2. The author has been strongly involved in the development of the processing code, as well as in its implementation through Graphical User Interfaces (GUI). Finally, Section 3.3 describes the characteristics of the CoASTS data.

3.1 CoASTS

The development and validation of the algorithms proposed in this study have been carried out on the basis of an extensive set of experimental data collected in the framework of the CoASTS program (Zibordi et al., 2002a). The aim of this project is to create a comprehensive time series of atmospheric and marine data representative of the northern Adriatic Sea for i) the development of site specific marine and atmospheric algorithms, ii) calibration and validation activities for ocean colour sensors and iii) time-series studies on the bio-geo-chemical measured quantities.

(a)

(b)

FIGURE 3.1: The CoASTS program has been relying on the Acqua Alta Oceanographic Tower (a), located in the Central Mediterranean Sea (b).

Since fall 1995, the CoASTS program has been relying on the use of the Acqua Alta Oceanographic Tower (Figure 3.1(a)) for the execution of field measurements. The AAOT site (Figure 3.1(b)) is located in the Norther Adriatic Sea, 8 nautical miles south-east off the Venice Lagoon (latitude 45°19′N, longitude 12°30′E) and exhibits water types close to the separation threshold suggested by Loisel and Morel (1998) for Case 1 and Case 2 waters (with roughly one third of the CoASTS data pertaining to Case 2 (Berthon et al., 2002)). The Case 2 waters are mostly determined by the river inputs (i.e., Piave, Livenza and Tagliamento). The aerosol type, occasionally maritime, is mostly continental (i.e., determined by atmospheric inputs from the close Po Valley). The former characters give to the site, the property of representing most of the northern Adriatic Sea and supports the use of the CoASTS data set for the development of regional ocean colour algorithms.

3.2 The CoASTS data processing system[1]

Whereas the CoASTS program occasionally produced above water radiance measurements (Zibordi et al., 2002b), the R_{rs} data used in this study are those derived from *in water* optical profiles. It is important to point out that environmental conditions

[1]Some elements of this section was presented in D'Alimonte and Zibordi (2001)

can influence the accuracy of the surface values extrapolated from optical profile measurements. This is particularly evident in coastal regions, where the water depth may present a limiting factor. In fact, the extrapolation layer for computing subsurface radiometric quantities has to be defined considering two concurrent factors: on one side it has to be reduced to limit the perturbations due to the optical stratification in the water column; on the other it is necessary to leave as many profile points as possible to reduce the wave effects for an effective regressional process (Zibordi et al., 2004).

A processing code was specifically developed to handle *in water* optical profile data collected in complex conditions in order to ensure the quality of the sub-subsurface values. The software package consists of two major programs supporting i) the calibration (i.e., conversion from digital number to physical units) and formatting of the raw input data and ii) the filtering of the data and the computation of the final products (i.e., remote sensing reflectance, diffuse attenuation coefficient,...). These programs ensure the possibility of storing in an information file the processing parameters selected by the user. The processing information can then be reloaded for an automatic data re-processing. The processing steps and the functionalities provided by the software toolbox are presented in the following sections.

3.2.1 Data calibration and formatting toolbox

This program ensures: i) the data calibration, ii) the screening of data as a function of an user-selectable tilt (i.e., inclination of the *in water* sensor) threshold, iii) the visualization of data as a function of depth and iv) the creation of output ASCII data files. Processing information (i.e., input file names, depth offset of the *in water* light sensors with respect to the pressure sensor, geographical location of the sampling site, water depth) are provided by the user through the GUI displayed in Figure 3.2.

The absolute calibration of the radiometric quantities is obtained from

$$\Im(\lambda, t) = C_c I_f [D(\lambda, t) - \overline{D}(\lambda)], \tag{3.1}$$

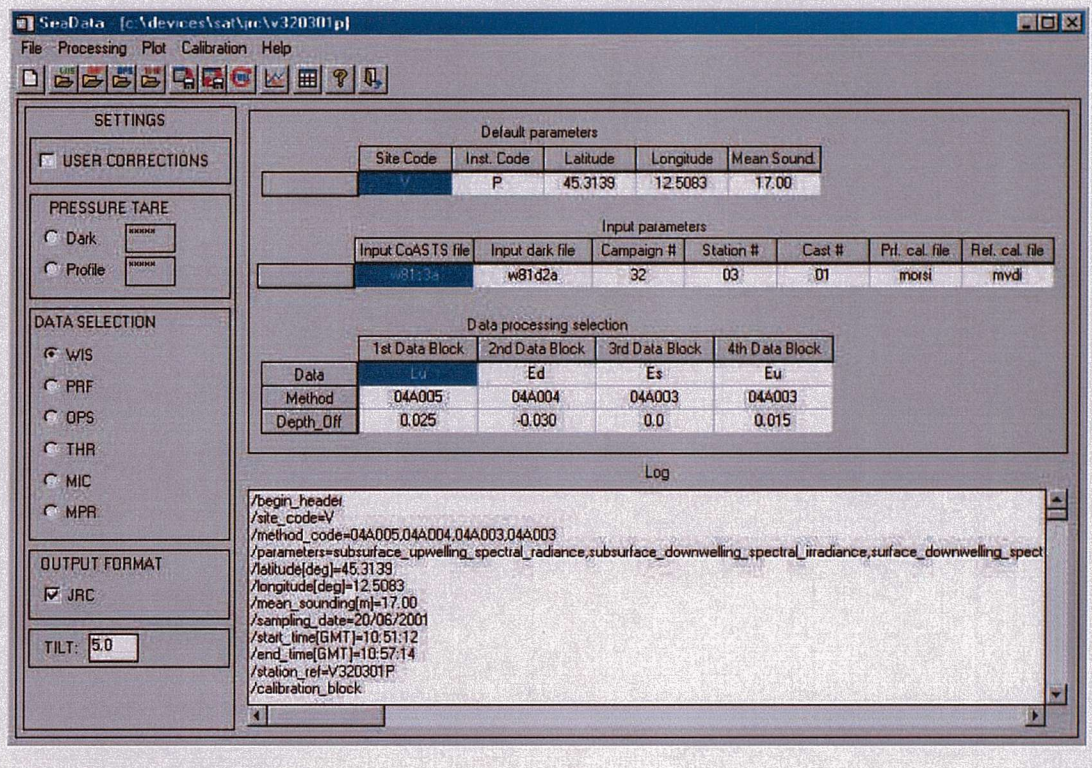


FIGURE 3.2: Graphical user interface of the software toolbox for the data calibration and formatting(D'Alimonte and Zibordi, 2001).

where $\Im$ indicates E_s, E_d, E_u and L_u (see Section 2). D, t and λ are the digital number, the time and at the center-wavelength of the sensor, respectively. $\overline{D}$ is the dark value. C_c is the *absolute calibration coefficient.* I_f is the *immersion coefficient* (i.e., the coefficient accounting for change in sensor responsivity when the in-air calibration is applied to *in water* measurements).

Depth values of the radiometric measurements are derived from the digital values provided by the pressure sensor, after removing the pressure-offset and applying the corresponding calibration coefficient. Radiometric data can then be displayed in logarithmic scale as a function of depth. The above water irradiance data are also displayed as a function of the depth of the corresponding *in water* profile data. This ensures the identification of artifacts, which may affect optical profile data.

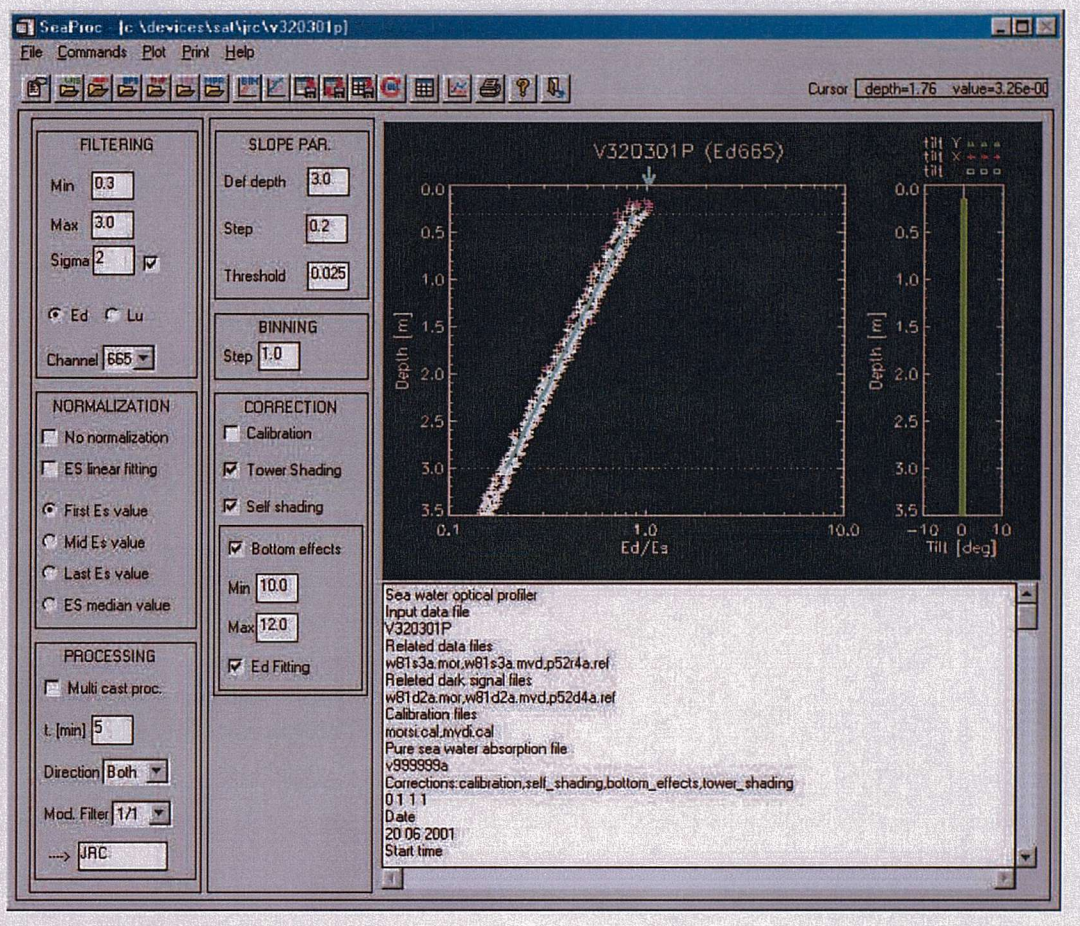


FIGURE 3.3: Graphical user interface of the software toolbox for the data filtering and for extrapolating sea surface radiometric quantities from optical profile measurements (D'Alimonte and Zibordi, 2001).

3.2.2 Data filtering and processing toolbox

The main functionality provided by this toolbox (Figure 3.3) are: i) the definition of extrapolation intervals for the computation of subsurface values and bottom reflectance, ii) the selection of normalization methods for *in water* data with respect to *above water* reference measurements, iii) the filtering of outliers in the subsurface extrapolation interval and the removal of data below the noise threshold, iv) the correction for perturbation effects (i.e., instrument self-shading, bottom effects and tower-shading, being the latter correction specific for measurements collected the AAOT site) and v) the visualization (Figure 3.4) of data as a function of depth together with extrapolated values (i.e., for both surface and bottom). Individual processing steps are hereafter presented.

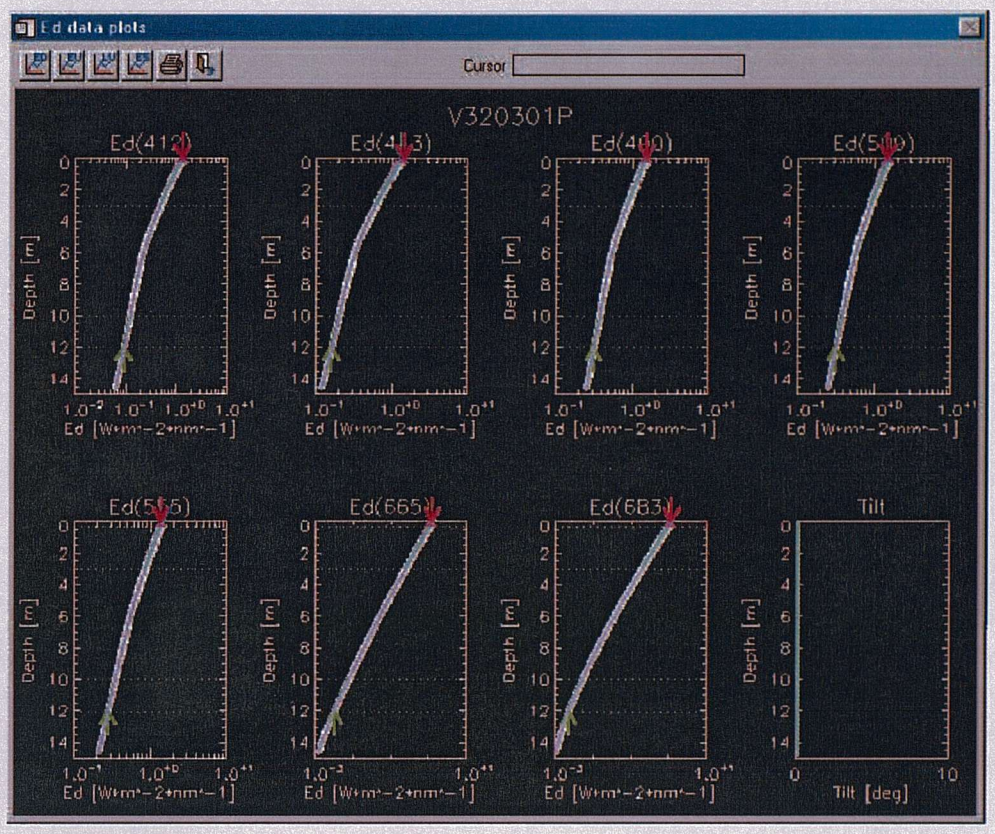


FIGURE 3.4: Example of visualization of E_d data as a function of depth together with extrapolated values.

3.2.2.1 Definition of the extrapolation intervals

The definition of the appropriate extrapolation interval for the extraction of subsurface data from *in water* optical profiles is a key element in coastal waters. Optical stratification may occur in the first few meters just below the surface, and a non appropriate selection of the extrapolation interval can produce large uncertainties in the final products. Because of this, even though default values are provided, the user can identify, on the basis of the visualization of the specific profile to be processed, an "optimum" extrapolation range that satisfies the requirement of a linear decay. In addition to the subsurface extrapolation interval, the toolbox also provides the possibility of selecting a bottom extrapolation interval (when correction for the bottom effect is required) to estimate the bottom reflectance from profile data.

3.2.2.2 Correction with respect to the above water irradiance

In-water radiometric quantities can be corrected with respect to above water irradiance data ($E_s(\lambda,t)$, with t expressing dependence with time), according to:

$$\Im(z,\lambda,t_0) = \Im(z,\lambda,t)\frac{E_s(\lambda,t_0)}{E_s(\lambda,t)} \tag{3.2}$$

where $\Im(z,\lambda,t_0)$ identifies radiometric quantities as they would have been taken at all depths z at the same time t_0, $E_s(\lambda,t)$ is the *above water* irradiance taken at the same time t as the *in water* data, while $E_s(\lambda,t_0)$ is the above water irradiance at time t_0 (where t_0 is generally chosen as to coincide with the start of the cast). Alternatively, the correction factor can be referred to the median of the above water irradiance data corresponding to the records included in the extrapolation interval. The correction process can be skipped if the reference data are not available or are strongly affected by noise (for instance, by the ship motion). Additionally, for the aforementioned case, it is also possible to apply a linear fit to the measurements of the above water irradiance and use the fitted values to correct the *in water* data.

3.2.2.3 Data filtering

The regression of the data in the surface layer is carried out using a least squares linear fit. This technique may produce biased results in presence of outliers (for instance produced by wave effects). This bias may increase as the number of optical profile points per unit depth decreases. Because of this, two complementary filters have been implemented.

The first filter ensures the removal of points distant by more than $n\sigma$ from the initial regression line obtained with all profile data (where n is user selectable and σ is the standard deviation of the distances of points from the initial regression line).

The second filter computes successive regressions by adding depth increments to the initial extrapolation interval. The slope of the resulting regression is then compared

with the slope of the initial regression. This filter removes all points below a certain depth if, for a given depth increment, the slope of the fitting line changes by more than a threshold value (user selectable) with respect to the initial value.

In order to verify the effects of the filtering, regression trials are possible choosing a specific quantity (i.e., L_u or E_d) and a specific λ. It is convenient to use E_d and $\lambda = 665$ nm. The use of E_d may ensure direct comparison of the extrapolated value $E_d(0^-,\lambda)$ with $E_s(\lambda)$, to immediately identify "difficult" situations (i.e., those showing very different $E_d(0^-,\lambda)$ and $E_s(\lambda)$ values). The use of a red channel (665 nm), where seawater is characterized by high absorption and sensors show a fast drop to noise levels as a function of depth, helps in excluding noisy data. All records (i.e., all radiometric quantities at all channels) taken at the same time of data flagged by the filter, are removed. This is done under the assumption that all data collected at the same time are affected by highly correlated perturbations.

3.2.2.4 Correction for perturbation effects

Measurement perturbations due to self-shading, bottom effects and tower-shading (the latter for the specific case of the WiSPER data (Zibordi et al., 2002a), see below) can be removed through single or multiple selection(s) assuming all perturbations are independent from each other.

Self-shading correction: The self-shading correction factor is computed using the scheme proposed by Gordon and Ding (1992) and the parameterizations suggested by Zibordi and Ferrari (1995) and Muller and Austin (1995). The computations are carried out using i) the sun zenith, ii) the seawater absorption, iii) the seawater single scattering albedo, iv) the ratio between the instrument radius and the sensor entrance optics radius and v) the diffuse over direct above water *downward irradiance ratio*, $r(\lambda)$, for instance obtained as

$$r(\lambda) = \frac{E_i(\lambda)}{E_s(\lambda) - E_i(\lambda)}, \tag{3.3}$$

where $E_i(\lambda)$ is the diffuse above water downward irradiance.

Bottom effect correction: The correction for the bottom effect is estimated with the analytical model of Zibordi et al. (2002a) adapted from Maritorena et al. (1994). Computations are performed using i) the *bottom reflectance*

$$R_b(\lambda) = \frac{E_u(z_b, \lambda)}{E_d(z_b, \lambda)}, \tag{3.4}$$

obtained from each profile at depth z_b near the bottom and ii) the mean diffuse attenuation coefficient for the depth interval $z_0 - z_b$ (where z_0 is the subsurface depth identifying the upper value of the extrapolation interval).

Tower-shading correction: The tower-shading correction is obtained from look-up tables (Doyle and Zibordi, 2002) resulting from Monte Carlo computations of the radiance and irradiance fields at the AAOT site and for the specific point where the optical profile measurements used in the present work are taken.

3.2.2.5 Sub-surface values

Sub-surface values (i.e. $E_u(0^-,\lambda)$, $L_u(0^-,\lambda)$ and $E_d(0^-,\lambda)$) are obtained from the exponent of the intercept given by the least-squares linear regression of the logarithm of optical profile measurements, (i.e., data not removed by filters and corrected for perturbation effects) versus z, within the extrapolation interval identified by $z_0 < z < z_1$. Generally for the CoASTS profiles, $0.3 < z_0 < 1$ m and $2.5 < z_1 < 4.5$ m. The negative value of the slope of the regression fit is the diffuse attenuation coefficient (specific for the extrapolation interval). An option, called "multi-cast processing", enables the processing of data from different profiles (collected within a user definable time interval, generally of few minutes) as a single optical profile. This solution, which cumulates profile data, increases the statistical significance of the regressional analysis applied to data characterized by a low depth resolution and affected by large perturbations (for instance caused by sea surface effects).

For each specific quantity it is possible to display profile data and inclination as a function of depth for all λ (Figure 3.4). Regression lines for the extrapolation layer and the bottom layer are also shown. In addition, arrows at the sea surface ($z = 0$) indicate the subsurface extrapolated value, the subsurface extrapolated value corrected by perturbation effects, and in the specific case of the optical profile of the downward irradiance, both $E_d(0^+,\lambda)$ and $E_s(t_0,\lambda)$ values.

3.2.2.6 Products

In-water data products: Major *in water* products obtained from optical profiles are $L_u(0^-,\lambda)$, $E_d(0^-,\lambda)$, $E_u(0^-,\lambda)$ and $K_d(\lambda)$ computed from regressions in the extrapolation interval.

Above-water data products: Above-water products are $E_d(0^+,\lambda)$, the remote sensing reflectance, $R_{rs}(\lambda)$, and the normalized water-leaving radiance, $L_{wn}(\lambda)$.

3.3 The CoASTS data

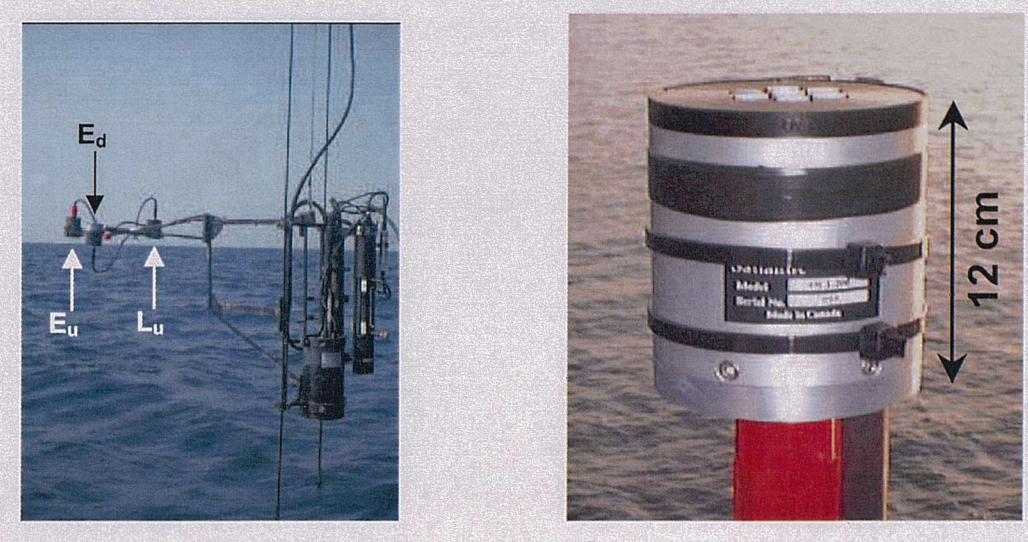
The CoASTS data set collection started in fall 1995. As the present study was carried out during 2002, only CoASTS data up to fall 2001 were here considered to ensure that all the applications were based on the same set of stations.

The CoASTS data set consists of the following *in situ* measured quantities:

- Optical profiles of upwelling radiance, $L_u(z,\lambda)$, downward irradiance, $E_d(z,\lambda)$, and upward irradiance, $E_u(z,\lambda)$, measured with the Wire Stabilized Profiling Environmental Radiometer (WiSPER, see Section 3.3.1).

- The above water irradiance, $E_s(\lambda)$, and diffuse sky irradiance, $E_i(\lambda)$, taken with a SATLANTIC Multichannel Visible Detector System.

- The direct solar irradiance $E_{sun}(\lambda)$ and the diffuse sky radiance $L_i(\theta,\phi,\lambda)$, taken at different zenith and azimuth angles with the CIMEL Electronique (Paris, France) CE-318 sun photometer.



(a) (b)

FIGURE 3.5: The WiSPER system for *in water* optical profile measurement (a) and the above water sensor for irradiance measurements used for the normalization of the radiometric profile data (b).

- Profiles of seawater beam attenuation $b(z, \lambda)$ and absorption $a(z, \lambda)$ coefficients, taken with the Western Environmental Technology Laboratories (WET-Labs), Inc. (Philomath, Oregon) AC-9.

- Ancillary data (i.e., profiles of water temperature and salinity, Secchi disk depth, tide level, atmospheric pressure, relative humidity, air temperature, wind speed and direction, cloud cover and sea state).

Laboratory analysis of water samples collected in correspondence of *in water* optical profiling, are:

- In-vivo pigmented, $a_{ph}(\lambda)$, and not pigmented, $a_{NPPM}(\lambda)$, particulate matter absorption coefficients, obtained from spectrometric analysis of particles retained on filters.

- The CDOM absorption coefficient, $a_{CDOM}(\lambda)$, obtained from spectrometric analysis of filtered seawater.

- Pigments concentration, from High Performance Liquid Chromatography (HPLC) analysis.

- The Total Suspended Matter concentration, TSM, determined with the dry weight technique.

The following sections present in more details the quantities later considered in this study for the algorithms development.

3.3.1 Remote sensing reflectance

The R_{rs} data, derived from the CoASTS data set and used in this study for the development of the ocean colour algorithms, have been computed from *in water* optical profile measurements. The radiometers for measuring $L_u(z,\lambda)$, $E_d(z,\lambda)$, and $E_u(z,\lambda)$, (with nominal center wavelengths corresponding to $\lambda = 412$, 443, 490, 510, 665 and 683 nm) are mounted on the Wire Stabilized Profiling Environmental Radiometer (WiSPER) system, shown in Figure 3.5(a). The WiSPER is installed on a metal frame which moves along two vertical wires fixed between the tower and the sea bottom. This provides a high degree of stabilization of the WiSPER instrument system and prevent almost any movement out of the vertical plane.

The WiSPER radiometers are SATLANTIC Ocean Colour Radiance and Irradiance sensors (OCR-200 and OCI-200, respectively). One OCR-200 is used to measure $L_u(z,\lambda)$, while two OCI-200 are used to measure $E_d(z,\lambda)$ and $E_u(z,\lambda)$. The light sensors are

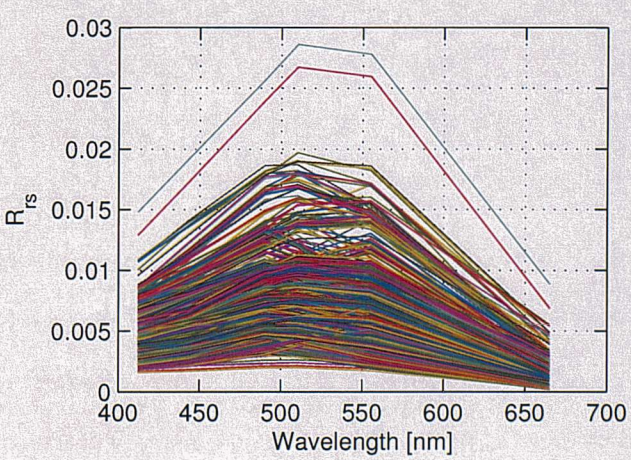


FIGURE 3.6: Spectral plot of the *in situ* remote sensing reflectance.

λ [nm]	412	443	490	510	555	665
μ	0.0043	0.0052	0.0069	0.0071	0.0062	0.0011
σ	0.0016	0.0022	0.0031	0.0033	0.0033	0.0009
min	0.0016	0.0017	0.0019	0.0020	0.0017	0.0002
max	0.0148	0.0191	0.0257	0.0286	0.0278	0.0089

TABLE 3.1: Variability of the *in situ* measured remote sensing reflectance corresponding to SeaWiFS center wavelengths (μ, σ, *min* and *max* represent the mean, standard deviation, minimum and maximum values, respectively).

mounted on an extension boom, which places them 1 m away from the main part of the frame, at approximately 7.5 m from the tower legs. The WiSPER is raised and lowered from the southeastern side of the tower by an electrical winch with a speed of approximatively 0.1 ms^{-1}.

The above water irradiance measurements, used for the processing of the *in water* optical profile data, are collected with a Satlantic Multichannel Visible Detector System (MVD, see Figure 3.5(b)).

The set of R_{rs} spectra derived from the CoASTS database is presented in Figure 3.6. The variability of the R_{rs} values is indicated in Table 3.1 for each individual wavelength.

3.3.2 Chlorophyll-*a* concentration

Phytoplankton pigment concentrations are determined using the HPLC analysis with a slightly modified Joint Global Ocean Flux Study protocol (JGOFS, 1994; Jeffrey et al., 1997). Seawater volumes, ranging from 1.0 to 2.5 L, depending on the quantity of material suspended in the seawater, are filtered immediately after the sample collection using glass fiber (GF) filters with a nominal 0.7 μm pore size. The filters are then immediately stored in liquid nitrogen for subsequent laboratory analysis. The use of the 0.7 μm pore size is justified by the diameter of living phytoplankton cells, which is generally higher than 1 μm (Stramsky and Kiefer, 1991). The measurement protocol applied for the determination of the various pigment concentration from the HPLC chromatogram is described in Zibordi et al. (2002a), while the *Chl-a* distribution for the set of CoASTS data used in this study is presented in Figure 6.1.

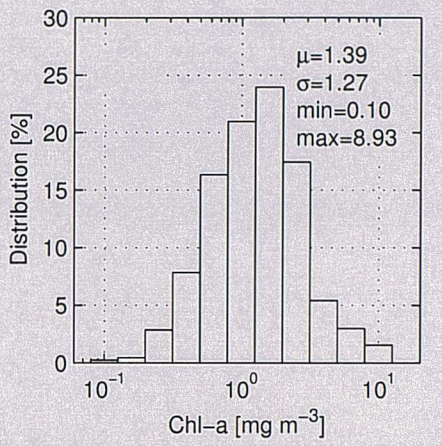


FIGURE 3.7: Distribution of the *Chl-a* within the set of CoASTS data used in the present study (μ, σ, *min* and *max* represent the mean, standard deviation, minimum and maximum values, respectively).

3.3.3 Pigmented and non pigmented particle absorption coefficients

The *in vivo* absorption coefficient, $a_p(\lambda)$, of aquatic particles retained on filters, is determined with a Perkin Elmer (Eremont, California) Lambda19 dual-beam spectrometer equipped with a 60 mm diameter integrating sphere. The deposit of the particles on filters is obtained by filtering the seawater samples on GF filters with a nominal pore size of 0.7 μm under low vacuum pressure (less than 120 mm Hg) to prevent particle breakage and pigment degradation. The filtered volume varies from 1.0 to 2.5 L, depending on the quantity of material suspended in the seawater. After filtration, the filters with the deposits are immediately placed on a Petri slide and stored in liquid nitrogen. The

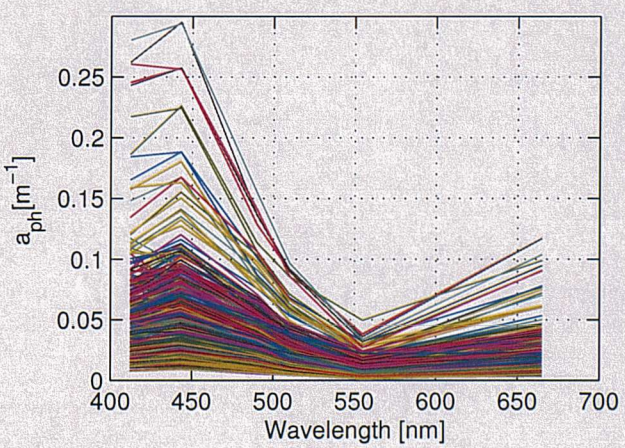


FIGURE 3.8: Spectral plot of the *in situ* pigmented particulate matter absorption coefficient.

λ [nm]	412	443	490	510	555	665
μ	0.047	0.054	0.033	0.023	0.009	0.020
σ	0.037	0.040	0.022	0.015	0.006	0.016
min	0.007	0.009	0.005	0.004	0.001	0.003
max	0.280	0.295	0.154	0.096	0.049	0.117

TABLE 3.2: The same as in Table 3.1 but for the pigmented particulate matter absorption coefficients $[m^{-1}]$.

total absorption coefficient of the equivalent particle suspension is computed according to Tassan and Ferrari (1995) in the 400-750 nm spectral range (with 1 nm resolution).

The coefficients of the the pigmented and non pigmented particulate absorption, $a_{ph}(\lambda)$ and $a_{NPPM}(\lambda)$, respectively, are obtained through bleaching the sample on the filter using a solution of sodium hypochlorite (NaClO) as an oxidizing agent. This oxidation acts rapidly on pigment molecules and slowly on detritus making possible a selective analysis of the absorption components of pigmented and non pigmented particles retained on the filter. A description of the NaClO bleaching technique, is presented in Tassan and Ferrari (1995) and in Ferrari and Tassan (1999).

Spectral plot at the wavelengths corresponding to the WiSPER nominal bands, and quantities describing their distribution, are presented in Figure 3.8 and Table 3.2 for a_{ph}, and in Figure 3.9 and Table 3.3 for a_{NPPM}.

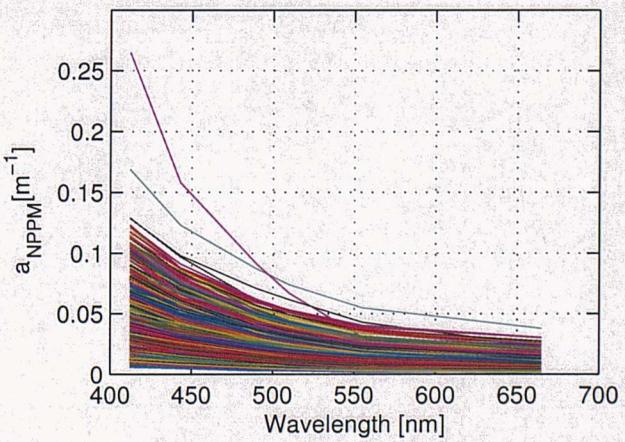


FIGURE 3.9: Spectral plot of the non pigmented particulate matter absorption coefficient.

λ [nm]	412	443	490	510	555	665
μ	0.049	0.036	0.025	0.021	0.015	0.011
σ	0.029	0.021	0.014	0.012	0.008	0.006
min	0.006	0.005	0.003	0.003	0.002	0.001
max	0.265	0.158	0.091	0.074	0.055	0.038

TABLE 3.3: The same as in Table 3.1 but for the non pigmented particulate matter absorption coefficients $[m^{-1}]$.

3.3.4 Coloured dissolved organic matter absorption coefficient

The absorption coefficient of coloured dissolved organic matter, $a_{CDOM}(\lambda)$, is determined using a dual-beam spectrometer Perkin Elmer Lambda-12. Water samples (generally from 0.7 to 1 L) are filtered through 0.22 μm cellulose filters and then refrigerated at 4°C in an amber glass bottle with the addition of a solution (0.5 mL for 100 mL of sample) of 10 $g\mathrm{L}^{-1}$ of sodium azide (NaN_3) to prevent the growth of any bacteria.

The spectrometric measurements are performed in the spectral range 350-750 nm with 1 nm resolution. They are performed by placing a 10 cm quartz cuvette containing Milli-Q water in the optical path of the reference beam, and a 10 cm quartz cuvette containing the filtered seawater sample in the optical path of the sample beam. The spectral absorption coefficient $a_{CDOM}(\lambda)$ is computed from the measured absorbance $A(\lambda)$ resulting from the difference between the sample absorbance and the reference absorbance (Ferrari et al., 1996). Spectral a_{CDOM} absorption coefficients, as resulting

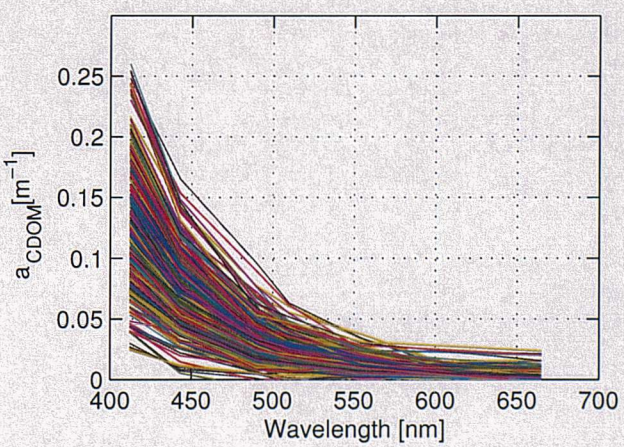


FIGURE 3.10: Spectral plot of the coloured dissolved organic matter absorption coefficient.

λ [nm]	412	443	490	510	555	665
μ	0.118	0.068	0.030	0.022	0.012	−0.000
σ	0.045	0.027	0.014	0.011	0.006	0.006
min	0.024	0.005	−0.006	−0.011	−0.004	−0.029
max	0.260	0.165	0.096	0.063	0.037	0.023

TABLE 3.4: The same as in Table 3.1 but for the yellow substance absorption coefficients $[m^{-1}]$. Negative values are only due to laboratory measurment uncertanties.

from the set of CoASTS data used in this study, are presented in Figure 3.10, while the corresponding parameters describing the distribution at specific wavelengths are presented in Table 3.4.

Chapter 4

Multi Layer Perceptron and Novelty Detection[1]

Multi Layer Perceptron (MLP) neural networks have been used in this study to retrieve various properties (i.e., concentration, absorption at a specific wavelength or absorption spectra) of the optically active seawater constituents as a function of the radiance leaving the sea surface. The basic elements of the MLP and the reasons for its suitability in developing ocean colour algorithms are presented in Section 4.1. The range of applicability of the algorithm implemented has been identified through the *novelty detection* technique (Bishop, 1994), presented in Section 4.2.

4.1 The Multi Layer Perceptron neural network

The MLP can be defined as a parameterized distributed algorithm consisting of a network of simple non-linear processing units and "may be viewed as a practical vehicle for performing a non-linear input-output mapping of general nature" (Haykin, 1998). The basic elements of the MLP, hereafter also indicated with the more general term of *network*, are presented in the following sections.

[1]Several concepts and diagrams presented in this section have been adapted from various textbooks and papers. Specifically, for what concerns the description of the MLP and the novelty detection technique, they were mostly adapted from Bishop (1994, 1995).

4.1.1 The MLP structure

The MLP consists of successive layers of units, with connections running from each unit in one layer to each unit of the next layer (Figure 4.1). The input units x_i represent, for $i = 1, \ldots, d$ (being d the dimension of the input space) the input variables, while the extra input $x_0 \equiv 1$ is added for the *bias* parameters (see below).

The hidden units compute a linear combination of the corresponding inputs through the network parameters, called *weight* and indicated with w, according to

$$a_j = \sum_{i=0}^{d} w_{ji}^{(1)} x_i, \tag{4.1}$$

where $w_{ji}^{(1)}$, for $i = 1, \ldots, d$, represents the weight linking the input unit i to the hidden unit j, $w_{j0}^{(1)}$ is the bias adaptive parameter of the hidden unit j and the superscript (1) is used to indicate the first layer of weights. The *activation* of the hidden unit j, identified with z_j and representing the input for the next layer, is then obtained by applying an

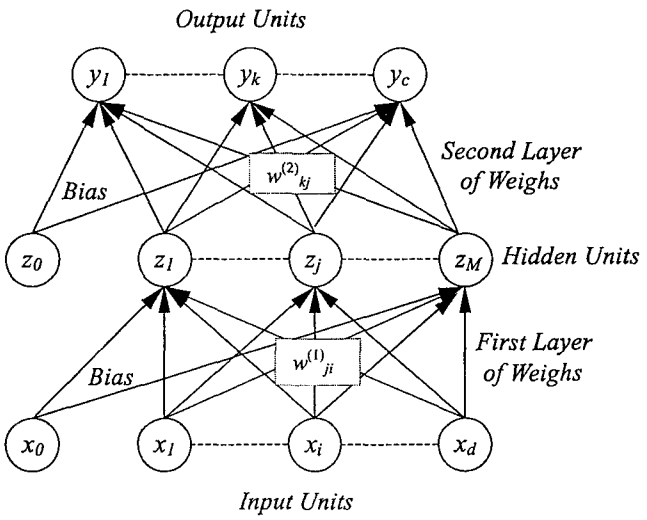


FIGURE 4.1: Example of MLP neural network having two layer of adaptive parameters. The bias parameters in the first layer are shown as weights from an extra input having fixed value of $x_0 = 1$. Similarly, the bias parameter in the second layer are shown as weights from an extra hidden unit, with activation again fixed at $z_0 = 1$.

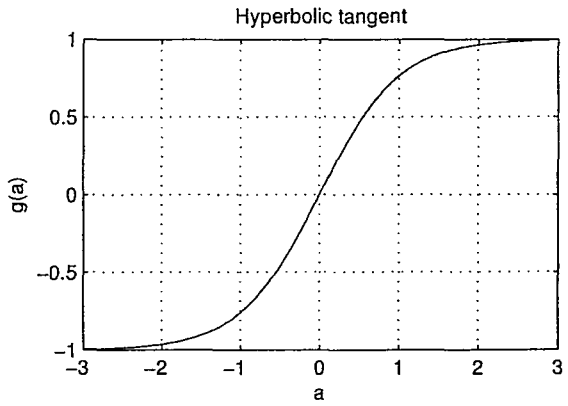


FIGURE 4.2: The hyperbolic tangent activation function.

activation function, $g(\cdot)$, to the summation of Equation 4.1

$$z_j = g\left(a_j\right). \tag{4.2}$$

Examples of activation functions are the *hyperbolic tangent* (Equation 4.3a, shown in Figure 4.2 and used in the present study) or the *logistic function* (Equation 4.3b)

$$g(a) = \frac{e^a - a^{-a}}{e^a + e^{-a}} \tag{4.3a}$$

$$g(a) = \frac{1}{1 + e^{-a}}. \tag{4.3b}$$

The MLP output y_k (where $k = 1, \ldots, c$, with c the dimension of the output space) is then obtained by applying a second activation function, $\widetilde{g}(\cdot)$, to the weighed combination through the second layer of weights $w_{kj}^{(2)}$ of the activations of the hidden units z_j

$$y_k = \widetilde{g}\left(\sum_{j=0}^{M} w_{kj}^{(2)} z_j\right), \tag{4.4}$$

where M indicates the number of hidden units and, as for the input layer, the hidden units $z_0 \equiv 1$ is added for the bias coefficients.

The activation function of the second layer depends on the network application. Networks for classification use the *logistic* (Equation 4.3b) activation function or, for outputs mutually exclusive, the *softmax* function (Nabney, 2001). Regressional networks, as those used in the present study, are simply based on the *identity* function.

The equation of the MLP used in the present work is then

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} g\left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i\right), \tag{4.5}$$

being g the hyperbolic tangent function. It is important to notice that this function satisfies the requirements prescribed by the *universal approximation theorem* (see Haykin, 1998). Provided that the number of hidden units is sufficiently high, Equation 4.5 can represent any continuous input-output mapping with arbitrary accuracy.

4.1.2 The error function

The objective of the MLP training is to create a statistical model of the underlying generator of the data. This objective is realized by means of a *training data set* representative of the relationship to be modelled (i.e., *supervised learning*). The training data set is here indicated with $(\mathbf{x}^n; \mathbf{t}^n)$ for $n = 1, \ldots, N$, where n and N are the subscription of the teaching input-output pair and the training data size, respectively. The bold letters $\mathbf{x}$ and $\mathbf{t}$ are used to indicate, in an array notation, the input and the target data, respectively. For the cases study here considered, the input variables are represented by the R_{rs} at different wavelengths, while the targets correspond to the property of the water components to be modelled.

The most general statistical descriptor of the process that generated the training data set is represented by the probability density $p(\mathbf{x}, \mathbf{t})$ in the input-target space. This joint probability can be decomposed in

$$p(\mathbf{x}, \mathbf{t}) = p(\mathbf{t}|\mathbf{x})\, p(\mathbf{x}), \tag{4.6}$$

where $p(\mathbf{t}|\mathbf{x})$ indicates the conditioned probability density of the target $\mathbf{t}$ on the input is $\mathbf{x}$, and $p(\mathbf{x})$ represents the unconditioned probability density of $\mathbf{x}$.

The probability density of having the entire training data set is given by the *likelihood* ($\mathcal{L}$) function

$$\begin{aligned} \mathcal{L} &= \prod_n p(\mathbf{x}^n, \mathbf{t}^n) \\ &= \prod_n p(\mathbf{t}^n|\mathbf{x}^n)\, p(\mathbf{x}^n)\,. \end{aligned} \tag{4.7}$$

Finally, the error $\mathcal{E}$ (i.e., the quantity to be minimized during the learning process by identifying the proper set of network adaptive parameters) is defined as the as the negative of the logarithm of the likelihood

$$\mathcal{E} = -\ln \mathcal{L} = -\sum_n \ln p(\mathbf{t}^n|\mathbf{x}^n) - \sum_n \ln p(\mathbf{x}^n)\,. \tag{4.8}$$

Algorithms for functional regression and classification aim at representing only the density probability of the output conditioned by the input. Thus, the second term of Equation 4.8 does not play any role in the learning process, and can thus be neglected. This reduces the error function to

$$\mathcal{E} = -\sum_n \ln p(\mathbf{t}^n|\mathbf{x}^n)\,, \tag{4.9}$$

and the statistical dependence of the target from input to

$$p(\mathbf{t}|\mathbf{x}) = \prod_{k=1}^{c} p(t_k|\mathbf{x})\,. \tag{4.10}$$

The explicit expression of the error $\mathcal{E}$ depends on the characteristics of the training data set. Here, it is assumed that each target output, t_k, is the result of a deterministic function, $h_k(\mathbf{x})$, plus the additive noise, ϵ_k, normally distributed with zero mean and

standard deviation equal to σ (Equation 4.11a and 4.11b, respectively)

$$t_k = h_k(\mathbf{x}) + \epsilon_k \tag{4.11a}$$

$$p(\epsilon_k) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\epsilon_k^2}{2\sigma^2}\right). \tag{4.11b}$$

The objective of the MLP is then to represent the function h_k. Substituting Equation 4.5 into Equation 4.11a, and combining the latter with Equation 4.11b, it follows that the conditional dependence of the output from the input can be expressed as

$$p(t_k|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\{y_k(\mathbf{x};\mathbf{w}) - t_k\}^2}{2\sigma^2}\right). \tag{4.12}$$

The expression of the error function can be rewritten, substituting Equation 4.12 into Equation 4.9, as

$$\mathcal{E} = \frac{1}{2\pi\sigma^2} \sum_{n=1}^{N} \sum_{k=1}^{c} \{y_k(\mathbf{x}^n;\mathbf{w}) - t_k^n\}^2 + Nc\ln\sigma + \frac{Nc}{2}\ln(2\pi). \tag{4.13}$$

Finally, removing from the latter all the terms not depending on the model adaptive parameters, $\mathbf{w}$ (here indicated in array notation), the quantity to be minimized during the learning process is the sum-of-squares errors

$$\mathcal{E} = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{c} \{y_k(\mathbf{x}^n;\mathbf{w}) - t_k^n\}^2. \tag{4.14}$$

4.1.3 Interpretation of the network outputs

According to the principle of the *maximum likelihood*, the objective of the algorithm learning is to determine the set of adaptive parameters that minimizes the sum of squares errors of Equation 4.14. In order to have a statistical interpretation of the network outputs, lets consider an hypothetical infinite training data set (i.e., $N \to \infty$). The

error $\mathcal{E}$ can then be expressed as

$$\begin{aligned}\mathcal{E} &= \lim_{N\to\infty}\frac{1}{2N}\sum_{n=1}^{N}\sum_{k=1}^{c}\{y_k(\mathbf{x}^n;\mathbf{w})-t_k^n\}^2 \\ &= \frac{1}{2N}\sum_{k=1}^{c}\int\int\{y_j(\mathbf{x};\mathbf{w})-t_k\}^2 p(t_k,\mathbf{x})dt_k d\mathbf{x}. \end{aligned} \tag{4.15}$$

Decomposing the joint probability into the products of the unconditional probability of the input and the conditioned probability of the output, Equation 4.15 becomes

$$\mathcal{E} = \frac{1}{2N}\sum_{k=1}^{c}\int\int\{y_j(\mathbf{x};\mathbf{w})-t_k\}^2 p(t_k|\mathbf{x})p(\mathbf{x})dt_k d\mathbf{x}. \tag{4.16}$$

Considering in Equation 4.16 the integration over t_k, and then substituting into the resulting equation the expressions (Equation 4.17a and 4.17b) of the conditional averages of the target data

$$\langle t_k|\mathbf{x}\rangle = \int t_k p\left(t_k|\mathbf{x}\right) dt_k \tag{4.17a}$$

$$\langle t_k^2|\mathbf{x}\rangle = \int t_k^2 p\left(t_k|\mathbf{x}\right) dt_k, \tag{4.17b}$$

it follows that the error can be written as

$$\begin{aligned}\mathcal{E} &= \frac{1}{2}\sum_{k=1}^{c}\int\{y_k(\mathbf{x};\mathbf{w})-\langle t_k|\mathbf{x}\rangle\}^2 p(\mathbf{x})d\mathbf{x} \\ &\quad +\frac{1}{2}\sum_{k=1}^{c}\int\{\langle t_k^2|\mathbf{x}\rangle-\langle t_k|\mathbf{x}\rangle^2\}p(\mathbf{x})d\mathbf{x}. \end{aligned} \tag{4.18}$$

Thus, the error can be decomposed into the sum of two terms, of which the first only depends on the adaptive parameters of the algorithm. Taking that the second term is non-negative, the Equation 4.18 shows the important result that **the error is minimized when the algorithm output corresponds to the mean value of the target data conditioned on the input**, being y_k the regression of $\langle t_k|\mathbf{x}\rangle$ (see Figure 4.3). This result was derived on the assumption of an infinite training data set and

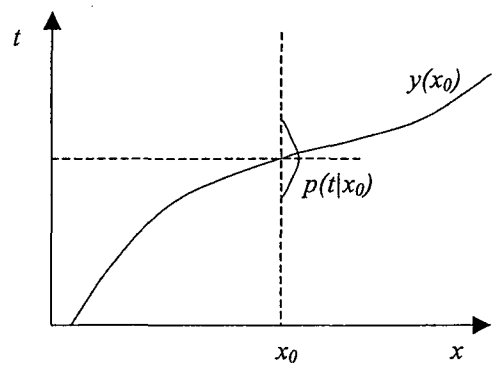


FIGURE 4.3: Schematic representation that the network resulting from the minimization of error function of Equation 4.18 corresponds to the regression of the average target data.

depends on the MLP capability of representing any continuous input-output mapping. When the size of the training data set is limited (as for the experimental data considered in the present study), it is not possible to represent the conditioned density of the target data with arbitrary accuracy and the network complexity has to be defined optimizing the algorithm *generalization capability*, as discussed in Section 4.1.5. It is important to highlight that Equation 4.18 also represents the basis of the *novelty detection* technique (see Section 4.2), adopted in the present study to define the range of applicability of the algorithms.

4.1.4 Optimization algorithms

The network learning consists in the minimization of the error function of Equation 4.14 with respect to the adaptive parameters of the network. Because the error is a non linear function of the adaptive parameters, this minimization can not be done analytically. The minimization of the error function is then performed through an iterative process

$$\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} + \Delta\mathbf{w}^{\tau}, \tag{4.19}$$

where, in the array notation, $\mathbf{w}$ indicates the set of adaptive parameters of the network and τ represents the step number. The initial set of adaptive parameters ($\mathbf{w}^{\tau=0}$) is chosen randomly from a predefined distribution (in this study represented by a multivariate

normal distribution with zero mean and unit standard deviation). The final vector represents the set of weights to be used for the operational application of the network.

In order to compute the iteration step $\Delta \mathbf{w}^\tau$, most of the learning algorithms use the gradient of the error $\nabla_{\mathbf{w}}\mathcal{E}$. The advantage of using $\nabla_{\mathbf{w}}\mathcal{E}$ relies on the fact that this quantity can be efficiently computed through the *back propagation* method (see Bishop, 1995). The simplest learning algorithm based on the gradient of the error function is the *gradient descendent*. This method prescribes to set the increment to $-\eta\nabla_{\mathbf{w}}\mathcal{E}^\tau$, where $\nabla_{\mathbf{w}}\mathcal{E}^\tau$ is the gradient of the error function at the iteration τ, and the coefficient η is called *learning rate*. The gradient descendent is one of the firsts and most common learning algorithms, but presents the disadvantage of generally requiring a large number of iterations to converge to a steady point of the error function.

More recent and effective learning algorithms were proposed, and some of the most effective ones compute the increment step according to

$$\Delta \mathbf{w}^\tau = \lambda^\tau \mathbf{d}^\tau, \tag{4.20}$$

where $\mathbf{d}$ is called *searching direction* and λ^τ is chosen to minimize the following error function (Equation 4.21) that, once $\mathbf{d}$ is defined, only depends on the parameter λ^τ

$$\mathcal{E}(\lambda) = \mathcal{E}\left(\mathbf{w}^\tau + \lambda^\tau \mathbf{d}^\tau\right). \tag{4.21}$$

There are different versions for these learning algorithms, according to the method to compute λ and $\mathbf{d}$. Among the optimization algorithms based on this approach, the *scaled conjugate gradients* (SCG) is that used in the present study. The SCG is characterized by the fact that each iteration step in computationally more demanding than other methods, but it requires much less iteration steps.

4.1.5 MLP generalization capability and performance assessment

The objective of the network implementation is to give the most effective prediction when new inputs are presented (*generalization capability*). In order to optimize the network generalization capability, it is necessary to take into account that the network adaptive parameters are defined through a minimization process based on a finite data set of teaching samples. If the data are noisy, as the case for the experimental data used in this study, the best generalization will not be obtained by the function perfectly fitting the training data but by the network capturing the underlying function from which the data were generated. Thus, instead of using more flexible networks with the scope of highly reducing the training error, it is necessary to control the network complexity to optimize the network generalization capability. The methodologies adopted in the present study to improve the network generalization capability are based on the *model selection* (Section 4.1.5.1) and on the use of a *weigh regularization term* (Section 4.1.5.2). The aforementioned considerations also indicate that an estimate of the network performance needs to be based on a set of input-output pairs (i.e., *validation data set*) different from that used for the training. To better exploit all of the available data, the network validation methodology adopted in the present study is the *cross-validation* (Section 4.1.5.3).

4.1.5.1 Model selection

The network generalization capability depends on the number of hidden units (here after also indicated as *network architecture*). If the architecture is too simple, the network may not have enough degrees of freedom to properly represent the function that generated the training data. On the other hand, if it is too complex, the network is capable to also fit noise and outliers.

The methodology adopted in the present study to identify of the most suitable network architecture consisted first of assessing the performance of the network made by a single

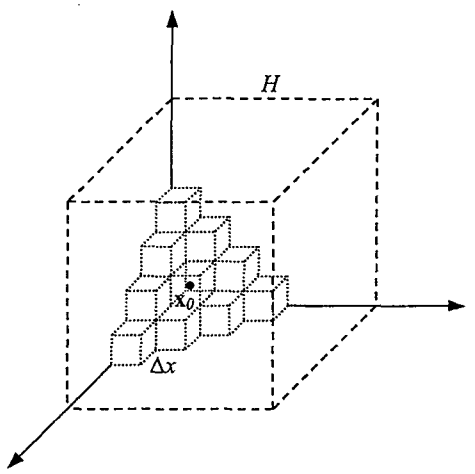


FIGURE 4.4: Schematic representation of the problem of the curse of dimensionality. The number of cells of side grows as the exponential of the input space dimensions (adapted from Bishop, 1995).

hidden unit and then adding new hidden units until the network performance was not showing any further appreciable improvement.

Together with the architecture, the network has also to be optimized with respect to the set of the input quantities (i.e., here indicated as *network configuration*). The use of a lower dimensional input space represents a way to address the problem of the *curse of dimensionality* (Bishop, 1995). This can be introduced as follows: suppose that the input data ($\mathbf{x}$) of training set (see Section 4.1.2) are uniformly distributed over an hypercube of side H (Figure 4.4). The quantity $\mathbf{x}_0$ indicates a generic point to be classified by the network and Δx is the side of a d-dimensional neighborhood centered in $\mathbf{x}_0$ (Figure 4.4). The expected number of teaching examples $\widetilde{N}$, filling the cell centered in $\mathbf{x}_0$, corresponds to the ratio between the total number of teaching examples (N) and the number of cells ($(H/\Delta x)^d$)

$$\widetilde{N} = \frac{N}{(H/\Delta x)^d}. \tag{4.22}$$

The teaching sampling density thus grows as the inverse of the exponential of the input space dimension and this penalizing behavior is named the curse of dimensionality. In practice, the former constraint is not so effective as in the example as i) the teaching

samples are not uniformly distributed but more concentrated in specific regions of the input space, and ii) the network approximates the unknown function using contextual, and not just local, information. Nevertheless, the reduction of the dimensions of the input space can significantly increase the density of the training data and thus also the capability of the network to more effectively represent the underlying generator of the data. An intuitive way to see the advantage of not using further input dimensions other than those ensuring the network performance, is given by the principle of the Occam razor[2].

The most suitable set of input quantities has been identified in the present study as follows. First, the performance of the network based on a minimal set of input quantities is assessed. Then, the performance of the former network is compared with that resulting from the network based on one more input quantity. The set of inputs producing the best network performance are retained and compared, in the next step, with the network using one more input quantity. This procedure is repeated until no significant improvement could be observed. This simple way of proceeding is preferred in respect to other approaches commonly applied for the reduction of the dimension of the input space (i.e., such as the Principal Component Analysis (PCA) and Independent Component Analysis (ICA)) because i) it ensures the use on the prior information on the known relevance of some specific inputs and also ii) it permits to preserve the information content associated to each individual input. Other advanced statistical approaches for the automatic determination of the relevance of the different input quantities (such as the ARD method, MacKay (1994)) were not used, being more effective for high dimensional input space and when the input quantities are characterized by a relative different importance more significant than that observed in the present case study.

[2] "No more things should be presumed to exist than are absolutely necessary." Principle attributed to W. Occam, 1280-1349

4.1.5.2 Weigh decay factor

Instead of limiting the number of the adaptive parameters, the smooth properties of the network can also be enhanced using the *regularization* technique. This technique prescribes to add to the error function (Equation 4.14) a *penalty term* Ω

$$\widetilde{\mathcal{E}} = \mathcal{E} + \nu\Omega, \tag{4.23}$$

where ν is called *regularization coefficient.* The network adaptive parameters are then determined by minimizing $\widetilde{E}$, while the value of ν controls the degree of influence of the penalty term Ω. In the present study, the penalty term applied to the error function is the *weight decay* (Equation 4.24), consisting of the sum of the squares of the adaptive parameters

$$\Omega = \sum_i w_i^2, \tag{4.24}$$

where the summation is over all the weights of the network.

The functioning of the weight decay depends on the fact that when the value of the model adaptive parameters are close to zero, the network outputs become an approximatively linear function of the input. In fact, the hyperbolic tangent is approximatively linear for a small value (i.e., close to zero) of its argument (Figure 4.2). Thus, the value of the regularization coefficient controls the effective complexity of the network. For a large value of ν the network is forced to be highly smoothed, while for a small value the network can overfit. In the present study the regularization coefficient has also been used to control the variability of the network performance for subsequent training-validation exercises of the same MLP configuration. This has been done testing different values of the regularization coefficient, and then keeping that allowing a more effective identification of the proper number of network adaptive parameters.

4.1.5.3 Cross-validation

The objective of the network implementation is to represent the underling generator of the data and not to represent the training data itself. For this reason, the accuracy of the network products has to be assessed with respect to a validation data set different from that used to define the adaptive parameters of the network. The simplest way to do this is to separate the available data into two parts, one to be used for the network training and the other for the validation. The disadvantage of this approach is that it significantly reduces the size of the data set used to determine the adaptive parameters of the network, and this may become a significant limitation when the availability of the teaching sample is limited as for the experimental data used in the present study.

In order to better exploit the set of experimental measurements, the method adopted to evaluate the algorithm performance is the *cross-validation* procedure. The cross-validation prescribes to partition the data set into different segments. One of the segments is then used for the validation of the algorithm, while the ensemble of the remaining data is used to train the model. Successively, a different segment is used for the validation and the network is trained with all the remaining segments. The procedure terminates when the network is applied to all the available data. The cross validation has been here applied dividing the training data set into three different segments (See Figure 4.5). Additional information on the criteria adopted for the separation of the experimental data set in the different segments, and on the implementation of the fi-

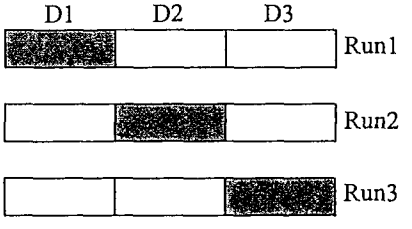


FIGURE 4.5: Schematic illustration of the partitioning of the data set into 3 segments (D1, ...,D3) for the cross-validation. The MLP is trained (run) 3 times, each time using a sub-set of the data set in which one of the segments (shown shadowed) is omitted. Each trained network is then tested on the data from the segment that was omitted during the training (adapted froms Bishop, 1995).

nal network for the operational processing of remotely sensed data, are described in Chapter 5, specifically presenting the algorithms developed in this study.

4.2 The novelty detection technique

One of the key aspects to be considered in applying neural network algorithms is that the reliability on the output depends on how the input data resemble the data used to define the network adaptive parameters. This consideration can be applied to distinguish between i) input data similar to the training data and for which the algorithms performance is expected to be close to that assessed by the validation process and ii) *novel data* (i.e, data substantially different from those used for the network training) for which the network is likely to produce less accurate results. The principle of novelty detection method (Bishop, 1994) is hereafter presented.

The definition of the adaptive network parameters depends on the minimization of the error of Equation 4.14, and given for the case of an hypothetical infinite data set by Equation 4.18. The latter is composed of two terms: the first related to the difference between computed and target (i.e., measured) quantities and the second depending on the distribution of the target data conditioned on the value of the input. The latter represents the intrinsic variability of the experimental measurements and does not play any role in the process of the algorithms training. Thus, it will not hereafter be considered. The first term, here rewritten for clarity

$$\mathcal{E} = \sum_{k=1}^{c} \int [y_k(\mathbf{x};\mathbf{w}) - \langle t_k|\mathbf{x}\rangle]^2 \, p(\mathbf{x}) d\mathbf{x}, \tag{4.25}$$

indicates that the difference (used in defining the adaptive parameters of the networks) between computed, $y_k(\mathbf{x};\mathbf{w})$, and corresponding regressional, target value $\langle t_j|\mathbf{x}\rangle$, is weighted on the probability density of the input quantity, $p(\mathbf{x})$. More precisely, in Equation 4.25 the probability density of the input data, $p(\mathbf{x})$, plays the role of accuracy

(i.e., $\sigma_y(\mathbf{x})$) of the computed values. This suggests that the $p(\mathbf{x})$ provides an appropriate quantity to associate the novelty of the inputs with the accuracy of the outputs.

In the present study, the range of applicability of the algorithms was identified as follows. First, the $p(\mathbf{x})$ of the training data has been estimated through a multivariate normal distribution

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{4.26}$$

Input data for which the $p(\mathbf{x})$ was lower than a threshold value were considered novel and thus outside of the range of applicability of the algorithm. It has anyway to be considered that this approach is more significant in the case that accuracy network outputs actually varies as $1/\sqrt{p(\mathbf{x})}$. In the case study here investigated, this condition is only in part fulfilled due to the relevance of specific combinations of water components, beside of the density of the input data, in determining the algorithm accuracy (see Chapters 6 and 7).

Chapter 5

Retrieval of the Optically Significant Seawater Components

This chapter focuses on the development and the performance assessment of the MLP algorithms for the retrieval of seawater components in optically complex conditions, and is organized into two papers submitted to *IEEE Transactions in Geoscience and Remote Sensing*. The references of these works are included in the thesis reference list for completeness.

The first paper concerns the retrieval of the phytoplankton seawater component. The accuracy of the MLP algorithms - developed using the CoASTS data set - in retrieving the *chlorophyll-a* concentration, was compared with the accuracy of the classical band ratio algorithm (also developed with CoASTS data) and with the accuracy of the OC4v4 algorithm (O'Reilly et al., 2000), routinely used for the operational processing of SeaWiFS data. Furthermore, the accuracy in retrieving the absorption of the pigmented particulate matter at 443 nm instead of the concentration of *chlorophyll-a*, was also investigated. The evaluation of the algorithms accuracy was made with respect to i) the entire CoASTS data set, ii) separately considering Case 1 and Case 2 water conditions and iii) additionally applying the algorithms to SeaWiFS derived R_{rs} data matching with *in situ* measurements.

The second paper is addressed to the retrieval of the coloured dissolved organic matter (CDOM) and the non pigmented particulate matter (NPPM) absorption coefficients. A new methodology is applied for the retrieval of these two components. This method prescribes i) the definition of the regressional parameters of the exponential functions interpolating the experimental data at different wavelengths, and then ii) the implementation of the MLP algorithms for the retrieval of the regressional parameters from the R_{rs}. The advantage of this approach with respect to previous studies is the determination of the absorption spectra and not just the absorption coefficients at some specific wavelength. The validation of the MLP algorithm assessed the accuracy in retrieving both the regressional parameters, as well as the absorption coefficients derived from the regressional parameters.

Phytoplankton Determination in an Optically Complex Coastal Region Using a Multi Layer Perceptron Neural Network[1]

Abstract

The determination of phytoplankton in seawater, quantified as *chlorophyll-a* concentration (*Chl-a*) or absorption of pigmented matter (a_{ph}), is a major objective of optical remote sensing. The accuracy of Multi Layer Perceptron (MLP) neural network algorithms in determining *Chl-a* and a_{ph} at 443 nm as a function of the multispectral remote sensing reflectance (R_{rs}), was investigated for optically complex waters. The implementation of the MLP algorithms was carried out relying on an experimental data set collected in a coastal region of the northern Adriatic Sea. The performance of the algorithms was assessed on both separate and combined Case 1 and Case 2 water types. The proposed MLP algorithms showed a better accuracy both with respect to other algorithms developed on the basis of the same data set as well as respect to independent algorithms operationally used for the processing of SeaWiFS (Sea-viewing Wide Field-of-view Sensors) data. The study also showed an high accuracy in determining $a_{ph}(443)$, and thus further confirmed the possibility of computing the inherent optical properties of seawater significant components from the R_{rs} spectra.

Index Terms

Neural network, bio-optical modelling, ocean colour.

[1]The content of this Section was submitted to *IEEE Transactions in Geoscience and Remote Sensing* by Davide D'Alimonte and Giuseppe Zibordi, in January 2003 and accepted for publication. The core theme of the paper is the retrieval of the phytoplankton component from R_{rs} data using MLP algorithms. The ideas, development and writing up of the paper were principal responsibility of the thesis author. The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into a based team research.

I. INTRODUCTION

BAND ratio algorithms currently used for the determination of chlorophyll-a concentration (*Chl-a*) from remote sensing data perform better in open ocean waters (generally identified as Case 1 (Morel and Prieur, 1977)) than in coastal waters (here identified as those permanently or occasionally considered as Case 2 (Morel and Prieur, 1977) because of bottom resuspension, coastal erosion, river inputs, or by simple anthropogenic impact). This different accuracy in retrieving the *Chl-a* is mostly due to the increased independent variability of the optically significant water components in Case 2 waters, while the validity of the band ratio approach relies on the correlation between phytoplankton and the other optically significant substances (i.e., the coloured dissolved organic matter and the non pigmented particulate matter).

In addition, the relationship between *Chl-a* and R_{rs} ratio varies not only from open ocean to coastal waters but also between different coastal areas. The ideal data set used for the implementation of the empirical algorithm should then represent all these specific regional conditions. Nevertheless, the inclusion within a single data set of too many different regional bio-optical conditions easily lead to a lower intrinsic accuracy of the band ratio algorithm, despite of its more general applicability.

As an attempt to overcome the limitations of the band ratio approach, different investigators used the Multi Layer Perceptron (MLP) neural network to produce ocean colour algorithms. Attractive properties of the MLP are the capability of i) representing complex input-output relationships and ii) efficiently handling a multi-dimensional input space represented by the R_{rs} at different wavelengths. An MLP for the retrieval of *Chl-a* in Case 1 waters was developed (Gross et al., 2000) on the basis of a synthetic data set created with an existing bio-optical model (Morel, 1988). A second study on the MLP performance in Case 1 waters (Keiner and

Brown, 1999) was made on the basis of the SeaBAM data set (O'Reilly et al., 1998), composed of 919 samples. These MLP algorithms, validated with *in situ* measurements, showed a better performance than the OC2 band ratio algorithm (O'Reilly et al., 1998) also based on the SeaBAM data set. Different authors, (Buckton et al., 1999; Shiller and Doerffer, 1997) with synthetic data only, showed the possibility of using the MLP to retrieve the phytoplankton pigment concentration, the suspended matter concentration and the the coloured dissolved organic matter in both Case 1 and Case 2 waters. Analogous results were obtained (Cipollini et al., 2001) with the *GRBF* (Generalized Radial Basis Function) neural network.

The aim of this work is to investigate the accuracy in retrieving the *Chl-a* and the absorption of the pigmented particulate matter at 443 nm ($a_{ph}(443)$) from R_{rs} in optically complex coastal waters. It is pointed out that the determination of the absorption $a_{ph}(443)$, as an alternative to *Chl-a*, is relevant for the modelling of marine bio-optical processes (Lee et al., 1996; Morel, 1991). The experimental data set used in this study was collected within the framework of the Coastal Atmosphere and Sea Time-Series (CoASTS) program carried out in the northern Adriatic Sea at the Acqua Alta Oceanographic Tower (AAOT) (Berthon et al., 2002; Zibordi et al., 2002). The sampling site ensures the presence of both Case 1 and Case 2 conditions (Berthon et al., 2002), making the CoASTS data set particularly appropriate for this analysis.

The performance of the MLP algorithms is here compared with i) band ratio algorithms developed on the basis of the CoASTS data set, and ii) the OC4v4 algorithm (O'Reilly et al., 2000) (based on multiple band ratios) developed on the basis of the SeaBAM data set and routinely used for the operational processing of the SeaWiFS (Sea-viewing Wide-Field-of view Sensor, Hooker and Esaias, 1993) data. The comparison with the OC4v4 is made more signifi cant by the fact that this algorithm has been developed with a data set containing a small subset of the CoASTS data. Thus, this analysis will also give some insight on the relevance of

including some points from an optically complex coastal region in a large data set to develop a more general algorithm.

The comparison of the accuracy of the different algorithms here analyzed, was made both on the basis of the CoASTS data set as a whole, as well as considering separately Case 1 and Case 2 data. In addition, R_{rs} derived from a set of SeaWiFS data coincident with *in situ* measurements was used to qualitatively discuss the response of different algorithms on actual remote sensing data.

II. Data Sets

A. The in situ data set

The measurements used in this study were collected within the framework of the Coastal Atmosphere and Sea Time-Series (CoASTS) program that supports ocean colour calibration and validation activities through a comprehensive data collection at the Acqua Alta Oceanographic Tower (AAOT) in the northern Adriatic Sea. The measurement site, located 8 nautical miles southeast off the Venice Lagoon (latitude 45°19′N, longitude 12°30′E), exhibits waters type close to the separation threshold suggested by Loisel and Morel (1998) for Case 1 and Case 2 waters (with roughly one third of the CoASTS measurement stations pertaining to Case 2, see Figure 1). Data used in this study were collected in the period from October 1995 up to December 2001.

In-water measurements of upwelling radiance, upward and downward irradiance from the Wire Stabilized Profi ling Environmental Radiometer (WiSPER) system were used to derive the R_{rs} (D'Alimonte and Zibordi, 2001) at the center wavelengths corresponding to the SeaWiFS visible channels (412, 443, 490, 510, 555, and 670 nm). *Chl-a* was obtained through High Performance Liquid Chromatography (HPLC) (Targa et al., 2000). The concentrations relevant to this study are in the range 0.1 to 9 mg m^{-3}, with mean and standard deviation equal to

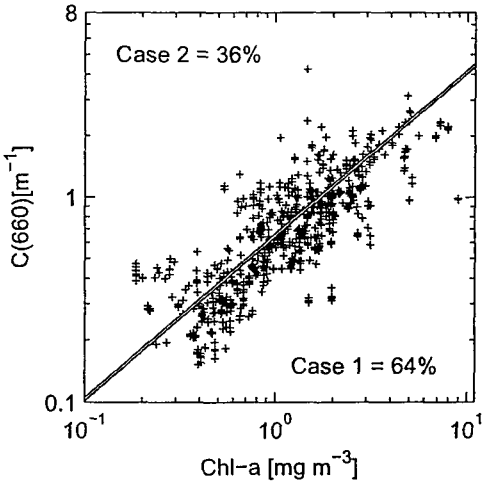


Fig. 1. Scatter plot of the beam attenuation coeffi cient C at 660 nm versus the *chlorophyll-a* concentration (*Chl-a*) for the data included in the CoASTS data set (October 1995 - December 2001). The line separates Case 1 from Case 2 waters conditions (Loisel and Morel, 1998).

1.4 and 1.3 mg m^{-3}, respectively.

The pigmented particulate matter absorption was measured through a spectrometric technique (Tassan and Ferrari, 1995) applied to the deposit obtained from the sea water sample fi ltration on fi ber glass fi lters with 0.7 μm pore size. The bleaching technique (Tassan and Ferrari, 1995) was used to separate the contribution of the pigmented and non pigmented fractions. The $a_{ph}(443)$ measurements relevant to this study are in the range 0.01 to 0.3 m^{-1}, with mean and standard deviation equal to 0.04 and 0.05 m^{-1}, respectively.

In order to ensure the highest independence between validation and training data, *in situ* measurements corresponding to satellite match-ups (31 SeaWiFS scenes) were not used for the algorithms development. The data set supporting the study was composed of 456 *Chl-a* and $a_{ph}(443)$ measurements. Each measurement produced from the analysis of water samples, was associated to one or more successive reflectance spectra measurements. The total number of pairs of *Chl-a* (or $a_{ph}(443)$) and corresponding R_{rs} spectra, is 906.

B. The SeaWiFS data set

A set of 31 SeaWiFS derived R_{rs} coincident with *in situ Chl-a* measurements was used to evaluate the algorithms performance with respect to actual remote sensing data. The R_{rs} is obtained by applying an approximate atmospheric correction scheme (Sturm and Zibordi, 2002) whose accuracy was theoretically (Bulgarelli and Zibordi, 2003) and experimentally (Mélin et al., 2002) assessed.

III. METHODS

This section presents the methodologies used for the development and the assessment of the MLP algorithms. For completeness, the band ratio algorithms later on compared with the MLP, are also presented.

A. MLP algorithms

MLP can be defined as a parameterized distributed algorithm consisting of a network of simple non-linear processing units called neurons and "may be viewed as a practical vehicle for performing a non-linear input-output mapping of general nature" (Haykin, 1998). The general principles of the MLP functioning, as well as the specific methods applied below, are presented in several textbooks (Bishop, 1995; Cherkassky and Muller, 1998). The MLP algorithms to derive the phytoplankton component from R_{rs} have been implemented by means of the Netlab toolbox for MATLAB (Nabney, 2001). The adopted MLP consisted of one layer of hidden neurons, a sigmoid activation function for the hidden layer and a linear activation function for the output layer. In fact, this configuration can represent any continuous input-output mapping (Haykin, 1998). The Scaled Conjugate Gradient (SCG) (Bishop, 1995) was used for the optimization of the algorithm parameters, while the number of training cycles was 100. Hereafter the discussion is focused on the methodologies used for the development, the performance assessment and the generalization improvement of the MLP.

1) Data pre-processing: The data pre-processing consisted in standardizing (i.e., subtracting their mean value and dividing by their standard deviation) the logarithm of both the R_{rs} (at each wavelength separately) as well as the *Chl-a* or the $a_{ph}(443)$, by assuming that the considered bio-optical quantities are lognormally distributed (Campbell, 1995).

2) Multiple training to identify the proper set of MLP parameters: The algorithm performance depends on the initial conditions of the weight values due to local minima of the error function. Thus, the simple strategy of training and testing three times the same MLP architecture was adopted to identify non effective learning processes.

3) Cross-validation: The objective of the algorithm development is to identify the general relationship that relates the input and output data. Thus, the algorithm validation has to assess the model generalization capability and not its capability of fitting the training points. This is particularly relevant for MLP, where the number of free parameters makes the *over fitting* a risk. The simplest way to assess the algorithm performance with points different from those used for its implementation is to divide the data set in two parts, one for the training and the other for the validation. This straight forward procedure is only applicable if the size of the data set is large with respect to the number of the algorithm parameters. When the data availability is limited, as for the *in situ* measurements here used, some alternative algorithm validation techniques should be used to better exploit all the available data. The *cross-validation* (Cherkassky and Muller, 1998) is the technique applied in this study.

The cross-validation requires to separate the data set into different segments: all except one are used for the algorithm development, while that discarded is used for the validation. The procedure is reiterated, each time using a different segment for the validation, until the algorithm performance has been evaluated over the entire data set. The main limitation of this method is its computational demand because of the several trainings required for the same MLP architecture.

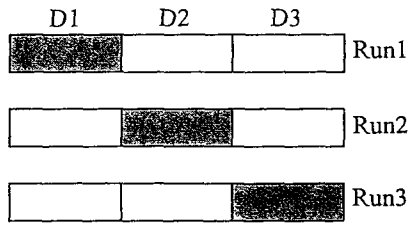


Fig. 2. Schematic illustration of the partitioning of the data set into 3 segments (D1, ...,D3) for the cross-validation. The MLP is trained (run) 3 times, each time using a sub-set of the data set in which one of the segments (shown shadowed) is omitted. Each trained network is then tested on the data from the segment that was omitted during the training (adapted from Bishop, 1995).

The cross-validation technique was here implemented dividing the CoASTS data set in three segments (Figure 2). The partition was created accounting that one water sample could correspond to various successive optical profiles (Section II-A). Then, to ensure the highest independence of the validation data, all the profiles from the same water sample were included into a unique segment.

The scattering and the bias of the algorithm results (y), with respect to corresponding measured values (t), were here expressed in terms of absolute (ε, Equation 1a) and signed (δ, Equation 1b) percentage differences, respectively:

$$\varepsilon = 100\frac{1}{N}\sum_{k=1}^{3}\sum_{i=1}^{N_k}\frac{|y_k^i - t_k^i|}{t_k^i} \tag{1a}$$

$$\delta = 100\frac{1}{N}\sum_{k=1}^{3}\sum_{i=1}^{N_k}\frac{y_k^i - t_k^i}{t_k^i} \tag{1b}$$

where the indices k and i indicate the validation segment and the sample in the validation segment, respectively; N_k is the number of samples in the validation segment k; and N is the total number of samples in the data set.

4) Complexity optimization: The MLP generalization capability depends on the algorithm architecture (i.e., the number of neurons in the hidden layer). When there are too many hidden

neurons, the MLP can reproduce non-meaningful details of the training data (i.e., such as noise and outliers). This affects the actual algorithm performance and indicates that the functional form is too complex with respect to the size of the training data set.

The algorithm complexity was optimized testing different MLP architectures having a minimum of a single neuron up to a maximum of 20 neurons into the hidden layer. Initial results showed some variability (especially for more complex MLP architectures) of the algorithm performance for subsequent retraining of the same MLP architecture. In order to reduce this variability, a *weight decay* regularization term was then added to the error function (Bishop, 1995). The weight regularization, penalizing high curvatures in the regressional function, makes more difficult to fit noise and outliers. Different weight decay coefficients (α) were tested in various trial and error exercises and the value $\alpha = 0.01$ was adopted in the current study.

5) Input selection: Besides the model architecture, also the set of input variables is relevant in the algorithm development, and was then specifically investigated. The analysis was addressed to identify the minimal set of input quantities that still ensures the algorithm accuracy. In fact, simpler MLP architectures are more robust and with a better generalization capability. In addition, the lower dimension of the input space also facilitates the identification of the region of applicability of the algorithm (D'Alimonte et al., 2003). A Bayesian statistical methodology (MacKay, 1994) is proposed in literature for the automatic determination of the relevance of input quantities. This approach is particularly efficient when applied to a high number of input variables having significantly different relative relevance. Here, due to the restricted number of variables and the correlation between the R_{rs} at different wavelengths, the identification of the appropriate set of input quantities evolved through different empirical steps. First, an MLP was developed using as input the set of R_{rs} at 490 and 555 nm commonly used for band ratio algorithms (O'Reilly et al., 1998). The performance of this model was

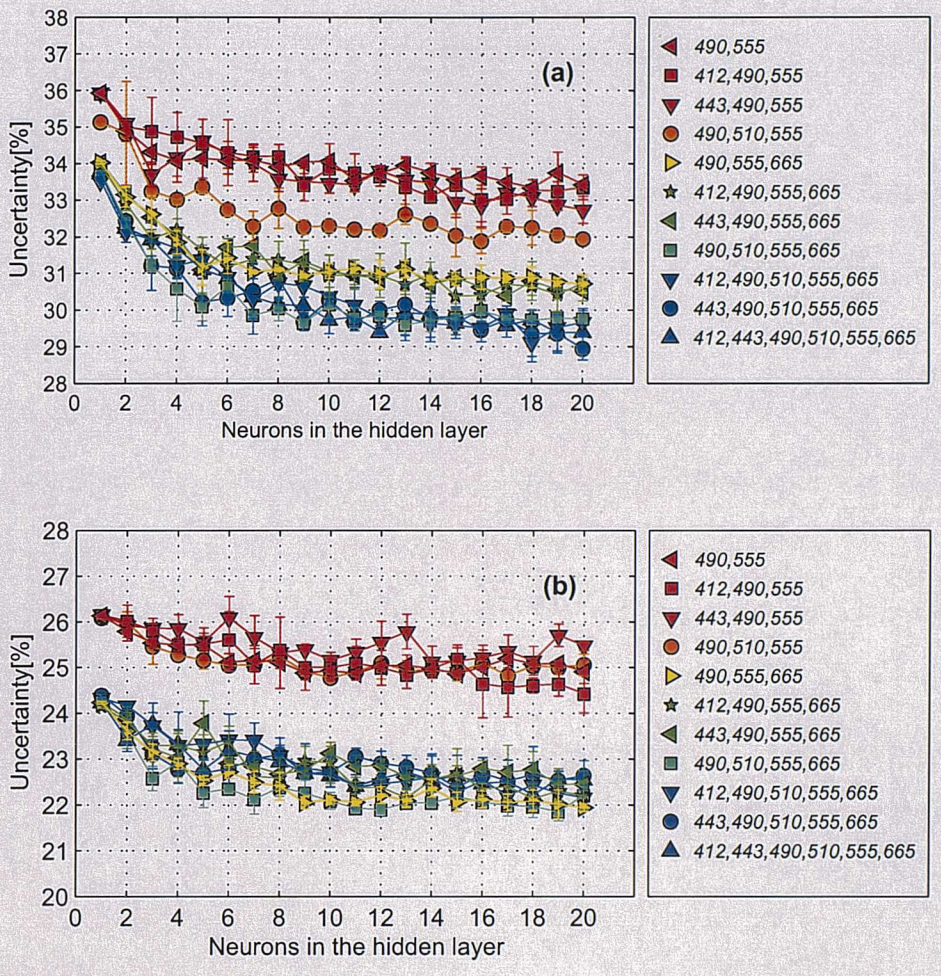


Fig. 3. Dependence of the MLP performance on the algorithm architecture for the different sets of input variables for the *Chl-a* (a) and a_{ph} (443) (b) retrieval . The error bars represent the MLP performance variability for different subsequent MLP training.

then compared with that of other MLPs developed adding to the pair of R_{rs} at 490 and 555 nm, the R_{rs} at one of the remaining wavelengths. This allowed to identify which wavelengths carry the most effective additional information. The process was then iterated adding each time the R_{rs} at one of the wavelengths not used in the configuration identified as optimal in the previous stage.

For the *Chl-a* retrieval, The dependence of the MLP performance on the algorithm architecture and on the input quantities (Figure 3(a)) showed that:

- The improvement of the MLP performance on the model architecture becomes negligible with approximatively ten hidden neurons.
- An improvement of *Chl-a* retrieval can be observed with the addition of R_{rs} at 665 to the pair R_{rs} at 490 and 555 nm. This can be explained by the combined absorbing and scattering properties of seawater optically significant constituents (Robinson, 1994): i) the phytoplankton, that on concentration increase induces a decrease in seawater reflectance in the blue and an increase in the red (when compared to pure seawater); ii) the non-pigmented particulate matter, that on concentration increase induces an increase in seawater reflectance at all wavelengths; and iii) the coloured dissolved organic matter (only absorbing), that on concentration increase induces a decrease in seawater reflectance rising from the red to the blue.
- The analysis of the MLP performance with four or more input bands, showed that some improvement can be obtained by adding the R_{rs} at 510 nm to the configuration based on 490, 555 and 665 nm. The additional use of the R_{rs} at 443 or at 412 nm was not producing any further significant improvement of the MLP performance.

In conclusion, the MLP configurations based on a minimal set of inputs and presenting best results are those using the R_{rs} at 490, 510, 555 and 665 nm or at 490, 555 and 665 nm as input. With the purpose of using the simplest configuration, the MLP applied in the comparison with the band ratio algorithms is made of ten hidden neurons and is based on the set of R_{rs} at 490, 555 and 665 nm (see Figure 4). This algorithm will be hereafter identified as $\mathrm{MLP}_{Chl\text{-}a}$.

A study similar to that made to develop the $\mathrm{MLP}_{Chl\text{-}a}$, was carried out to define the most

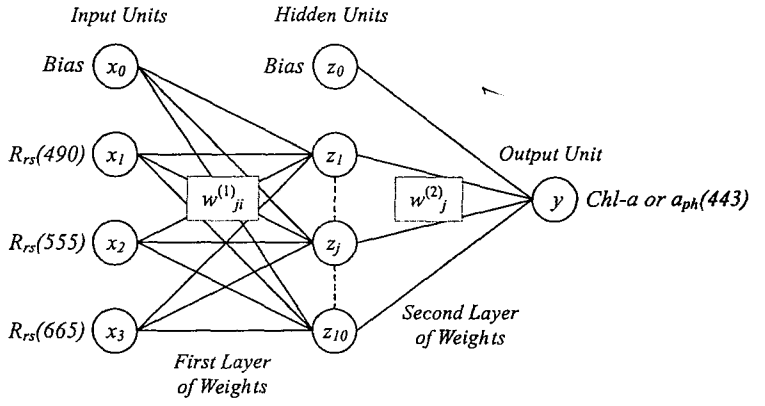


Fig. 4. Schematic representation of the MLP adopted for the retrieval of *Chl-a* and $a_{ph}(443)$. The MLP architecture consists of three input, ten hidden units and one output. The input units x_i correspond, for $i = 1, \ldots, 3$, to the R_{rs} at 490 555 and 665 nm. The extra input $x_0 \equiv 1$ is added for the *bias* parameters Bishop (1995). The hidden units combine the inputs according to $z_j = g\left(\sum_{i=0}^{3} w_{ji}^{(1)} x_i\right)$, where $w_{ji}^{(1)}$ represents the weight linking the input unit i to the hidden unit j; $w_{j0}^{(1)}$ is the bias adaptive parameter of the hidden unit j; the superscript (1) indicates the first layer of weights and $g(a) = \frac{e^a - a^{-a}}{e^a + e^{-a}}$ is the *hyperbolic tangent*. The MLP output y (corresponding to *Chl-a* or $a_{ph}(443)$) is then obtained through $y = \sum_{j=0}^{10} w_j^{(2)} z_j$, where $w_j^{(2)}$ represent for $j = 0, \ldots, 10$ the second layer of weights. Analogously to the input unit, the hidden unit $z_0 \equiv 1$ is added for the bias coefficients.

appropriate MLP for the retrieval of $a_{ph}(443)$. The results presenting its performance are given in Figure 3(b). This ($\text{MLP}_{a_{ph}}$) showed the best performance when the R_{rs} at 490, 555 and 665 nm are used as inputs (also highlighting the relevance of $R_{rs}(665)$, as for the retrieval of *Chl-a*). The additional use of the R_{rs} at 510 nm did not improve the algorithm performance, while the use of other channels seemed to reduce it. As for the $\text{MLP}_{Chl\text{-}a}$, the most appropriate algorithm architecture for the retrieval of $a_{ph}(443)$ was given by ten hidden neurons. Thus the same MLP architecture for the *Chl-a* retrieval was also adopted to model $a_{ph}(443)$.

B. Band ratio algorithms

Band ratio algorithms derived from the CoASTS data set or used for the operational *Chl-a* retrieval were applied to provide a comparison term to the MLP accuracy. Elements on these algorithms are hereafter given.

1) The two-band polynomial regressional algorithm ($AD2_{Chl\text{-}a}$): This algorithm for the retrieval of the *Chl-a* as a function of the subsurface R_{rs} at 490 and 555 nm, was implemented on the basis of a subset of the CoASTS data set (Berthon et al., 2002). Data used for the algorithm development were almost equally representing Case 1 (49%) and Case 2 (51%) waters. The algorithm equation is:

$$Chl\text{-}a = 10.0^{(0.091-1.620R_{2S}-1.148R_{2S}^2-4.949R_{2S}^3)} \quad (2)$$

where R_{2S} indicates the R_{rs} ratio at 490 and 555 nm. Hereafter this algorithm is identified as $AD2_{Chl\text{-}a}$. In this study, the $AD2_{Chl\text{-}a}$ accuracy is assessed without applying the cross-validation method (See Section III-A.3). This is assumed to not affect the algorithm validation results as i) $AD2_{Chl\text{-}a}$ was developed with a subset of the CoASTS data and ii) polynomial regressional algorithms are much less subject to over fitting problems than the MLP algorithms.

2) The OC4v4 *algorithm:* The OC4v4 algorithm was developed on the basis of a SeaBAM data set (O'Reilly et al., 2000). The initial version of this data set was made of 919 R_{rs} spectra and coincident *chlorophyll-a* concentration collected in different oceanic provinces. A first polynomial band ratio algorithm, namely OC2, based on the initial SeaBAM data set demonstrated a good performance (with an accuracy of $\pm 35\%$) in Case 1 waters with *Chl-a* between 0.03 to 1 mg m^{-3}, while a general overestimate was observed in eutrophic waters (O'Reilly et al., 2000). In order to improve the *Chl-a* retrieval, the SeaBAM data set was enlarged including new data (higher *Chl-a* and additional Case 2 water conditions). Some of these additional data (35 points) were from the CoASTS data set.

The OC4V4 algorithm, based on the enlarged version of the SeaBAM data set (2853 points), is the result of a fourth order polynomial interpolation of the logarithm of the *chlorophyll-a* concentration as a function of:

$$R_{4S} = \max_i \left\{ R_{555}^{\lambda_i} \right\} \tag{3}$$

where $R_{555}^{\lambda_i}$ represents the ratio with respect to R_{rs} at 555 nm of R_{rs} at 443, 490 and 510 nm (Maximum Band Ratio approach (MBR)). The MBR approach was adopted to use the R_{rs} ratio at the most suitable wavelengths for different *chlorophyll-a* concentration ranges. Specifically the algorithm switches from the 443 nm channel in oligotrophic waters to the 510 nm channel in eutrophic waters. The OC4v4 equation is:

$$Chl\text{-}a = 10.0^{(0.366 - 3.067 R_{4S} + 1.930 R_{4S}^2 + 0.649 R_{4S}^3 - 1.532 R_{4S}^4)} \tag{4}$$

3) Four-band polynomial regressional algorithms for retrieving the Chl-a ($AD4_{Chl\text{-}a}$) and the $a_{ph}(443)$ ($AD4_{a_{ph}}$): Polynomial regressional algorithms derived from the CoASTS data set to determine the *Chl-a* and the $a_{ph}(443)$ were developed to specifically support the analysis of the MLP performance . These algorithms, here identified as $AD4_{Chl\text{-}a}$ and $AD4_{a_{ph}}$ (for *Chl-a* and $a_{ph}(443)$, respectively), use the same set of inputs (R_{rs} at 443, 490, 510 and 555 nm) and the MBR approach applied for OC4v4. The accuracy of these algorithms was evaluated with the same cross-validation methodology applied to the corresponding MLP models. Versions of the $AD4_{Chl\text{-}a}$ (Eq. 5a) and $AD4_{a_{ph}}$ (Eq. 5b) algorithms, created using the entire CoASTS data set, are given by:

$$Chl\text{-}a = 10.0^{(0.236 - 3.331 R_{4S} + 2.386 R_{4S}^2 + 4.2834 R_{4S}^3 - 5.816 R_{4S}^4)} \tag{5a}$$

$$a_{ph}(443) = 10.0^{(-3.360 - 2.107 R_{4S} + 2.806 R_{4S}^2 - 2.735 R_{4S}^3 - 1.174 R_{4S}^4)} \tag{5b}$$

where R_{4S} is defined in Equation 3.

		All (n=906)			Case1 (n=497)			Case2 (n=274)			SeaWiFS (n=31)		
		ε(%)	δ(%)	r^2	ε(%)	δ(%)	r^2	ε(%)	δ(%)	r^2	ε(%)	δ(%)	r^2
Chl-a	$MLP_{Chl\text{-}a}$	31	4	0.69	27	-11	0.70	31	15	0.70	30	-18	0.80
	$AD2_{Chl\text{-}a}$	35	-14	0.57	37	-31	0.64	33	3	0.58	35	-11	0.72
	OC4v4	60	51	0.56	40	24	0.63	80	77	0.58	65	56	0.66
	$AD4_{Chl\text{-}a}$	36	8	0.56	29	-10	0.62	41	27	0.56	41	12	0.66
a_{ph}	$MLP_{a_{ph}}$	22	3	0.77	23	1	0.76	20	4	0.79	27	-18	0.67
	$AD4_{a_{ph}}$	26	5	0.64	28	1	0.65	26	10	0.70	32	4	0.65

TABLE I. Summary of the algorithms accuracy in retrieving *Chl-a* and $a_{ph}(443)$.

IV. Results and Discussion

A. The retrieval of Chl-a

This section analyzes the accuracy of the MLP and of the different band ratio algorithms here proposed for the retrieval of *Chl-a*. The algorithm accuracy is assessed over the entire CoASTS data set (Figure 5), as well as i) separately considering Case 1 and Case 2 water conditions for stations including measurements of the beam attenuation at 660 nm (i.e., needed for the water type classifi cation) and ii) with SeaWiFS derived R_{rs} (Table IV-A.1).

1) $MLP_{Chl\text{-}a}$ *:* The algorithm validation over the entire CoASTS data set shows an accuracy $\varepsilon = 31\%$ with a mean overestimation $\delta = 4\%$ (Figure 5(a1) and (b1), respectively). When Case 1 and Case 2 conditions are separately considered (Table IV-A.1), the algorithm accuracy becomes $\varepsilon = 27\%$ (with an average underestimation $\delta = -11\%$) for the former, and $\varepsilon = 31\%$ (with an average overestimation $\delta = 15\%$) for the latter. An analogous accuracy ($\varepsilon = 30\%$) is also observed with SeaWiFS derived R_{rs}, even though associated to a mean underestimation $\delta = -18\%$ (probably explained by the additional R_{rs} uncertainty due to the atmospheric correction of satellite data).

A further test, made adding the R_{rs} at 510 nm as input, demonstrated that this channel

slightly improved the *Chl-a* retrieval in Case 1 water but reduced the algorithm performance in Case 2 conditions. This furthermore confirms the validity of the set of inputs previously identified.

2) $AD2_{Chl\text{-}a}$: The $AD2_{Chl\text{-}a}$ algorithm shows an uncertainty $\varepsilon = 35\%$ with respect to the entire CoASTS data set and a tendency to underestimate ($\delta = -14\%$) *Chl-a*, especially in oligotrophic conditions (Figures 5(b1) and 5(b2)). The assessment of the model accuracy separating Case 1 and Case 2 waters conditions shows a better results for the latter ($\varepsilon = 33\%$, $\delta = 3\%$) than for the former ($\varepsilon = 37\%$, $\delta = -31\%$). This can be explained by the different representativeness of Case 1 and Case 2 waters within the development (half Case 1 and half Case 2 waters) and the validation data sets (two third of Case 1 waters), as highlighted in Sections III-B.1 and II-A, respectively. The $AD2_{Chl\text{-}a}$ application to SeaWiFS derived R_{rs} gives $\varepsilon = 35\%$ and $\delta = -11\%$.

3) OC4v4: The application of the OC4v4 algorithm (Figure 5(c)) generally shows over-estimated values, especially for low *Chl-a* concentration. For the Case 1 waters included in the CoASTS data set, the retrieved *Chl-a* is more accurate ($\varepsilon = 40\%$) than for the optically complex conditions ($\varepsilon = 80\%$) given by the Case 2 waters (see Table IV-A.1). Also, the OC4v4 results are more biased in Case 2 waters ($\delta = 77\%$) than in Case 1 waters ($\delta = 24\%$). The low performance of the OC4v4 algorithm, especially in Case 2 water condition, can be explained by the different waters types represented by the CoASTS and SeaBAM data sets and highlights the importance of developing specific algorithms for optically complex coastal regions. This result is made more significant by the fact that a small subset of the CoASTS data was used for the OC4v4 development.

4) $AD4_{Chl\text{-}a}$: The MBR polynomial algorithm specifically developed using the CoASTS data shows, with respect to the entire CoASTS data set, an uncertainty of 36% (Figure 5(d1)). The analysis restricted to Case 1 water shows $\varepsilon = 29\%$ and $\delta = -10\%$, while it exhibits a

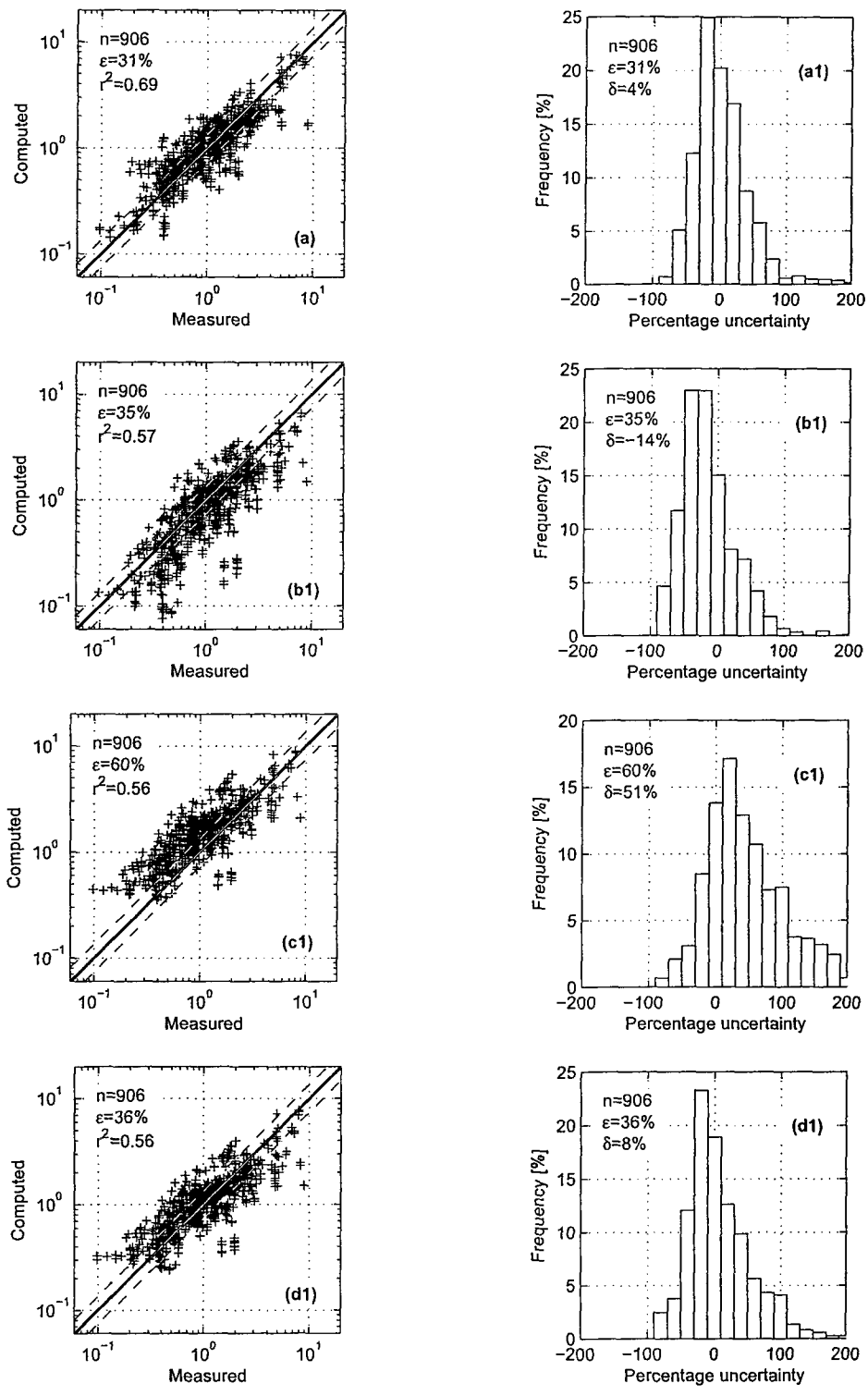


Fig. 5. Scatter plot of computed versus measured *Chl-a* (left column) and corresponding distribution of uncertainty (right column) for the different algorithms analyzed ($MLP_{Chl\text{-}a}$, $AD2_{Chl\text{-}a}$, OC4v4 and $AD4_{Chl\text{-}a}$ in (a), (b), (c) and (d), respectively). The *chlorophyll-a* concentration is expressed in mg m^{-3}. The dashed lines represent the uncertainty of ±35% with respect to the 1 to 1 line.

lower accuracy in Case 2 waters ($\varepsilon = 41\%$ and $\delta = 27\%$). In particular, it can be observed that the algorithm tends to overestimate the *Chl-a* in oligotrophic conditions. The comparison of these results with those from the $MLP_{Chl\text{-}a}$ and $AD2_{Chl\text{-}a}$ algorithms indicates that i) band ratio algorithms developed on the basis of the CoASTS data set, whereas generally less performing than the $MLP_{Chl\text{-}a}$, can anyway give acceptable results (uncertainty less than 35%) for the represented coastal waters and ii) band ratio algorithms are less performing than MLP in handling, within a single regressional model, both Case 1 and Case 2 conditions.

B. $a_{ph}(443)$ retrieval

The same approach used to assess the accuracy of the MLP and the MBR algorithms in retrieving the *Chl-a*, were used to verify how these functional forms are suitable to develop algorithms for the retrieval of $a_{ph}(443)$ (see Figure 6).

1) $MLP_{a_{ph}}$: The accuracy in determining the $a_{ph}(443)$, shows that this quantity can be retrieved with an higher accuracy than *Chl-a*. In fact, the uncertainty over the entire CoASTS data set is $\varepsilon = 22\%$ for $a_{ph}(443)$, while it is $\varepsilon = 31\%$ for *Chl-a*. The separate analysis of the algorithm accuracy in Case 1 and Case 2 waters shows uncertainties $\varepsilon = 23\%$ with $\delta = 1\%$ and $\varepsilon = 20\%$ with $\delta = 4\%$, respectively. The $a_{ph}(443)$ values resulting from the application of the $MLP_{a_{ph}}$ algorithm to SeaWiFS derived R_{rs} exhibit uncertainty $\varepsilon = 27\%$ with $\delta = -18\%$. Among the algorithms considered, the $MLP_{a_{ph}}$ presents the largest difference between the accuracy determined using the *in situ* and the SeaWiFS derived R_{rs}. This can be explained considering that the relationship between $a_{ph}(443)$ and the R_{rs} spectra is more deterministic than that between *Chl-a* and the R_{rs} spectra. In other words, the MLP is more effective in reproducing the $a_{ph}(443)$ versus R_{rs} relationships, but the resulting algorithm is also more sensitive to the uncertainty affecting the input quantities, i.e.,R_{rs}.

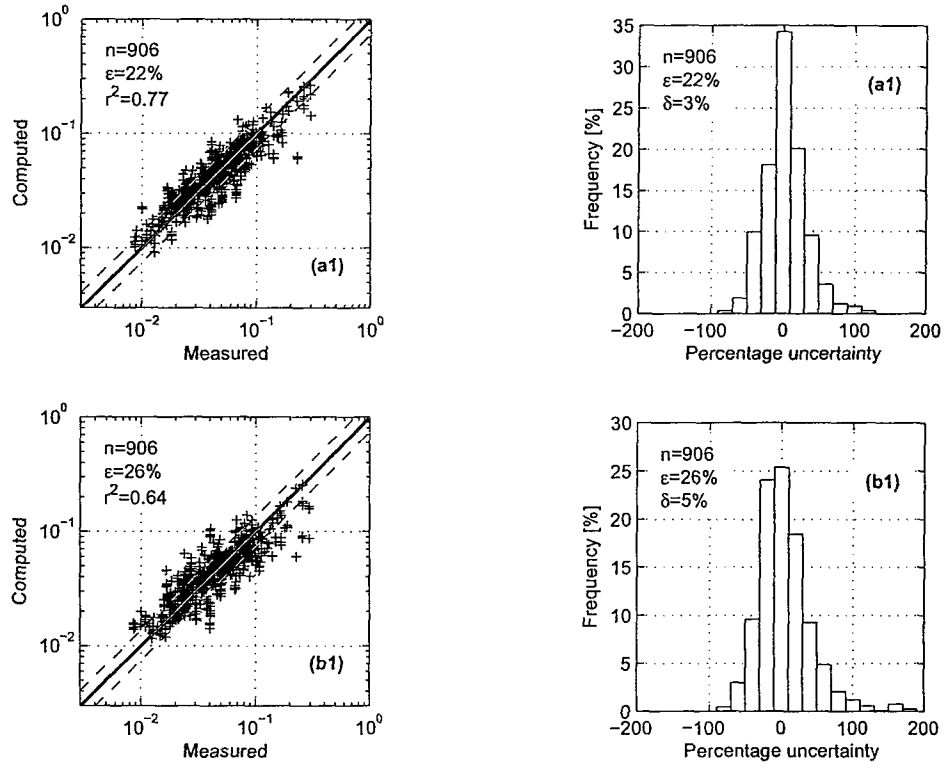


Fig. 6. Scatter plot of computed versus measured $a_{ph}(443)$ (left column) and corresponding uncertainty distribution (right column) as resulting from the application of the $MLP_{a_{ph}}$ and$AD4_{a_{ph}}$ models ((a) and (b), respectively). The absorption of the pigmented particulate matter is expressed in m^{-1}. The dashed lines represent the uncertainty of ±35% with respect to the 1 to 1 line.

2) $AD4_{a_{ph}}$: Whereas the $AD4_{a_{ph}}$ accuracy is not as good as for the $MLP_{a_{ph}}$, it confirms the improved accuracy in retrieving the $a_{ph}(443)$ with respect to the *Chl-a*, shown by the $MLP_{Chl\text{-}a}$ algorithm. Over the entire CoASTS data set, $AD4_{a_{ph}}$ shows an uncertainty $\varepsilon = 26\%$ associated to a general overestimate $\delta = 5\%$. The differences between Case 1 and Case 2 waters conditions ($\varepsilon = 28\%$ with $\delta = -1\%$ and $\varepsilon = 26\%$ with $\delta = 10\%$, respectively) are lower than those observed for the *Chl-a* retrieval adopting the same functional form. As already observed for the $MLP_{a_{ph}}$, a significant reduction of the algorithm accuracy characterizes the application of the $AD4_{Chl\text{-}a}$ to SeaWiFS derived R_{rs} with respect to products derived from

in situ R_{rs}.

V. Summary and Conclusion

The capability of the MLP and of other band ratio algorithms to determine the phytoplankton components (i.e., *Chl-a* and $a_{ph}(443)$) from R_{rs} was evaluated on the basis of the CoASTS data set, representative of optically complex coastal waters. The MLP algorithm complexity was optimized testing different MLP architectures and different sets of input R_{rs}. The MLP configuration identified as the most suitable for the retrieval of both *Chl-a* and $a_{ph}(443)$ is made by ten hidden neurons and is based on the R_{rs} at 490, 555 and 665 nm.

The analysis of the accuracy of the MLP and of the other band ratio algorithms showed that:

- The MLP is more effective than band ratio algorithms in modelling both the *Chl-a* and the $a_{ph}(443)$. This resulted from an accuracy analysis carried out with the entire CoASTS data set as well as separately considering Case 1 and Case 2 water conditions, and with SeaWiFS derived R_{rs}.
- The MLP and MBR algorithms, both demonstrated that $a_{ph}(443)$ can be determined with a better accuracy than *Chl-a* (with an improved accuracy of about 10%). This suggests that models for primary production study could benefit of this higher accuracy on the a_{ph} determination when assimilating ocean colour products.
- The band ratio algorithms developed on the basis of the CoASTS data set showed a limited capability in handling within a single algorithm complex relationships relating *chlorophyll-a* and R_{rs}, both Case 1 and Case 2 waters.
- The OC4v4, when applied to the CoASTS data set, shows the highest uncertainty in determining *Chl-a*. This is highlighted by the results obtained for Case 2 waters (see Table IV-A.1) and for oligotrophic cases (see Figure 5(c1)), and further on confirms the

Determination of CDOM and NPPM Absorption Coefficient Spectra from Coastal Water Remote Sensing Reflectances[1]

Abstract

Multi Layer Perceptron (MLP) neural network algorithms were developed to retrieve the absorption coefficient spectra of the colored dissolved organic matter (CDOM) and non pigmented particulate matter (NPPM) from the remote sensing reflectance R_{rs} of optically complex waters. The two MLP algorithms, consisting of one hidden layer with ten neurons and requiring R_{rs} at 412, 490 and 665 nm as inputs, were trained with a comprehensive experimental data set of the Northern Adriatic Sea coastal waters. The products of the proposed regional MLP algorithms showed lower uncertainties than regional band-ratio algorithms, and exhibited average values of 20% and 25% in the determination of CDOM and NPPM absorption coefficients at 412 nm, respectively.

Index Terms

Neural network, bio-optical modelling, ocean colour.

[1]The content of this Section was submitted to *IEEE Transactions in Geoscience and Remote Sensing* by Davide D'Alimonte and Giuseppe Zibordi, in March 2004. The core theme of the paper is the retrieval of the IOPs from R_{rs} data using MLP algorithms. The ideas, development and writing up of the paper were principal responsibility of the thesis author. The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into a based team research.

I. INTRODUCTION

SEAWATER components are conventionally grouped on the basis of their optical properties and bio-geo-chemical origin in: *pigmented particulate matter*, *colored dissolved organic matter* (CDOM) and *non-pigmented particulate matter* (NPPM). In the open ocean water, generally indicated as Case 1 (Morel and Prieur, 1977), CDOM and NPPM are mainly of phytoplankton origin and thus well correlated with it. For this reason, algorithms for the determination of seawater components from remote sensing data of open ocean waters, are mostly restricted to phytoplankton chlorophyll *a*. Differently, specific algorithms for the determination of CDOM and NPPM in coastal regions (here identified as permanently or occasionally Case 2 (Morel and Prieur, 1977) due to bottom resuspension, coastal erosion, river inputs, or high anthropogenic impact), are still matter of investigations due to the reduced correlation among seawater components (Sathyendranath, 2000).

Sensitivity studies based on synthetic data addressed the accuracy in retrieving the seawater inherent optical properties using an oceanic radiance model (Hoge and Lyon, 1996, 1999). On the applicative side a semi-analytical model was proposed for the retrieval of combined colored dissolved organic matter and detritus particulate (equivalent to NPPM) absorption coefficients from Moderate-Resolution Imaging Spectrometer (MODIS) data (Carder et al., 1999) assuming the absorption exponentially decreases with wavelength according to a constant decay slope. Various empirical algorithms were additionally proposed (Buckton et al., 1999; Kahru and Mitchell, 2001; Keiner and Brown, 1999; Shiller and Doerffer, 1997; Tassan, 1994) for the independent retrieval of CDOM and NPPM absorption coefficients (a_{CDOM} and a_{NPPM}, respectively) from remote sensing data at specific wavelengths. These models and algorithms cannot however provide the spectral dependence of a_{CDOM} and a_{NPPM}, which may vary from region to region, according to the specific bio-geochemical components of CDOM and

NPPM (Berthon et al., 2002; Ferrari, 2000).

Aiming at improving the optical characterization of CDOM and NPPM in coastal waters through remote sensing, a methodology is here presented to retrieve the a_{CDOM} and a_{NPPM} spectra, instead of their values at a single wavelength, only making assumptions on the functional form of their wavelength dependence. The proposed methodology is based on a Multi Layer Perceptron (MLP) neural network for the determination of the parameters defining the a_{CDOM} and a_{NPPM} spectra from the remote sensing reflectance R_{rs} at different wavelengths. The study, restricted to the development of regional MLP algorithms, was carried out with a comprehensive data set collected in an optically complex coastal region of the northern Adriatic Sea.

This work complements a previous study based on the same experimental data set and addressing the effectiveness of MLP algorithms in retrieving the phytoplankton component in optically complex waters (D.D'Alimonte and G.Zibordi, 2003).

II. DATA SET

The data set used in this study was collected within the framework of the Coastal Atmosphere and Sea Time-Series (CoASTS) program. This, since 1995, supports ocean color calibration and validation activities through a comprehensive monthly collection of data at the Acqua Alta Oceanographic Tower (AAOT) in the northern Adriatic Sea (Zibordi et al., 2002). The measurement site, considered well representing the northern Adriatic Sea coastal region, is located approximately 8 nautical miles south east of the Venice Lagoon (45.314 N, 12.508 E) and is characterized by the presence of water types close to the separation threshold suggested by Loisel and Morel (1998) for Case 1 and Case 2 waters (with roughly one third of the CoASTS measurement stations pertaining to Case 2 (Berthon et al., 2002)).

The data set used in this study is composed of a total number of 906 pairs of a_{CDOM} –or

a_{NPPM}– and corresponding R_{rs} spectra, collected from October 1995 to December 2001. A short description of the methods applied for the production of the R_{rs}, a_{CDOM} and a_{NPPM} data is given in the following subsections.

A. Remote Sensing Reflectance

In-water profiles of upwelling radiance and above water downward irradiance data from the Wire Stabilized Profiling Environmental Radiometer (WiSPER) system (Zibordi et al., 2002) were used to compute the R_{rs} at the Sea-Viewing Wide Field-of-view Sensor (SeaWiFS) center wavelengths (i.e., λ = 412, 443, 490, 510, 555, and 665 nm instead of 670 nm) of relevance for remote sensing applications.

WiSPER is a winched system deployed through a custom-built profiling rig from the main structure of the AAOT at a speed of 0.1 m s^{-1} . The processing of WiSPER data was made using an Optical Processor developed for the analysis of the CoASTS data (D'Alimonte and Zibordi, 2001). The processing was carried through successive steps: *i.* the in water profile data were calibrated and corrected for the changes in the illumination conditions using the above water downward irradiances; *ii.* the extrapolation interval satisfying the requirement of linear decay of the logarithm of the upwelling radiances as a function of depth, was determined on a cast-by-cast basis; *iii.* the sub-surface upwelling radiance $L_u(0^-, \lambda)$ was extrapolated from the linear fit of the logarithm of upwelling radiances versus depth, after removing the self-shading, bottom effects, and tower-shading perturbations (Zibordi et al., 2002); *iv.* finally the remote sensing reflectance was computed according to

$$R_{rs}(\lambda) = \frac{0.54\ L_u(0^-, \lambda)}{E_d(0^+, \lambda)} \tag{1}$$

where $E_d(0^+, \lambda)$ is the downward irradiance at a chosen reference time.

The R_{rs} data relevant to this study exhibit R_{rs} (490) values in the range of 0.002–0.026 sr^{-1}, with mean and standard deviation equal to 0.007 and 0.003 sr^{-1}, respectively.

B. CDOM and NPPM Absorption Coefficients

The a_{CDOM} values were obtained from spectrometric analysis of seawater filtered with 0.22 μm cellulose filters. The analysis was performed by placing a 10 cm quartz cuvette containing Milli-Q water in the optical path of the reference beam and a 10 cm quartz cuvette containing the filtered seawater sample in the optical path of the sample beam, of a Perkin Elmer Lambda-10 Spectrometer. The a_{CDOM} spectra (determined with 1 nm resolution in the 350–750 nm spectral range) were computed from the measured absorbance spectra A_{CDOM} resulting from the difference between the sample and the reference absorbance (Ferrari et al., 1996) from

$$a_{CDOM}(\lambda) = 2.3\frac{A_{CDOM}(\lambda)}{L_c} \tag{2}$$

where L_c is the path length of the cuvette (in units of m). The instrument background was removed using measurements performed with Milli-Q water in both the sample and the reference cuvettes.

The a_{CDOM} data relevant to this work are the values at the center-wavelengths of the WiSPER radiometers. The required a_{CDOM} values were computed averaging the a_{CDOM} data falling within ± 3 nm interval centered at the wavelength of interest. The a_{CDOM} data applied in this study exhibit values in the range 0.02–0.26 m^{-1} at 412 nm, with mean and standard deviation equal to 0.12 and 0.05 m^{-1}, respectively.

The a_{NPPM} values were produced from the spectrometric analysis of the deposit obtained from seawater filtration on fiber glass filters of 0.7 μm pore size. The analysis first required the determination of the total seawater particulate matter absorption coefficient, a_p, and successively the discrimination of the components related to the pigmented, a_{ph}, and the non pigmented, a_{NPPM}, fractions. The applied method (Tassan and Ferrari, 1995) is particularly suitable for determining the absorption components of water samples characterized

by highly backscattering mineral particles or by highly absorbing sediments likely to occur in coastal regions. Specifically, a_p spectra were computed from the equivalent particle suspension absorbance A_{sus} (determined with 1 nm resolution in the 400–750 nm spectral range) using

$$a_p(\lambda) = 2.3A_{sus}(\lambda)\frac{F_a}{V_w} \tag{3}$$

where V_w is the volume of filtered water (in units of m^3), F_a is the filter clearance area (in units of m^2), and $A_{sus}(\lambda)$ is the equivalent particle suspension absorbance derived from the transmission and reflection measurements performed with a Perkin Elmer Lambda-19 equipped with an integrating sphere (Tassan and Ferrari, 1995). The a_{ph} and a_{NPPM} fractions of the total particulate absorption coefficient, a_p, were separated through bleaching of the sample on the filter using a solution of sodium hypochlorite (NaClO) as an oxidizing agent (Ferrari and Tassan, 1999). This oxidation acts rapidly on pigment molecules and slowly on detritus, thereby permits a selective analysis of the absorption components of non pigmented particles retained on the filter.

Similarly to a_{CDOM}, the a_{NPPM} data relevant to this work are those at the center-wavelengths of the WiSPER radiometers. The required a_{NPPM} values were computed averaging the a_{NPPM} spectral data falling within ± 3 nm interval centered at each wavelength of interest. The a_{NPPM} data applied in this study, exhibit values in the range of 0.01–0.13 m^{-1} at 412 nm (excluding two outliers showing much higher values), with mean and standard deviation equal to 0.05 and 0.03 m^{-1}, respectively.

C. Feature extraction

III. Methods

The proposed methodology aims at determining the CDOM and NPPM absorption spectra from R_{rs} data in the visible spectral region. The study relies on experimental data with the

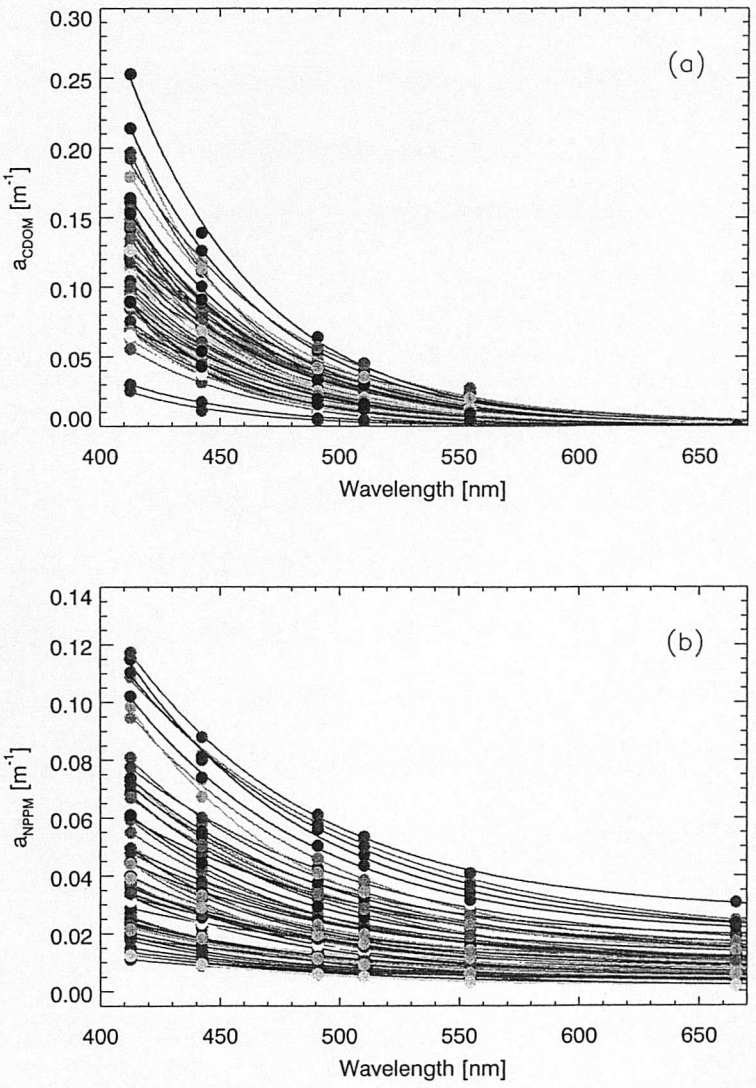


Fig. 1. Sample absorption spectra of a_{CDOM} (a) and of a_{NPPM} (b). Dots represent experimental measurements at the SeaWiFS center-wavelengths, while lines correspond to the exponential regressional functions. Only a subset (one record each twenty) of the CoASTS data was here displayed to allow for a better identification of the individual regression lines.

assumption that the a_{CDOM} and a_{NPPM} spectra can be described through an exponential function (Berthon et al., 2002; Ferrari, 2000). This section presents the methodologies applied

in *i.* determining the regressional parameters of the functions describing the a_{CDOM} and a_{NPPM} spectra (i.e., *feature extraction*) and *ii.* implementing the MLP algorithms to relate the the R_{rs} spectral data to the regressional parameters.

The a_{CDOM} and a_{NPPM} spectral values can be modeled through the general exponential function

$$a_{xx}(\lambda) = a_{0xx}e^{-S_{xx}(\lambda-\lambda_0)} + cb_{xx}, \quad (4)$$

where λ indicates discrete wavelengths in the 412–665 nm interval, λ_0 is the reference wavelength here chosen at 412 nm, xx indicates $CDOM$ or $NPPM$, a_{0xx}, S_{xx} and cb_{xx} are the regressional parameters, hereafter referred to as the *coefficient*, the *slope* and the *background*, respectively. The *background* for a_{CDOM} is taken null (i.e. $cb_{CDOM} = 0$) because the CDOM absorption generally assumes negligible values above 650 nm (Ferrari, 2000). Differently, the NPPM absorption of complex coastal waters is not negligible in the red and near-infrared (Tassan and Ferrari, 2004), and the use of $cb_{NPPM} \geq 0$ in equation 4 was shown appropriate to parameterize a_{NPPM} (Berthon et al., 2002).

The regressional parameters describing the CDOM and NPPM spectral absorption were determined fitting equation 4 through non-linear least squares regressions –based on a gradient-expansion algorithm– of the measured absorption coefficients. Figure 1 shows, through the experimental values and the corresponding interpolation curves, the absorption coefficient spectra of CDOM and NPPM (panels (a) and (b), respectively) at the SeaWiFS center-wavelengths.

A. MLP implementation

Various neural networks were recently applied to support remote sensing studies. Among these, the Multi Layer Perceptron (MLP) (D'Alimonte, 2003) and the Generalized Radial Basis Function (GRBF) (Cipollini et al., 2001) neural networks were explicitly used for

ocean color applications. Attractive property of the MLP is its capacity to efficiently represent complex relationships between multidimensional input-output spaces (Bishop, 1995). The MLP is here used to relate R_{rs} to the regressional parameters (see equation 4) defining the spectra of a_{CDOM} and a_{NPPM}. The following sections describe the methods applied for the implemention[2] of the two distinct MLP algorithms proposed for the determination of a_{CDOM} and a_{NPPM} from R_{rs}.

1) MLP architecture: One relevant aspect in developing an MLP algorithm is the *over fitting* problem. In fact MLP algorithms based on too many adaptive parameters may fit the noise and outliers included in the training data set and this reduces the capability of correctly interpreting new input data (i.e., *generalization*). To improve the MLP generalization capability it is necessary to define a suitable MLP architecture (i.e., number of MLP adaptive parameters), and possibly to also apply a *regularization term* (Bishop, 1995) to penalize high curvatures of the MLP regressional function driven by noise and outliers. Former studies (D'Alimonte, 2003; D.D'Alimonte and G.Zibordi, 2003), focused on the use of MLP for the development of ocean color algorithms and making use of the current experimental data set, already addressed both aspects. These studies investigated the performance of different configurations with a minimum of a single neuron up to a maximum of 20 in one hidden layer. Results showed the suitability of a MLP architecture composed of one hidden layer of ten neurons with an hyperbolic tangent activation function for the hidden layer and the identity function for the output layer. In addition, a regularization weight decay term with a coefficient equal to 0.01 was identified through trial and error exercises. The same MLP configuration was also applied in the present study, and its suitability was a posteriori supported by the assessed performance of the resulting algorithms.

[2]Neural network algorithms have been implemented through the Netlab (Nabney, 2001) toolbox for Matlab.

2) Assessment of the algorithm performance through cross-validation: To assess the MLP performance, it is necessary to apply the algorithm to *validation* data different from those used for the training of the MLP. The technique here adopted to assess the accuracy of the MLP products was the *cross-validation* (Cherkassky and Muller, 1998). This technique consists in separating the available data into different segments, alternatively used for the MLP training and validation. The cross validation was here applied separating the CoASTS data set into three different and independent segments. The ensemble of the data of two segments was used for the training process while the resulting algorithm was validated with the data of the excluded segment. This process was iterated three times, each time using data from a different validation segment. It is pointed out that the CoASTS data set was partitioned taking into account that one water sample used to produce a_{CDOM} and a_{NPPM} measurements, can correspond to multiple optical profiles and thus to multiple R_{rs}. Thus, to ensure the independence of the validation with respect to the training data, all the data related to the same water sample were included into a unique segment.

The performance of the MLP was quantified through the absolute value of the relative percent difference, ε, between the regressional coefficients t_k (i.e., a_{0xx} or S_{xx} or cb_{NPPM}) determined fitting equation 4 to the experimental data and the corresponding values y_k retrieved through the MLP

$$\varepsilon = 100\frac{1}{N}\sum_{h=1}^{3}\sum_{i=1}^{N_h}\left|\frac{y_k^i - t_k^i}{t_k^i}\right|, \tag{5}$$

where h and i are the validation segment and the sample in the validation segment, respectively; N_h indicates the number of samples in the validation segment h; and N is the total number of samples in the data set.

3) MLP input selection: In agrement with a recent study (D.D'Alimonte and G.Zibordi, 2003) showing that R_{rs} at three different and properly identified wavelengths can still represent

a suitable choice for the determination of phytoplankton absorption and concentration, the current work investigated the performance of a subset of the available R_{rs} wavelengths for the retrieval of a_{NPPM} and a_{CDOM}. The objective of this process was to reduce the number of R_{rs} inputs to a subset still ensuring a relatively high accuracy in determining CDOM and NPPM absorption coefficients. This aimed at improving the generalization capability of the MLP algorithms and at allowing for an effective determination of their range of applicability when operationally used (D'Alimonte et al., 2003). In order to identify the proper subset of input R_{rs} wavelengths, the performance of the MLP algorithms was assessed using the R_{rs} at all wavelengths. Then, the results were compared with those obtained removing from the set of MLP inputs R_{rs} at one wavelength at a time (i.e., by incrementally dropping one input at a time). The configuration based on one less input and giving the best result was retained and the procedure reiterated.

To select the set of inputs on the basis of a more accurate evaluation of the MLP performance, the problem represented by a non effective learning process was also considered. In fact, the iterative process determining the adaptive parameters of the MLP (here, the Scaled Conjugate Gradient (SCG) learning algorithm (Bishop, 1995)) can be staked into a poor local minimum of the error function. In this case, the resulting algorithm performance does not express the true capability of the applied configuration. To avoid this problem, each MLP configuration was initialized, trained and validated three times to verify the consistency of the algorithms resulting from independent learning processes.

B. Uncertainty in retrieving the absorption functional form

The uncertainties in the retrieved regressional parameters defining the fitted a_{CDOM} and a_{NPPM} spectra are presented in Table I and II for different sets of input wavelengths. These uncertainty values show only a general slight difference with the number of inputs. The

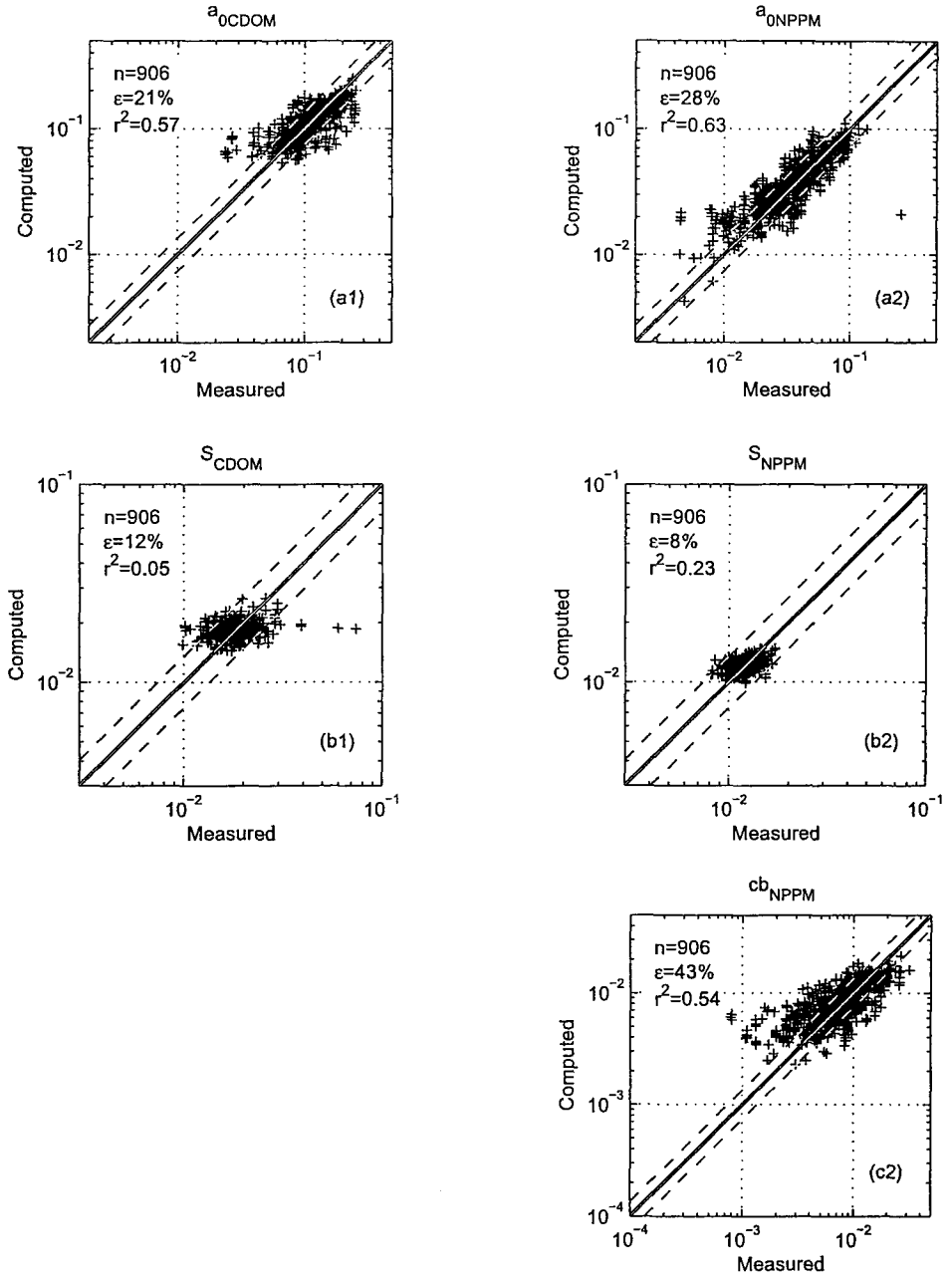


Fig. 2. Scatter plots of the computed regressional parameters (*coefficient* in units of m^{-1}, *slope* in units of λ^{-1} and *background* in units of m^{-1}, in (a),(b) and (c), respectively) versus the corresponding values derived from experimental measurements for the a_{CDOM} (left column) and a_{NPPM} (right column). The dashed lines indicate the ±35% uncertainty threshold.

$\lambda[nm]$	a_{0CDOM}	S_{CDOM}
412, 443, 490, 510, 555, 665	20.1 ± 0.2	12.11 ± 0.20
412, 443, 490, 555, 665	20.1 ± 0.3	11.93 ± 0.17
412, 443, 490, 665	20.4 ± 0.1	12.16 ± 0.15
412, 490, 665	20.8 ± 0.1	12.19 ± 0.05

TABLE I. Summary of the uncertainties in percent, for the computed a_{CDOM} regressional coefficients. The values presented after the $\pm$ sign represent the variability of the results obtained from three subsequent initialization, training and validation of the same MLP. These values were used to identify non effective learning processes.

$\lambda[nm]$	a_{0NPPM}	S_{NPPM}	cb_{NPPM}
412, 443, 490, 510, 555, 665	28.3 ± 0.6	8.13 ± 0.04	43.4 ± 0.8
412, 443, 490, 555, 665	27.7 ± 0.6	8.03 ± 0.07	42.9 ± 0.4
412, 490, 555, 665	27.5 ± 0.8	8.08 ± 0.08	42.7 ± 0.2
412, 490, 665	27.7 ± 0.5	8.30 ± 0.02	42.9 ± 0.1

TABLE II. As in Table I, but for the a_{NPPM} regressional coefficients.

scatter plots of the *coefficient*, *slope* and *background* (for a_{NPPM} only) retrieved by the MLP algorithms using R_{rs} at 412, 490 and 665 nm (for both a_{CDOM} and a_{NPPM}) versus the corresponding values derived from the experimental data, are presented in figure 2.

1) Retrieval of the a_{CDOM} regressional parameters: The analysis showed that the subset of three inputs for which the MLP algorithm for a_{CDOM} produces the most accurate results is composed of R_{rs} at 412, 490 and 665 nm. The difference in accuracy between this configuration and that using the full set of R_{rs}, is $\sim 0.7\%$ for the *coefficient* and negligible for the *slope*. The evaluation of the accuracy in retrieving the a_{CDOM} regressional parameters showed that the *coefficient* and the *slope* (see Section II-C) can be determined with an uncertainty of about 21% and 12%, respectively (see Table I). The analysis also highlighted

the relevance of R_{rs} at both 412 and 665 nm. In fact the removal of R_{rs} at one of the two wavelengths produced a significant reduction in the accuracy of the retrieved a_{CDOM} regressional parameters (result not presented here).

The scatter plots in panels (a.1) and (b.1) of figure 2 show that some retrieved *coefficients* of a_{CDOM} exhibit a reduced accuracy in correspondence of the low values while some *slopes* exhibit a decreased accuracy in correspondence of the low and high values. This is explained by the difficulty of accurately modelling a few spectra with the adopted model (see equation 4).

2) Retrieval of the a_{NPPM} regressional parameters: The subset of three inputs identified as the most appropriate for the retrieval of the a_{NPPM} regressional parameters is the same as that identified for a_{CDOM} (i.e., R_{rs} at 412, 490 and 665 nm). The accuracy in retrieving the a_{NPPM} regressional parameters using the selected R_{rs} subset, instead of the full set of R_{rs} inputs, did not show appreciable differences. The retrieval of the regressional parameters for a_{NPPM} showed uncertainty values of 28%, 8% and 43% for the *coefficient*, the *slope* and the *intercept*, respectively (Table II). The uncertainty of the *coefficient* and the *slope* parameters computed for a_{CDOM} are respectively 7% lower and 4% higher than the corresponding values obtained for a_{NPPM} (see Table I and Table II).

IV. Results and Discussion

This section presents and discusses the results obtained in retrieving a_{CDOM} and a_{NPPM} by applying the proposed MLP regional algorithms. Similarly to the analysis presented for the regressional parameters, the accuracy in retrieving a_{CDOM} and a_{NPPM} at specific wavelengths is here discussed using the absolute values of the relative percentage difference between measured and computed values.

Before separately presenting the accuracy analysis for a_{CDOM} and a_{NPPM}, it is relevant

	Measured		Computed	
	S_{CDOM}	S_{ad}	S_{CDOM}	S_{ad}
μ	0.0182	0.0122	0.0182	0.0122
σ	0.0041	0.0015	0.0019	0.0008

TABLE III. Statistical values of a_{CDOM} and a_{NPPM} *slope* regressional coefficients in units of nm^{-1}, where μ is the average value and σ is the standard deviation.

to highlight that no appreciable correlation (i.e., $r^2 = 0.01$) was observed between the experimental *slope* parameters of a_{CDOM} and a_{NPPM}, and this was also shown by the corresponding values computed with the MLP algorithms (i.e., $r^2 = 0.10$). Additionally, the computed S_{CDOM} and S_{NPPM} exhibit the same mean values as the corresponding quantities derived from the experimental data, respectively, (even though the former show a variability reduced by a factor of 2 (see Table III)). This, in addition to the validation results presented in panels (a.2) and (b.2) of figure 2, support the capability of the MLP algorithms, whereas based on the same set of input quantities, to discriminate between the CDOM and NPPM absorption spectral forms.

A. Spectral CDOM absorption coefficients

The uncertainty in retrieving a_{CDOM} from the regressional coefficients computed with the MLP is presented at 412, 443 and 490 nm (i.e., at those wavelengths where the a_{CDOM} values are more significant) through scatter plots of computed versus measured values (see figure 3). It can be observed that the uncertainty significantly increases with wavelength (from 20% at 412 nm to 47% at 490 nm). At the considered wavelengths, the more scattered data (generally exhibiting overestimated values) correspond to those characterized by low absorption values. This result can be explained by observing that a_{CDOM} exponentially

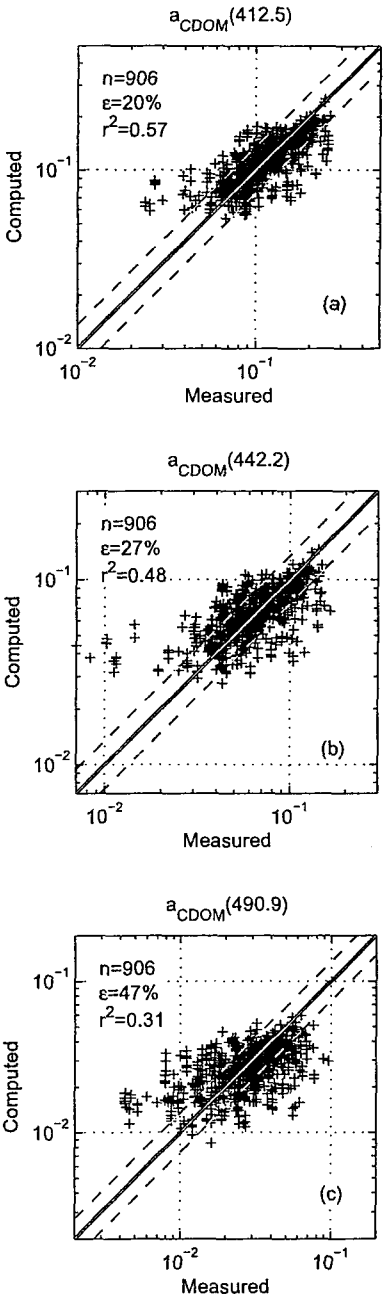


Fig. 3. Scatter plots of computed versus measured a_{CDOM} at 412, 443 and 490 nm (panels (a), (b), and (c), respectively) in units of m^{-1}. The computed a_{CDOM} were derived applying the regressional coefficients calculated using the MLP with R_{rs} at 412, 490 and 665 nm. The dashed lines indicate the ±35% uncertainty threshold.

decreases with wavelength and that the regressional parameters can be better retrieved when a_{CDOM} more significantly influences the R_{rs} spectra (that is for high absorption).

Aiming at independently assessing the quality of the a_{CDOM} values retrieved with the proposed MLP, an empirical band-ratio algorithm was produced with the same data set used for the development of the neural network algorithm. The empirical algorithm giving a_{CDOM} at 412 nm and based on the use of R_{rs} at 443 and 510 nm, as already proposed for the retrieval of CDOM absorption (Kahru and Mitchell, 2001), is

$$a_{CDOM}(412) = 10^{\{-1.136-1.387[log10\frac{R_{rs}(443)}{R_{rs}(510)}]\}} \quad (6)$$

and it was determined with $r^2 = 0.37$. The comparison of $a_{CDOM}(412)$ from the empirical algorithm with the corresponding value from the data set shows average uncertainties $\varepsilon = 24\%$ with $r^2 = 0.40$. By using equation 6 with a constant *slope* $S_{CDOM} = 0.0182$ (equal to the average value characterizing the data set) and , the a_{CDOM} values at 443 and 490 nm were retrieved with uncertainties of 32% and 57%, respectively. These values are higher than those presented in figure 3 and their differences increase with wavelength. This result shows that, regardless of the low determination coefficient characterizing the retrieval of S_{CDOM} (see figure 2), the MLP algorithm performs appreciably better than the considered band-ratio algorithm.

B. Spectral NPPM absorption coefficients

The uncertainty in retrieving a_{NPPM} using the regressional coefficients determined with the MLP, was estimated at 412, 555 and 665 nm (i.e., at wavelengths representative of the whole spectral region of interest). At 412 nm the uncertainty is comparable to that computed for a_{CDOM} (i.e., 25% versus 20%). It can also be observed that the uncertainty in retrieving a_{NPPM}, varying from 25% at 412 nm to 32% at 665 nm, presents a lower dependence on wavelength than that of a_{CDOM}. This is still explained by *i.* the more pronounced spectral

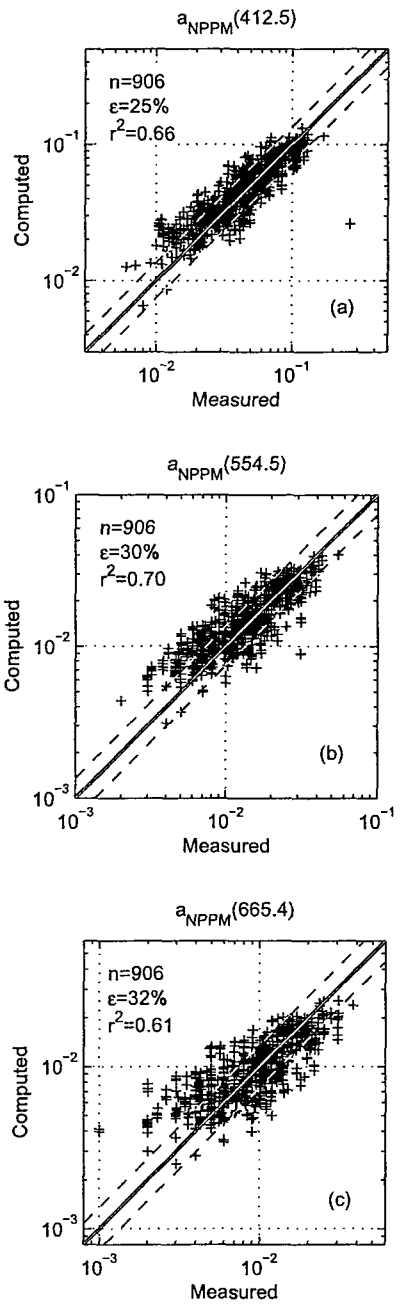


Fig. 4. Scatter plots of computed versus measured a_{NPPM} at 412, 555 and 665 nm (panels (a), (b), and (c), respectively) in units of m^{-1}. The computed a_{NPPM} were derived applying the regressional coefficients calculated using the MLP with R_{rs} at 412, 490 and 665 nm. The dashed lines indicate the $\pm 35\%$ uncertainty threshold.

dependence of a_{CDOM}, *ii.* the fact that at $\lambda > 650$ nm, a_{CDOM} exhibits values at the noise level while a_{NPPM} can still be appreciable (see figure 1), and finally *iii.* the lower uncertainty characterizing the retrieval of S_{NPPM} than S_{CDOM}.

Similarly to the case of a_{CDOM}, aiming at independently assessing the quality of the a_{NPPM} values retrieved with the proposed MLP, an empirical band-ratio algorithm was produced with the same data set used for the development of the neural network algorithm. The empirical algorithm giving a_{NPPM} at 412 nm and based on the use of R_{rs} at 510 and 665 nm, already proposed for the retrieval of the total suspended matter (Berthon et al., 2002), is

$$a_{NPPM}(412) = 10^{\{-0.355-1.177[log10\frac{R_{rs}(510)}{R_{rs}(665)}]\}} \tag{7}$$

and it was determined with $r^2 = 0.68$. The comparison of a_{NPPM} (412) from the empirical algorithm with the corresponding value from the data set, shows average uncertainties $\varepsilon = 29\%$ with $r^2 = 0.59$.

By using equation 7 with constant *slope* $S_{NPPM} = 0.0122\ m^{-1}$ and *background* $cb_{NPPM} = 0.0086\ m^{-1}$ (equal to the average values characterizing the data set), the a_{NPPM} values at 555 and 665 nm were retrieved with uncertainties of 42% and 49%, respectively. Similarly to a_{CDOM}, also these uncertainty values are appreciably higher than those presented in figure 4 and thus also support the better performance of the MLP with respect to the considered band-ratio algorithm.

V. Summary and Conclusions

Two MLP regional algorithms for the northern Adriatic Sea coastal waters were developed to retrieve, from the seawater remote sensing reflectance R_{rs}, the parameters of the functions defining the spectra of CDOM and NPPM absorption coefficients, a_{CDOM} and a_{NPPM}, respectively. The MLP algorithms, consisting of ten hidden neurons, were trained with experimental data from the CoASTS data set.

The study showed that:

- The accuracy in retrieving the parameters of the exponential function describing the a_{CDOM} and a_{NPPM} spectral dependence (i.e., *coefficient*, *slope*, and the *background* for NPPM) weakly decreases by reducing the number of inputs. Accurate results, compared to the complete set of inputs, can still be obtained using R_{rs} at three wavelengths with the most appropriate MLP configuration relying on R_{rs} at 412, 490 and 665 nm for both a_{NPPM} and a_{CDOM}.
- The MLP permits to better determine the *coefficient* of the exponential function describing a_{CDOM} than the coefficient of the function describing a_{NPPM}. Differently, the *slope* can be slightly better determined for a_{NPPM} than for a_{CDOM}. The retrieval of the *background* for a_{NPPM} shows the highest uncertainty.
- The uncertainty in retrieved CDOM and NPPM absorption coefficients increases with wavelength and exhibits average values of 20% and 25% at 412 nm, respectively.
- The effectiveness of the MLP algorithms is supported by statistical uncertainties of CDOM and NPPM absorption coefficients determined with regional band-ratio algorithms. These, developed with the same data set used for training the neural networks, show average uncertainties appreciably larger and increasing at a higher rate with wavelength, than those of the MLP algorithms.

It is finally recalled that this study only aimed at presenting the effectiveness of a methodology focused on the use of R_{rs} for the determination of the regressional parameters that describe the absorption coefficient spectra of CDOM and NPPM. Then, because of the regional origin of the applied experimental data set, the extendibility of the results to a different area needs to be proven or supported by similarity in the seawater type characterizing the applied data set and the target area.

Chapter 6

Range of applicability of the empirical ocean colour algorithms

This chapter deals with the problem of identifying the range of applicability of empirical ocean colour algorithms. The results of the study are presented through a paper submitted to *IEEE Transactions in Geoscience and Remote Sensing*. The references of this work are included in the thesis reference list for completeness.

The relevance of this analysis relies on the fact that optically complex regions require the implementation of specific regional algorithms. But, the range of applicably of regional algorithms can not be defined on purely geographical basis. Moreover, a simple geographical identification would easily introduce discontinuities in final product maps. The range of applicability of the empirical algorithms was here investigated on the basis of the statistical representativeness (expressed through the novelty detection method (Bishop, 1994), see Section 4.2) of the data used for the algorithm development. An extension of the novelty detection was also proposed to blend different algorithms avoiding discontinuities at the boundaries of their specific regions of applicability.

The quantity considered in this study was the *chlorophyll-a* concentration. The applicability of the MLP algorithm derived from the CoASTS data set was compared with that of a second algorithm derived from a subset of the SeaBAM data (O'Reilly et al.,

1998), and representative of Case 1 waters only. The study on the applicability of both algorithms was applied to the northern Adriatic Sea (i.e., where the CoASTS data set was collected) and other major European seas.

Use of the Novelty Detection Technique to Identify the Range of Applicability of Empirical Ocean Colour Algorithms[1]

Abstract

Novelty detection is used to identify the range of applicability of empirical ocean colour algorithms. This method is based on the assumption that the level of accuracy of the algorithm output depends on the representativeness of inputs in the training data set. The effectiveness of the novelty detection method is assessed using two data sets: one representative of the northern Adriatic Sea coastal waters and the other representative of open sea waters. The two data sets are independently used to develop neural network algorithms for the retrieval of chlorophyll a (*Chl-a*) concentration. The range of applicability of the individual algorithms is presented using remote sensing data derived from the Sea-viewing Wide-Field-of view Sensor (SeaWiFS) for three selected regions: the central Mediterranean Sea, the North Sea and the Baltic Sea. An extension of the novelty detection technique is also proposed to blend the individual algorithms and to avoid discontinuities in the resulting *Chl-a* maps.

Index Terms

Neural network, bio-optical modelling, ocean colour.

[1]The content of this Section was submitted to *IEEE Transactions in Geoscience and Remote Sensing* by Davide D'Alimonte, Frédéric Mélin, Giuseppe Zibordi, Jean-François Berthon, in Genuary 2003 and accepted for publication. The core theme of the paper is the retrieval of the phytoplankton component from R_{rs} data using MLP algorithms. The ideas, development and writing up of the paper were principal responsibility of the thesis author. The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into a based team research.

I. INTRODUCTION

OCEAN colour algorithms use upwelling radiance spectra (input) to predict one or more properties of the optically significant seawater components (output). They can be grouped into two main categories: i) analytical algorithms, based on the inversion of a forward radiance model, and ii) empirical algorithms, derived from a data set representative of the conditions to be modeled. The development of an empirical algorithm can be seen as a regression problem: once the functional form of the algorithm is chosen, the parameters of the function are derived from a set of input-output pairs (training data set). Training data can be generated with a theoretical model (synthetic data) or can be obtained from field measurements (*in situ* data). Synthetic data are obviously not subject to measurement uncertainties and can represent a virtually unlimited range of conditions, but their capability to represent actual conditions should always be assessed. On the other hand, *in situ* data intrinsically represent actual conditions, but the products computed from the derived algorithm are only reliable when the input data are similar to those used for the algorithm development. Therefore, the definition of the range of applicability of algorithms developed on the basis of *in situ* data is a task of primary importance and represents the objective of the present study.

Various empirical ocean colour algorithms built on *in situ* data have been developed for the retrieval of the chlorophyll a concentration (*Chl-a*). Some have been proposed for global applications, (e.g., Aiken et al., 1995; Clark, 1999; Gordon et al., 1983; Kishino et al., 1997; O'Reilly et al., 1998; OReilly et al., 2000) and have been specifically addressing the open ocean waters with various degrees of success (Gordon et al., 1983; Hooker and McClain, 2000). These waters, characterized by upwelling radiances depending on phytoplankton and its covarying constituents of biological origin, are designated as Case 1 (Morel and Prieur, 1977). However, the accuracy of these algorithms often appears low for coastal waters and

more in general for those waters designated as Case 2 (Morel and Prieur, 1977) where optically significant constituents do not co-vary. The latter situations have suggested the development of local empirical algorithms (e.g., Berthon et al., 2002; Kahru and Mitchell, 1999; Siegel et al., 1994; Tassan, 1994) generally restricted to specific regions.

So far, remote sensing applications rely on global or local algorithms on a geographical basis, that is in conflict with the dynamics of the water masses and can result in large discontinuity in boundary regions. And, whereas a scheme for the identification of different oceanographic regions on the basis of a fuzzy logic approach was investigated (Moore et al., 2001), no operational attempt has been yet presented in order to identify the range of applicability of empirical algorithms developed using quality assured data representative of specific water types.

Here, the *novelty detection* Bishop (1994) is proposed to identify the range of applicability of empirical algorithms as an alternative to geographic criteria. This method assumes that the reliability of the algorithm products depends on how well the input is represented within the training data set. The range of validity of the algorithm is then identified by first modeling the unconditional probability density function (*pdf*) of the training data and then by excluding those inputs (i.e., *novel data*) for which the *pdf* is lower than a predefined threshold value.

The study focuses on the determination of the *Chl-a* as a function of the remote sensing reflectance (R_{rs}), which is defined as the ratio of the upwelling radiance and the downwelling irradiance measured just above the sea surface, and that can be derived from satellite data. The study is carried out using two experimental data sets, one representative of the northern Adriatic Sea coastal waters (Zibordi et al., 2002), the other representative of open ocean waters (O'Reilly et al., 1998). Here are identified as coastal waters those which can be permanently or occasionally considered Case 2 because of bottom resuspension, coastal erosion, river inputs, or by simple anthropogenic impact.

The two data sets have been separately used to develop neural network algorithms for the retrieval of *Chl-a*. Then, the range of applicability of the individual algorithms is presented for three selected regions (i.e., the central Mediterranean Sea, the North Sea and the Baltic Sea) using maps of R_{rs} derived from the Sea-viewing Wide-Field-of view Sensor (SeaWiFS) (Hooker and Esaias, 1993). An extension of the novelty detection technique is finally proposed to blend the individual algorithms avoiding discontinuities in the resulting *Chl-a* maps.

II. Data Sets

The following sections present the two independent *in situ* data sets and the satellite imagery archive used in this study.

A. Open ocean data set (SeaBAM)

The first data set is obtained from the SeaBAM database (O'Reilly et al., 1998) that was used for the development of the OC2v2 algorithm (O'Reilly et al., 1998). It contains 919 coincident *in situ Chl-a* and R_{rs}, at different wavelengths, collected during various cruises and at different locations. The concentrations of *chlorophyll-a* have been obtained by fluorometric or High Performance Liquid Chromatography (HPLC) measurements. The use of these data is here restricted to the stations i) not considered as pertaining to Case 2 waters (as labelled in O'Reilly et al. (1998)), and ii) for which the remote sensing reflectances at 490, 555, and 665 nm are available (exceptions are 166 spectra for which the R_{rs} at 665 is not given and the R_{rs} at 683 is instead used). The remaining 631 pairs of *Chl-a* and R_{rs} spectra are representative of open ocean waters with *Chl-a* values comprised in the range $0.025 - 28$ mg m^{-3}, with mean and standard deviation equal to 0.8 and 2.9 mg m^{-3}, respectively (Figure 1(a)). This data set, as well as the associated neural network based algorithm, will be hereafter referred to as SeaBAM.

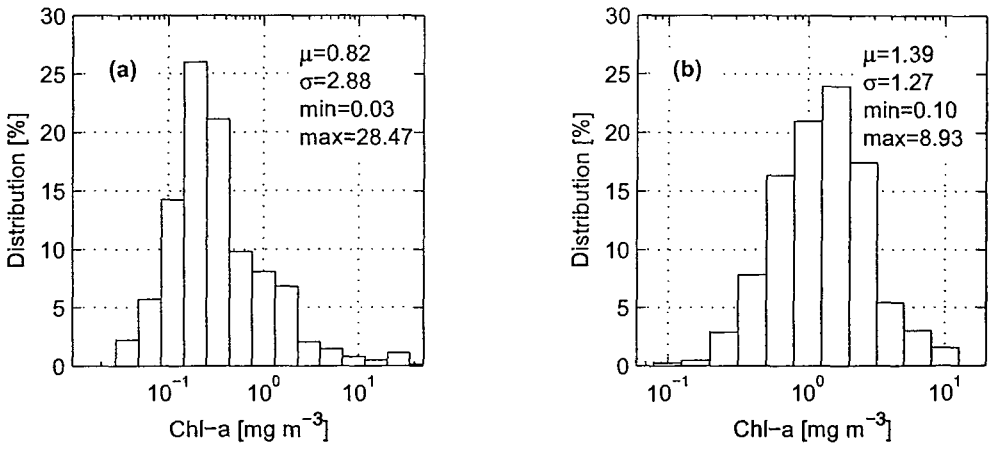


Fig. 1. Distribution of the *Chl-a* for the SeaWiFS and the CoASTS data sets ((a) and (b), respectively): μ indicates the mean value; σ is the standard deviation; m and M the minimum and the maximum *chlorophyll-a* concentrations.

B. Adriatic Sea data set (CoASTS)

The CoASTS data set has been collected at the Acqua Alta Oceanographic Tower (AAOT) in the northern Adriatic Sea in the framework of the CoASTS (Coastal Atmosphere and Sea Time Series) sampling program started in 1995 (Zibordi et al., 2002). The dynamics of the measurement site, located 8 nautical miles southeast of the Venice Lagoon (latitude 45°19′N, longitude 12°30′E), lead most of the CoASTS data set to be close to the separation threshold suggested in Loisel and Morel (1998) for Case 1 and Case 2 waters (with one third of the measurements strictly pertaining to Case 2 (Berthon et al., 2002)).

In-water measurements of upwelling radiance, upward and downward irradiance from the Wire-Stabilized Profiling Environmental Radiometer (WiSPER) system (Zibordi et al., 2002) are used to derive the R_{rs} at the center wavelengths corresponding to the SeaWiFS visible channels (412, 443, 490, 510, 555, and 665 nm instead of 670 nm). *Chl-a* has been obtained through HPLC technique (Targa et al., 2000). The concentrations are in the interval 0.1 –

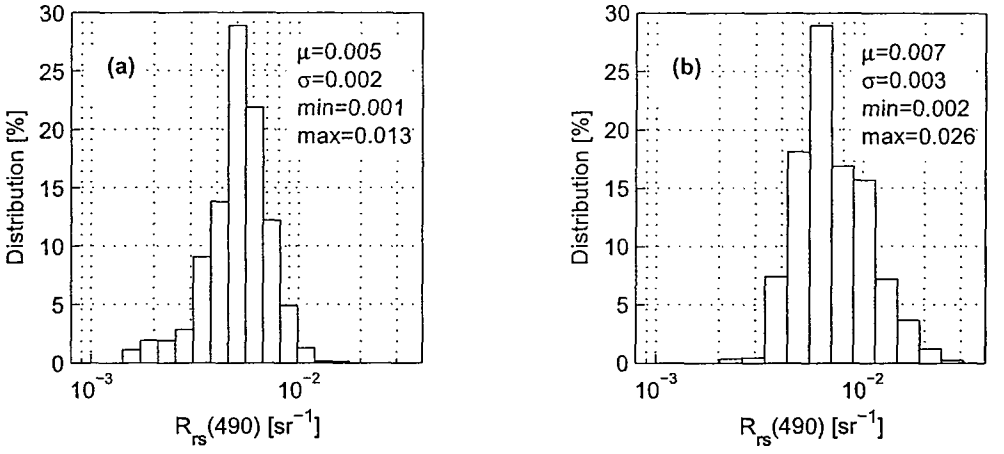


Fig. 2. Distribution of the R_{rs} at 490 nm for the SeaWiFS and the CoASTS data sets ((a) and (b), respectively): μ indicates the mean value; σ is the standard deviation; m and M the minimum and the maximum R_{rs} values at 490 nm.

9 mg m^{-3}, with mean and standard deviation equal to 1.4 and 1.3 mg m^{-3}, respectively (Figure 1(b)).

The data set applied in the study is composed of 456 *Chl-a* measurements. Each *Chl-a* is associated with one or more successive R_{rs} spectra. This association is only justified by the fact that repeated optical profiles always produce slightly different R_{rs} spectra because of environmental perturbations (sea state effects, changes in the water column and illumination conditions during data collection). Because of this the intrinsic noise characterizing pairs of *Chl-a* and R_{rs} having the same *Chl-a* value, is assumed to support the MLP learning process in better discriminating the signal from environmental perturbations. The total number of pairs of *Chl-a* and corresponding R_{rs} spectra is 906. This data set, as well as the associated neural network algorithm, will be hereafter referred to as CoASTS.

C. Satellite data set

The satellite products used in this study are SeaWiFS derived monthly maps of R_{rs} for the European seas. Monthly averages have been chosen to ensure a complete geographical coverage of the considered areas. For each SeaWiFS scene, the remote sensing reflectance is obtained by applying the atmospheric correction scheme described in Sturm and Zibordi (2002), and experimentally assessed in M´elin et al. (2003). The scenes are then combined in space and time following M´elin et al. (2002).

III. METHODS

This section describes the development of the neural network algorithms for deriving the *Chl-a* and the definition of their range of applicability through the novelty detection technique. A method for blending different algorithms is also formulated.

A. MLP algorithms for the Chl-a retrieval

The Multi Layer Perceptron (MLP) neural network can be defined as a parameterized distributed algorithm consisting of a network of simple non-linear processing units and "may be viewed as a practical vehicle for performing a non-linear input-output mapping of general nature" (Haykin, 1998). MLP have been extensively used to solve classification and function regression tasks in several research fields, including ocean colour (Buckton et al., 1999; Gross et al., 2000; Keiner and Brown, 1999; Shiller and Doerffer, 1997). An ocean colour application based on Generalized Radial Basis Function model can be found in Cipollini et al. (2001). The general principles of the MLP as well as the specific methods applied below are presented in several textbooks (Bishop, 1995; Duda et al., 2001; Ripley, 1996).

The MLP algorithms to derive *Chl-a* from R_{rs} have been implemented by means of the Netlab toolbox for Matlab (Nabney, 2001). The adopted MLP consisted of one layer of hidden

neurons, a sigmoid activation function for the hidden layer and a linear activation function for the output layer. In fact, this configuration can represent any continuous input-output mapping (Haykin, 1998). Methods applied for *1)* the data preprocessing, *2)* to improve the algorithm generalization capability and *3)* to assess the algorithm performance, are hereafter presented.

1) Pre-processing: The pre-processing of the input and output data consisted in standardizing the logarithm of R_{rs} (at each wavelength separately) and of *Chl-a*, i.e., subtracting the mean value and dividing by the standard deviation. In fact in agreement with results from a different study (Campbell, 1995), both the SeaBAM and CoASTS data set display a lognormal distribution of the *Chl-a* (Figure 1) and the R_{rs} (Figure 2).

2) MLP complexity optimization: The algorithm generalization capability has been optimized: *a)* selecting the proper set of input quantities, *b)* applying a weight regularization term and *c)* testing different MLP architectures.

a) Inputs selection: The set of MLP inputs was restricted to R_{rs} at 490, 555 and 665 nm. R_{rs} at 490 and 555 nm are commonly used for the determination of *Chl-a* in the open ocean (Maritorena and O'Reilly, 2000; O'Reilly et al., 1998; OReilly et al., 2000), while R_{rs} at 665 nm has been shown relevant for deriving concentrations of optically significant constituents like total suspended matter Berthon et al. (2002); Tassan (1994) or *Chl-a* itself (Siegel, 1991). A specific analysis (D'Alimonte and Zibordi, 2003) has also shown that the use of the R_{rs} at 412 or 443 nm, in addition to the set of R_{rs} at 490, 555 and 665 nm, does not generally improve the retrieval of *Chl-a*. The use of the 665 nm center wavelength may suggest caution due to the difficulty in producing accurate R_{rs} in the red part of the solar spectrum. However, a recent study (M´elin et al., 2003) on SeaWiFS products relying on in situ data from the AAOT coastal site, has shown the possibility of determining the normalized water leaving radiance (an consequently R_{rs}) at 670 nm with an average accuracy comparable to that obtainable at the other SeaWiFS center wavelengths.

b) Regularization term: Initial results showed some variability (especially when the number of hidden neurons was increasing) of the algorithm performance for subsequent retraining of the same MLP architecture. In order to reduce this variability, a weight decay regularization term was added to the error function (Bishop, 1995). The principle of the weight regularization functioning is to penalize high curvatures in the regressional function. Different weight decay coefficients (α) were tested in various trial and error exercises and the value $\alpha = 0.01$ was finally adopted.

c) Architecture selection: Instead of applying any general rule based on the size of the training data set, the MLP architecture was here identified verifying the performance of different configurations with a minimum of a single neuron up to a maximum of 20 in the hidden layer. For both SeaBAM and CoASTS algorithms it was observed that the improvement of the MLP performance was becoming negligible with approximatively ten hidden neurons, and thus this configuration was adopted. The resulting number of free parameters that ensures the MLP generalization capability is relatively high with respect to the size of the training data set. This can be explained by the use of the weight regularization term and the characteristics of the training data set.

3) MLP performance assessment: The selection of the MLP architecture was based on the assessment of the performance of each algorithm through the *cross-validation* method, by partitioning each data set in three segments. The ensemble of the data of two segments was used for the training process while the resulting algorithm was applied to the data of the excluded segment (see Figure 3). This process was iterated three times, and the computed *Chl-a* was derived from all available data. Note that the CoASTS data set was partitioned taking into account that one water sample can correspond to successive optical profiles (see Section II-B). To ensure the independence of the validation part with respect to the training data, all the data from the same water sample were included into a unique segment.

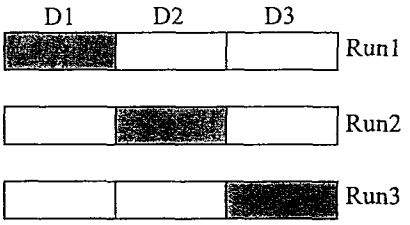


Fig. 3. Schematic illustration of the partitioning of the data set into 3 segments (D1,,D3) for use in cross-validation. The MLP is developed (run) 3 times, each time using the data of two segments (shown in white), while those of omitted segment (shown shadowed) are used to assess the algorithm performance (adapted from Bishop, 1995)

The algorithm capability in retrieving the *Chl-a* was evaluated in terms of absolute percentage difference ε between the measured (t) and the modeled (y) *Chl-a*:

$$\varepsilon = 100\frac{1}{N}\sum_{k=1}^{3}\sum_{i=1}^{N_k}\frac{|y_k^i - t_k^i|}{t_k^i} \tag{1}$$

where the indices k and i indicate the validation segment and the sample in the validation segment, respectively; N_k is the number of samples in the validation segment k; and N is the total number of samples in the data set. The assessed uncertainty in retrieving *Chl-a* is 24% and 31% for the SeaBAM and CoASTS algorithms, respectively. The scatter plots of modeled versus *in situ* measured *Chl-a* for the SeaBAM and CoASTS algorithms are given in Figure 4(a) and 4(b), respectively (the corresponding distribution of the *in situ* measured *Chl-a* was provided in Figure 1).

The cross-validation technique, applied to assess the MLP performance for the model selection, creates three different algorithm versions (one for each run, Figure 3) for each MLP architecture. For the analysis it was preferred to apply a single MLP algorithm instead of combining the three MLP algorithm versions. Thus, a new training for the final MLP algorithm version was performed on each entire data set with the optimal MLP architecture

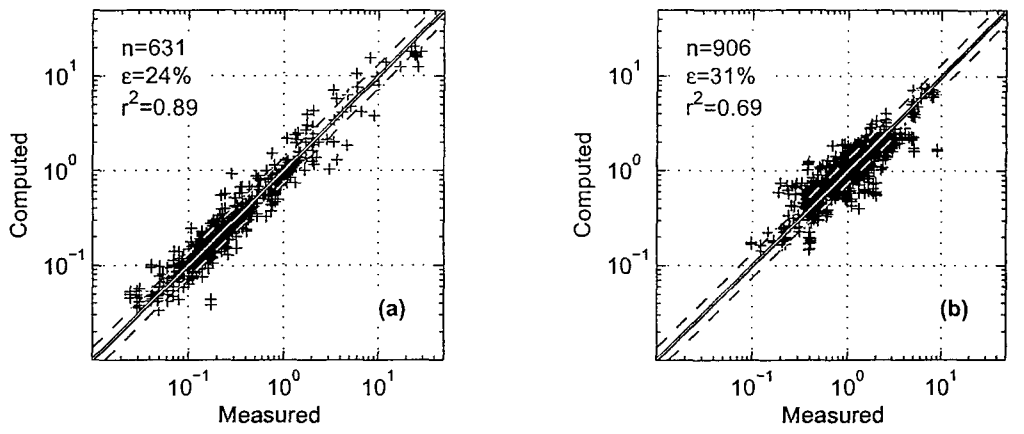


Fig. 4. Scatter plot of modelled versus *in situ* measured *Chl-a* [mg m^{-3}] for the SeaWiFS and the CoASTS MLP algorithms ((a) and (b), respectively): n indicates data set size; ϵ is the absolute percentage difference between modelled and measured values and r^2 is the coeffi cient of determination. The dashed lines represent the uncertainty of ±35% with respect to the 1 to 1 line.

previously defined.

The accuracy of the final algorithm developed on the basis of the entire data set is likely to be better than that assessed with the cross-validation process. To estimate the difference, the algorithm developed with all the available data was verified using the training data set itself. The actual accuracy of the final algorithm (i.e., that developed using all the available data) is between that assessed with the cross validation and that resulting applying the algorithm to the same data used for its implementation. This exercise showed that the performance of the final algorithm can be at most 2% better than that assessed with the cross-validation.

B. Novelty detection

The novelty detection is a candidate method to identify those inputs for which the performance of the algorithm is likely to be unreliable (Bishop, 1994). The basic assumption of this method is that the algorithm performs correctly when the input data are represented in

the training data set. Conversely, significant uncertainties can result from inputs substantially different from those used to set the algorithm parameters. In order to define the range of applicability of the algorithm, the training data set is first used to model the unconditional probability density function (*pdf*) of the input data. Then, during the operational use of the algorithm, those input data for which the *pdf* is lower than a threshold are considered novel, and thus out of the range of applicability.

The first step in identifying novel data is to separately model the unconditional distribution of the algorithm inputs for each training data set. As stated earlier, the distribution of R_{rs} was verified to be lognormal. Thus, a multivariate Gaussian function was used to represent the distribution of the logarithm of R_{rs} at 490, 555 and 665 nm:

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{2}$$

where $\mathbf{x}$ indicates the input vector formed by the logarithm of R_{rs} at 490, 555 and 665 nm; $\boldsymbol{\mu}$ and Σ are the mean value and the covariance matrix of the experimental data, respectively; d is the dimensionality of the input space; || and T represent the matrix determinant and transpose, respectively. Once the unconditional distribution of the input data is modeled, the next step is to set the *pdf* threshold that defines the range of applicability.

Validation exercises showed that the algorithm accuracy was not only depending on the unconditioned density of the training data but also on the type and concentration of the various optically significant components (i.e., phytoplankton component, coloured dissolved organic matter and inorganic suspended matter). It was here assumed that the algorithm accuracy, in a first approximation, can be considered the same over each entire experimental data set (see Figure 4(a) and 4(b)). Then, it was adopted the heuristic scheme of identifying the range of applicability of the empirical algorithm with that region of the input space containing 90% of the training data. It is pointed out that this simplified classification scheme was aiming at

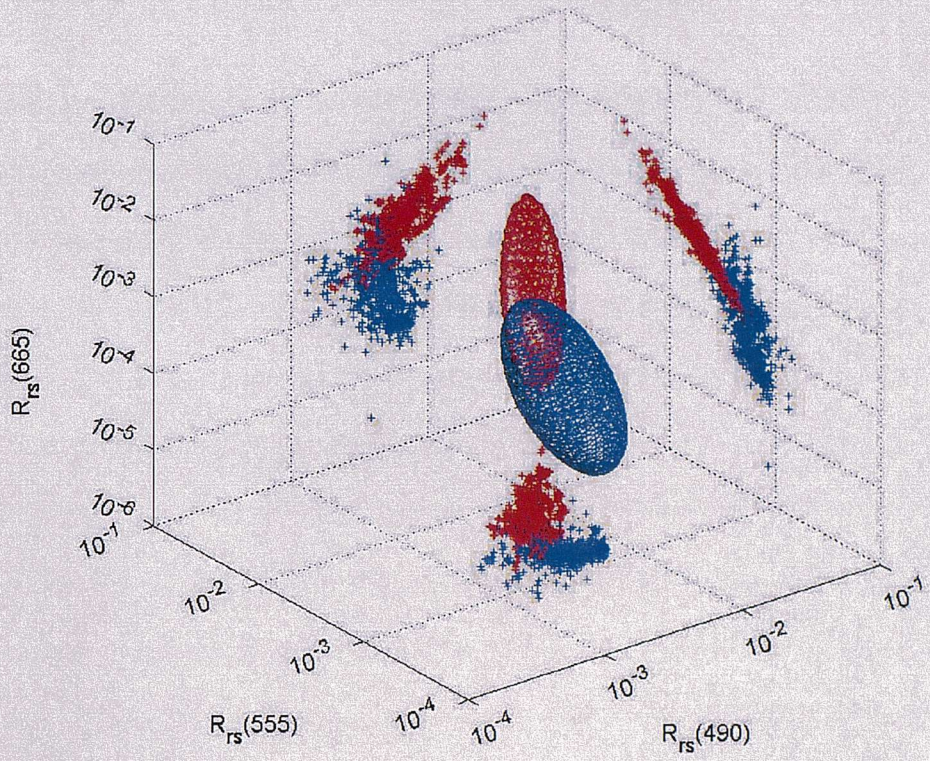


Fig. 5. Three-dimensional representation of the range of applicability of the SeaWiFS and CoASTS algorithms. Each axis of the plot corresponds to one of the algorithm input quantity (R_{rs} at 490, 555 and 665 nm). Both the boundary of the range of applicability (hyperellipsoid) as well as the projection of the data onto the principal plane are coloured in blue and red to represent the SeaWiFS and COASTS data sets, respectively (see text).

dividing the input space into two regions: one representing those data for which the algorithm performance corresponds to that assessed through the cross-validation and one where the algorithm products are likely to be unreliable.

By modeling the distribution of the input quantities with a multivariate Gaussian function, the contours of constant probability density are hyperellipsoids in which the Mahalanobis distance, $\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$, is a constant. Figure 5 highlights, through hyperellipsoids and the projection of the two data sets onto principal planes, the three-dimensional representation of the range of applicability of the SeaBAM and CoASTS algorithms. Each axis

of the plot corresponds to one of the algorithms input quantities (R_{rs} at 490, 555 and 665 nm). On the basis of the *pdf* threshold criterion previously defined, the range of applicability of each algorithm corresponds to that region of the input space for which the Mahalanobis distance from the center of the training data set distribution is less than 6.25 (Press et al., 1992). The implementation of the novelty detection method, analogously to the MLP development, was carried out using the Netlab software (Nabney, 2001).

It is pointed out that the *pdf* threshold was not applied to identify and remove any outliers from the training data sets. In fact, this work is based on experimental quality assured data and all points have been assumed to have equal validity and represent a true condition. The exclusion of some training data on the basis of a *pdf* threshold would have meant not only losing precious real cases but also a possible reduction of the algorithm accuracy close the boundaries of the region of applicability.

C. Algorithm blending

Once the range of applicability are separately defined for the SeaBAM and CoASTS data sets, some of the inputs may be found non-novel for both algorithms. In this case the subsequent method for *Chl-a* estimation is proposed:

$$y = \frac{y_{CoASTS}(\mathbf{x}) \cdot pdf_{CoASTS}(\mathbf{x}) + y_{SeaBAM}(\mathbf{x}) \cdot pdf_{SeaBAM}(\mathbf{x})}{pdf_{CoASTS}(\mathbf{x}) + pdf_{SeaBAM}(\mathbf{x})} \tag{3}$$

where $\mathbf{x}$ is again the triplet of the logarithm of R_{rs} at 490, 555 and 665 nm; y is the computed *Chl-a*; *pdf* is the probability density of finding $\mathbf{x}$ within the training data set, and the subscripts indicate the data set used to estimate the probability density function or its associated algorithm for *Chl-a* retrieval. This blending approach gives more weight to the algorithm for which the probability of finding the input data in the training data set is higher and ensures the shift from one algorithm to another without introducing discontinuities. The proposed blending approach is straight forward. However, it doesn't represent the optimal solution because the

blending coefficients are not optimized to minimize the difference between the modeled (i.e., as resulting after the blending) and the measured values. A possible alternative approach to overcome this limit would be to address the algorithm blending with a Mixture of Expert (ME) approach (Bishop, 1995; Haykin, 1998) as discussed in Section V-B.2.

IV. Results

The novelty detection method and the algorithms for *Chl-a* determination have been applied to monthly maps of remote sensing reflectance for three regions – the central Mediterranean Sea, the North Sea and the Baltic Sea. It is stressed that the use of a monthly map of reflectance for the classification of water type and calculation of *Chl-a* serves only an illustrative purpose. Ideally, the proposed steps should be applied to daily imagery, but this simplification does not affect the interpretation of the results presented and discussed below.

Chl-a maps resulting from the individual CoASTS and SeaBAM algorithms are presented over the entire European seas investigated to fully highlight the differences resulting from the application of the two algorithms. The region of applicability of the individual algorithms and the difference between the products of the SeaBAM and CoASTS algorithms are then provided in a separate map. Differently, the *Chl-a* derived from data novel for both SeaBAM and CoASTS data sets is not presented on the maps of the blended products and the corresponding region is flagged (i.e., coloured of gray). This was done to avoid considerations that may result too speculative, being only derived from novel data where both *pdf* tend to zero.

A. Central Mediterranean

The first area considered in this study is the central part of the Mediterranean Sea in April 1999 (Figure 6). It contains the CoASTS site in the northern Adriatic Sea, and also includes regions typical of Case 1 conditions, with low *Chl-a*, particularly in the Ionian and Tyrrhenian

basins, or with high *Chl-a*, during the spring bloom in the Ligurian Sea (Antoine et al., 1995; Bricaud et al., 2002; Morel and Andr´e, 1991).

The maps showing the applicability of algorithms (Figure 6(a) and 6(c)) display an almost complementary classification. The CoASTS data are found representative of the northern Adriatic Sea, of the coastal zones and of the waters transported southeast of the Sicilian coast. About 6% of the area shown is characterized by reflectances represented in the CoASTS data set. Conversely, almost all the offshore waters are represented by the SeaBAM data set, including the bloom conditions of the Ligurian basin (Figure 6(b)). The associated range of applicability covers 94% of the area. This distribution supports the fact that reflectance spectra associated with a very low *Chl-a* (i.e., below 0.1 $\mathrm{mg\,m^{-3}}$) or bloom conditions are not observed at the AAOT site, whereas coastal zone conditions are not represented in the considered SeaBAM data. The maps of *Chl-a* resulting from the two separate MLP algorithms are shown in Figure 6(b) and 6(d), whereas Figure 6(e) highlights the differences between the two maps. The CoASTS algorithm gives values of *Chl-a* lower than the SeaBAM algorithm for the coastal regions and for the Ligurian Sea bloom, and higher values in low *Chl-a* offshore waters (e.g., southern Adriatic Sea, Tyrrhenian Sea). The discrepancy between the *Chl-a* maps derived from the two algorithms can be quantified in an average overestimation by the CoASTS algorithm of about 26% in the central Mediterranean Sea. Obviously, these differences stem from the respective characteristics of the two training data sets. Moreover, it can be noticed that the MLP algorithms used in this work did not predict *Chl-a* values much out of the range of variability of the training data sets. An example of this is illustrated by the absence of low *Chl-a* in the map resulting from the CoASTS algorithm.

Through a weighting based on the *pdf*, the blending of the algorithm products combines these two maps (Figure 6(f)). It is observed that some areas very close to the Adriatic coast south of the Po estuary are not identified by any of the classification schemes (grey regions

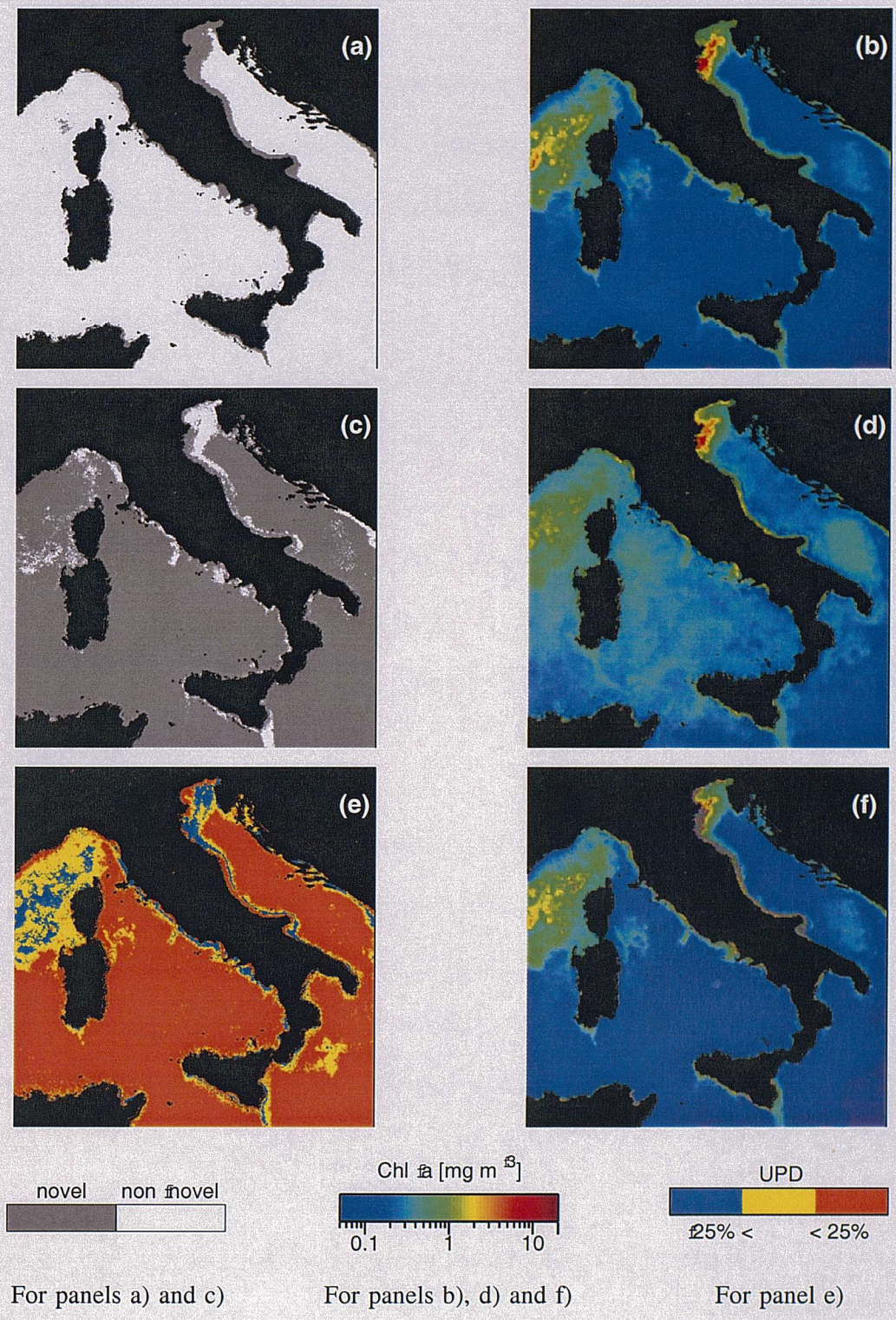


Fig. 6. Central Mediterranean Sea. Range of applicability of the SeaBAM (a) and CoASTS (c) algorithms. The *Chl-a* maps resulting from the use of the two algorithms are shown in b) for SeaBAM and d) for CoASTS. In e) the Signed Unbiased Percentage Difference (UPDS) in the *Chl-a* between the two algorithms is shown, with $UPDS = 200 * (y_{CoASTS} - y_{SeaBAM})/(y_{CoASTS} - y_{SeaBAM})$ where y indicates the algorithm result while the subscription indicates the corresponding training data set. In f) the *Chl-a* map resulting from the SeaBAM and CoASTS blending approach (regions containing data out of the region of applicability of both algorithms have been flagged in gray) is shown.

on Figure 6(f)). For instance, close to the Po river outlet, the spectra present values of the reflectance at 490 nm that are relatively low with respect to the training data sets, whereas the reflectances at 555 and 670 nm are relatively high. Actually, the AAOT site is not under appreciable influence of the Po river discharge and is not characterized by very high absorption values in the blue part of the spectrum or by very high scattering values (Berthon et al., 2002). In general, it is noticed that the areas classified as novel are among regions of high values of the backscattering coefficient (see Mediterranean map presented in Loisel et al. (2001)). The overall area of applicability of the blended algorithms indicates that 96% of the central Mediterranean Sea in April 1999 is represented in at least one of the two data sets. The resulting *Chl-a* map is very similar to that predicted by the SeaBAM algorithm except for the coastal zones and particularly the northern Adriatic Sea.

B. North Sea

Figure 7 shows similar maps for the North Sea and adjacent areas (April 1999). As for the Central Mediterranean, a complementarity in the range of applicability of the two data sets is observed (Figure 7(a) and 7(b)). It is interesting to see how the CoASTS reflectance data are found representative of most of the Irish Sea, the English Channel and the German Bight (27% of the marine area). The classification indicates that the area between Scotland and Norway as well as the Atlantic waters are well represented in the SeaBAM data set (71% of the marine grid points).

When compared to the *Chl-a* map derived from the SeaBAM algorithm (Figure 7(c)), the map obtained from the CoASTS algorithm (Figure 7(d)) generally shows higher concentrations in the Atlantic and North of the German Bight and lower concentrations in waters presumably pertaining to Case 2, for example along the Dutch and German coasts. Figure 7(e) shows the differences between the latter maps and highlights an overall underestimation in the North

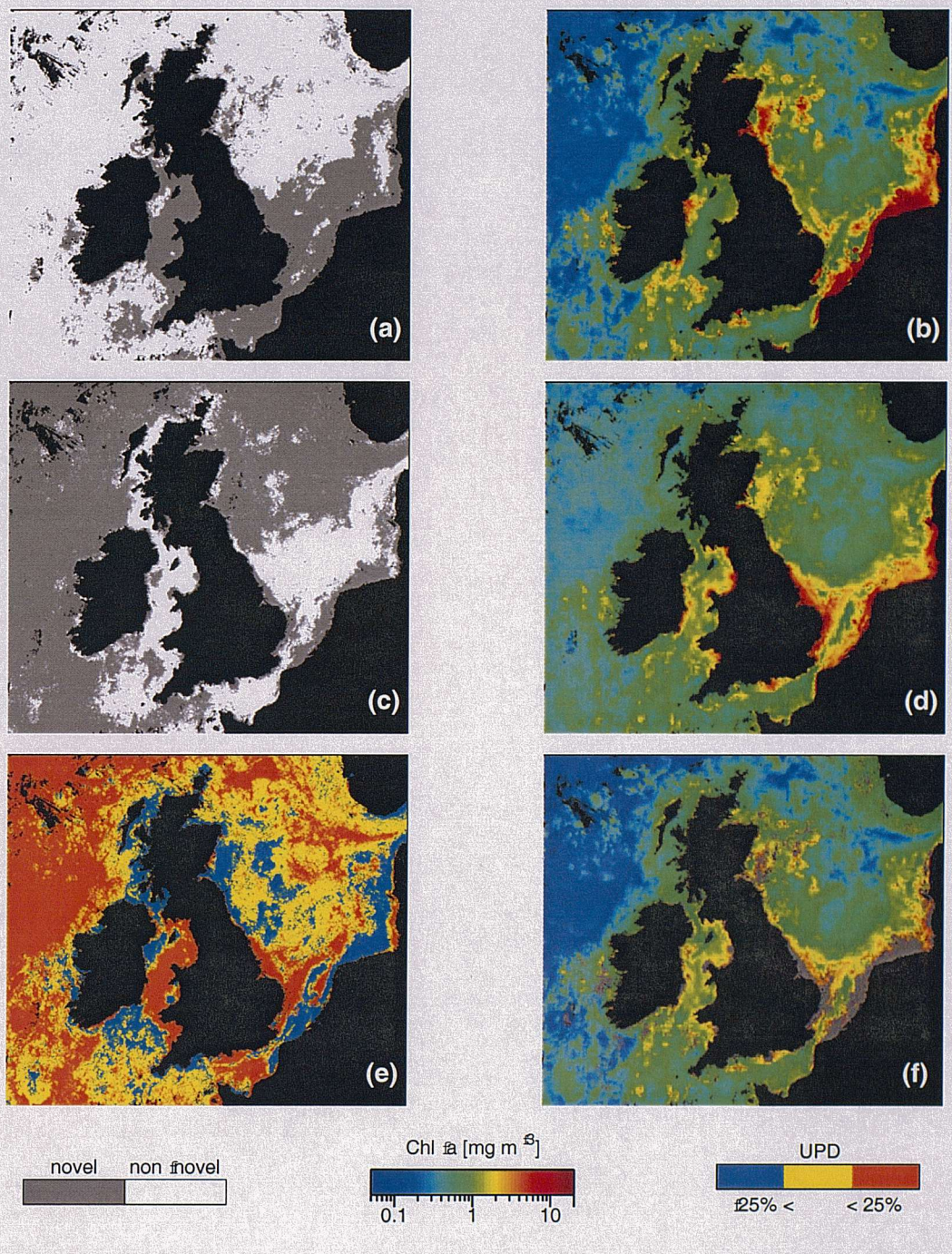


Fig. 7. As in figure 4 for the North Sea.

Sea of about 17% of the CoASTS based *Chl-a* distribution with respect to that predicted by SeaBAM.

Figure 7(f) shows the output of the blending of the algorithm products and highlight in gray those regions out of the range of applicability of both algorithms. The reflectance spectra are represented by at least one of the training data sets for 89% of the map. However, the coastal regions of southeast England and of the German Bight, characterized by high concentrations of dissolved and particulate matter (Aarup et al., 1996; Doerffer and Fischer, 1994), are identified as novel data (areas in grey). For these regions, the reflectance at 490 nm appears relatively low and the reflectance at 670 nm relatively high with respect to the training data sets. Those areas are influenced by river discharge (Thames River on one hand, Elbe, Ems, Rhine delta, Weser on the other) and resuspension of bottom sediments due to tidal currents in shallow waters (Jago et al., 1994).

C. Baltic Sea

For the Baltic Sea, the month of July 1999 is retained as an example because it maximizes the availability of ocean colour data. Compared to the other two areas, the analysis in the Baltic Sea shows the highest fraction of novel data. For example, in the Baltic Sea proper, the number of marine grid points represented in at least one data set is clearly insignificant (Figure 8(b) and 8(c)). Thus, the application of the present empirical algorithms to this area and period of the year would result in a high uncertainty. This restricted range of applicability is explained by a value of R_{rs} that is relatively low mostly at 490 nm but also at 555 nm with respect to the training data sets. This basin is indeed characterized by high absorption values in the blue particularly due to the presence of yellow substance (Aarup et al., 1996; Ferrari et al., 1996; Jerlov, 1976; Kowalczuk, 1999; Siegel et al., 1994).

These results do not hold all year round. Parts of the basin can be confidently classified in

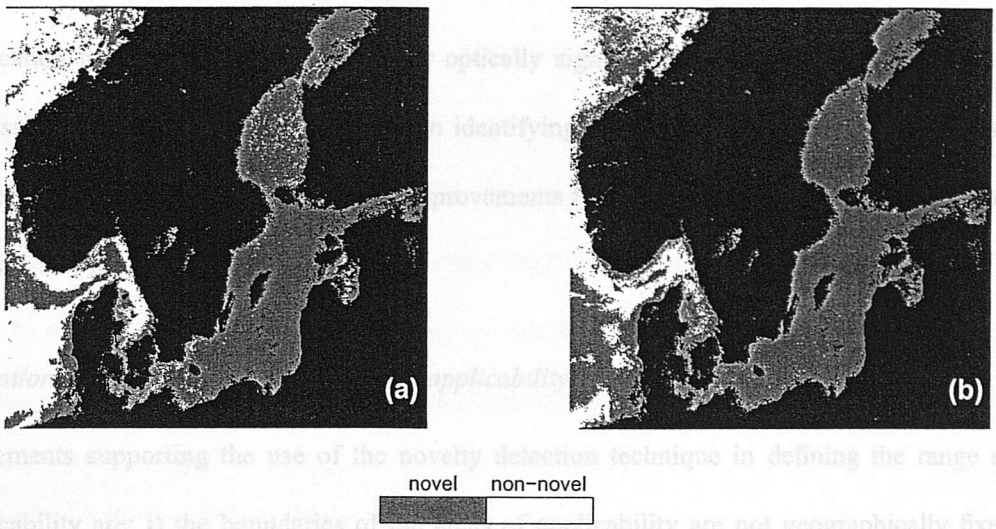


Fig. 8. Region of applicability of the SeaBAM, a), and CoASTS, b), algorithms for the Baltic Sea. None of the two algorithms has been identifi ed as applicable in this region.

some months, especially the southern sector in spring (March, April), even if the global range of applicability of the algorithms remains limited (not shown). The temporal variability of the representativeness of the two selected *in situ* data sets should be studied but this analysis goes beyond the scope of this work. In any case, the inclusion of *in situ* data collected in the Baltic Sea would allow for the extension of the range of applicability of the proposed empirical algorithms.

V. Discussion

The accuracy of the *Chl-a* maps obtained with the proposed approach relies on: i) the accuracy of the *in situ* and remote sensing data used to develop the algorithms and produce the maps, respectively; ii) the existence of ideal and unique relationships linking *Chl-a* to R_{rs} for different marine areas; iii) the assumption of negligible uncertainties induced in R_{rs} data by the atmospheric correction of remotely sensed radiances. The following discussion will then be restricted to elements linked with the general concept of novelty detection and its

application to remote sensing of seawater optically significant constituents. Specifically the discussion will address: *A)* the rationale in identifying the algorithm applicability by means of the novelty detection and *B)* areas of improvements for future activities which may deserve focused study.

A. Rationale in identifying the algorithm applicability with the novelty detection

Elements supporting the use of the novelty detection technique in defining the range of applicability are: i) the boundaries of the areas of applicability are not geographically fixed and follow the changes of the water properties so that both seasonal and spatial variations as well as occasional events can be dealt with, ii) the areas for which new data should be included in the training data sets in order to more efficiently extend the range of applicability of the algorithms are highlighted and iii) the individual range of applicability of the CoASTS and SeaBAM algorithms can be associated with the corresponding uncertainty of the computed *Chl-a* (see section III-A.3). The latter outcome could not be achieved if the two data sets were merged to create a single algorithm.

An alternative approach to identify the range of applicability would be to accept only those algorithm outputs that are included within the range of variability of the training data. This approach can be clearly misleading especially if the algorithm functional form can model non-linear relationships as it is the case for the adopted MLP. In fact, with this class of functions, saturation may characterize intervals close to the limits of the range of variability of the training data, and the algorithm output could erroneously be considered correct.

An example of what is stated can be found in the maps of the central Mediterranean Sea given in section IV (Fig.4). The *Chl-a* derived using the CoASTS algorithm for the Ligurian basin bloom is well included in the range of variability of the associated training data set, but this area is not part of the range of applicability of the algorithm. On the other hand, the

SeaBAM algorithm predicts higher values of *Chl-a* for this bloom and the confidence in the validity of these results is high.

The use of the novelty detection to define the range of applicability avoids dealing with algorithm saturation. It is pointed out that other MLP architectures or other functional forms could be used to reduce the saturation trend outside the range of applicability. Nevertheless, this would not increase the confidence in the algorithm products derived from novel data.

B. Areas of improvement for future activities

1) Additional elements for the identification of the range of applicability of algorithms: In the present work, the distribution of the R_{rs} of the training data sets was modeled using a single multivariate Gaussian function. The water types classification (see Section IV-A and IV-B) resulting from the proposed scheme using the SeaBAM and CoASTS data sets corresponds to the expected (i.e., as suggested by the prior knowledge) distribution of coastal and open ocean waters. And this supports the validity of the proposed classification scheme. Nevertheless a single Gaussian may not provide accurate description of the distribution of the training data. In this case, a better identification of the range of applicability could be obtained by modeling the *pdf* through a Gaussian Mixture Model (GMM) (Nabney, 2001).

Validation exercises (see Section III-B) showed that the algorithm accuracy not only depends on the density of the training data but also on the type and concentration of the various optically significant components. Thus, an alternative approach for estimating the range of applicability of the empirical algorithms would be to directly express the dependence of the algorithm accuracy as a function of the R_{rs} data. Then, estimates of the algorithm accuracy could be obtained i) developing an MLP for which the input is the R_{rs} and the output is the difference between computed and measured *Chl-a* or ii) implementing in a Bayesian framework a single algorithm that provides both the *Chl-a* as well as its accuracy.

Chapter 7

Summary and conclusions

This work has addressed the problem of the retrieval of the optically active seawater components in coastal waters and the range of applicability of coastal regional algorithms. The relevance of these objectives depends on any limitation of the operational algorithms for Case 1 waters when applied in optically complex coastal waters, due to the reduced correlation between seawater components. Algorithms for coastal areas are likely to have only a regional validity, that can not be identified on the basis of a geographical partitioning but vary according to the specific characteristics of Case 2 waters (Sathyendranath, 2000). The relevance for effective regional[1] algorithms, together with a corresponding range of applicability, is motivated by a growing need to identify and monitor specific biological habitats in coastal waters, as well as by the importance of these areas for human activities. Outcomes of this work were based on *in situ* experimental data, and thus they provide a significant reference to evaluate the suitability of proposed algorithms for operational applications.

7.1 Summary of the major results

The results obtained from this study have been presented through three papers submitted to *IEEE Transactions in Geoscience and Remote Sensing.* The most relevant results

[1]The term regional is here used to indicate *in situ* data collected from specific coastal water types, and not the geographical separation of the seawater.

produced within this works are hereafter summarized.

7.1.1 Modelling of the phytoplankton component

The phytoplankton component has been modelled in terms of i) *chlorophyll-a* concentration and ii) absorption of the pigmented particulate matter at 443 nm. The study has shown that the MLP configuration most suitable for the retrieval of both *Chl-a* and $a_{ph}(443)$ is made by ten hidden neurons and requires R_{rs} at 490, 555 and 665 nm. The MLP algorithms have been compared with i) band ratio algorithms developed on the basis of the CoASTS data set, and ii) the OC4v4 algorithm, (O'Reilly et al., 2000) routinely used for the operational processing of the SeaWiFS data.

Among the algorithms applied to the CoASTS data, OC4v4 was the least successful (the unsigned relative percentage difference between the *Chl-a* computed with OC4v4 and the corresponding measured values was 60%). This highlights the limits of including some points from an optically complex coastal region in a large data set in the scope of developing more general algorithms. This is disconcerting as some of the data used for the implementation of the OC4v4 algorithms were taken from the CoASTS data set used in this study, and confirms the need to implement specific regional algorithms for optically complex coastal areas. The MLP has been shown more effective than band ratio algorithms in modelling both the *Chl-a* and the $a_{ph}(443)$. Finally, it is evident that $a_{ph}(443)$ can be better retrieved than *Chl-a* (with uncertainties of 22% and 31%, respectively).

7.1.2 Coloured dissolved organic matter and non pigmented particulate matter absorption coefficiens

The possibility of retrieving the absorption coefficient spectral shape of seawater optically significant components has been investigated on the basis of synthetic data by Hoge and Lyon (1996, 1999). In addition, various algorithms have been proposed (Tassan,

1994; Shiller and Doerffer, 1997; Buckton et al., 1999; Keiner and Brown, 1999; Kahru and Mitchell, 2001; Berthon et al., 2002) for the retrieval of CDOM and NPPM absorption coefficients (a_{CDOM} and a_{NPPM}, respectively) at specific wavelengths. However, these algorithms do not provide a description of the spectral dependence of a_{CDOM} and a_{NPPM}, which may vary from region to region, according to the specific bio-geochemical components of CDOM and NPPM (Ferrari and Tassan, 1999; Ferrari, 2000). In the present study, the CoASTS data set has been used to developed MLP regional algorithms to retrieve the spectra of a_{CDOM} and a_{NPPM}, and not just their values at some specific wavelength. Specifically, the proposed approach consisted in firstly defining the regressional parameters (i.e, the *coefficient*, *slope* and *background*) of the exponential functions fitting a_{CDOM} or the a_{NPPM} spectral measurements. Then, these regressional parameters have been related to the corresponding remote sensing reflectance measurements, through the MLP.

For both a_{CDOM} and a_{NPPM}, the proper set of MLP inputs was identified with the R_{rs} at 412, 490 and 665 nm. The study has shown that the *coefficient* of the exponential function describing a_{CDOM} can be better retrieved than that of the function describing a_{NPPM}. Conversely, the *slope* can be better identified for a_{NPPM} than for the a_{CDOM}. The retrieval of the *background* values of a_{CDOM} presented the highest uncertainty. These results have been explained by the exponential decrease of a_{NPPM} and a_{CDOM} with wavelength, and by the fact that the absorption of the coloured dissolved organic matter is almost negligible (i.e., close to the instrument noise level) in the red region of the spectra. The uncertainty on the retrieval of a_{CDOM} and a_{NPPM} values, using the regressional parameter computed by the MLP, are 20% and 25% respectively (412 nm), and increase with wavelength.

7.1.3 Algorithms applicability

This study addresses the problem of the identification of the range of applicability of the empirical ocean colour algorithms towards the use of the *novelty detection* technique

(Bishop, 1994). The principle of the novelty detection is that the reliability of the algorithm product depends on how well the input is represented within the training data set used for the algorithm implementation. The range of applicability of the algorithm is then identified by first modelling the unconditional probability density function (*pdf*) of the training data and then by excluding those inputs (i.e., *novel data*) for which the *pdf* is lower than a predefined threshold value.

It is important to highlight that the definition of the range of applicability of ocean colour regional algorithms is still an open problem. In fact, whereas a scheme for the identification of different oceanographic regions on the basis of a *fuzzy logic* approach was investigated by Moore et al. (2001), no operational attempt has yet been presented to identify the range of applicability of empirical algorithms developed through experimental data from specific water types.

The analysis of the algorithm applicability has made use of a subset of the SeaBAM data (O'Reilly et al., 1998) representative of Case 1 water conditions only, in addition to the CoASTS data (Zibordi et al., 2002a) mostly representative of conditions close to the separation line between Case 1 and Case 2 waters (Loisel and Morel, 1998). Two MLP algorithms for the *Chl-a* retrieval (i.e., one from each data set) have been implemented. Then, the range of applicability of each algorithm has been identified with the region of the input space (i.e., represented by R_{rs} at 490, 555 and 665 nm) containing 90% of the corresponding training data set.

The range of applicability of each algorithm has been analyzed for three selected regions (i.e., the central Mediterranean Sea, the North Sea and the Baltic Sea) using maps of R_{rs} derived from the SeaWiFS data. An extension of the novelty detection technique has been proposed to blend the individual algorithms, thus avoiding discontinuities in the resulting *Chl-a* maps.

For the considered monthly maps (April, 1999), it has been observed that the overall region of applicability (i.e., data non novel for at least one algorithm) covers most of the Central Mediterranean Sea and the North Sea. Data not included in the range of

applicability have been mostly found in coastal waters and in areas influenced by river discharge (Thames, Elbe, Ems, Rhine, Weser for the North Sea and the Po river for the North Adriatic Sea). Conversely, the Baltic Sea (monthly map of July 1999), has been recognized to be mostly characterized by data novel for both algorithms. This, highlights the need of collecting and using additional *in situ* measurements from the Baltic Sea.

The study also suggests that the accuracy of the algorithms, beside of the density of the input data, is also dependent on additional factors such as the the type and concentration of the specific optically significant components.

7.2 Accomplishment of the study objectives

This study has demonstrated the suitability of the MLP for the implementation of regional algorithms developed on the basis of data collected in optically complex coastal regions. Various authors (Shiller and Doerffer, 1997; Keiner and Brown, 1999; Buckton et al., 1999; Gross et al., 2000; Cipollini et al., 2001; Dransfeld, 2003) investigated the potentiality of neural networks in developing Case 2 waters algorithms, but these studies were primarily based on synthetic data only. Thus, a main achievement of the present work is to assess the suitability of a limited set of experimental data collected in optically active complex waters for the MLP development. The study also highlighted the relevance of adopting techniques to better exploit the available data and to optimize the algorithm generalization capability (i.e., *cross-validation*, *complexity optimization* and *weight decay* regularization term). In light of the results obtained, it is seen that the use of the MLP, as a complementary approach to the band ratio polynomial regression for the implementation of empirical ocean colour algorithms, is of benefit. This is also supported by the availability of effective routine libraries and advanced optimization algorithms (Nabney, 2001) for the MLP implementation.

The range of applicability of the empirical algorithms has been investigated, for the specific case of the *chlorophyll-a* concentration, through the novelty detection technique

Bishop (1994). The novelty detection technique has also been extended to blend different algorithms avoiding discontinuity between the boundaries of individual regions of applicability. Identified elements supporting the proposed scheme for the determination of the range of applicability of algorithms are: i) the boundaries of the regions of applicability are not geographically fixed and they follow the changes of water properties so that both seasonal and spatial variations as well as occasional events can be dealt with, ii) the areas where new data should be included in the training data sets, to more efficiently extend the range of applicability of the algorithms, are highlighted and iii) the accuracy of the computed *Chl-a* maps (as resulting from the algorithm validation process) can be directly associated to the range of applicability of the corresponding algorithm. The novelty detection has been recognized to be an effective approach to evaluating the algorithm applicability. Thus, both the validity of the results obtained, as well as the simplicity of its implementation, suggest the operational use of the novelty detection.

7.3 Future work

The present study has investigated effective approaches for better exploiting *in situ* data collected in optically complex coastal regions. The scientific area of the current research is both vast and complex. It is inevitably the subject of continuing research, some aspects of which are discussed hereafter.

7.3.1 The empirical versus the analytical approach

Whereas O'Reilly et al. (1998) already investigated the effectiveness of some analytical and semi-analytical models, showing that the empirical ones were providing more accurate results, nevertheless relevant issues leaved open by the present study are i) to extend the comparison between the neural network models here developed and other analytical solutions derived from the CoASTS data set, and ii) to extend such a comparison to the retrieval of optically active water components other than the pigmented

part (i.e., CDOM and NPPM). The relevance of such a study is that the analytical solution is based on a physical model of the underlying radiative transfer process, while the empirical algorithms do not depend on any physical assumption regarding the data generating process.

An element that may partially justify the higher accuracy shown, up to now, by the empirical solution is the complexity of the analytical approach with respect to the empirical one. In fact, the analytical approach requires i) to introduce some simplification in order to make the forward model computationally tractable, ii) to identify a set of parameters to characterise the model on the basis of a set of experimental measurements (and in such a case the model is also called *semi-analytical*) and finally iii) to identify an effective numerical approach for the inversion of the forward model. Each step inevitably introduces some additional uncertainty that contributes to reduce the accuracy of the final solution. Another factor possibly limiting the application of the analytical approach is that the inversion of the forward model has to be carried on a pixel by pixel basis, and this can be time consuming when processing large remote sensing images. Efforts to improve the accuracy of the analytical solution have to address all the here above mentioned items.

It is once more pointed out that, whereas less straight forward, the analytical solution has to be in principle preferred because it is part of a theoretical framework in which data are used to validate hypotheses aiming to explain a natural process. Contrary, empirical solutions such as the band ratio polynomial regression or the neural network based methods, even when more accurate, only allow to optimally exploit the statistical information content embodied in the data but without improving the physical understanding of the data generating process.

7.3.2 Algorithms assessment with data from other oceanographic areas

The study has addressed the applicability of empirical algorithms without investigating aspects related to the accuracy of atmospheric corrections in other coastal regions (the

processing scheme applied to the SeaWiFS images, proposed by Sturm and Zibordi (2002), was assessed by Mélin et al. (2003) for the northern Adriatic Sea). Moreover, the application of ocean colour algorithms developed with the CoASTS data to areas other than the Northern Adriatic Sea presumes the uniqueness of the relationship between R_{rs} and the optically active seawater components. These are key aspects to be verified with experimental data from other coastal regions before relying on algorithm products. High quality *in situ* measurements collected in coastal European waters to be used for further validation exercises are represented by the COLORS[2] and the COASTlOOK[3] data sets.

7.3.3 Reduction of the dimensionality of the input space

The identification of the lowest dimension of the input space still ensuring the algorithm performance, is motivated by the problem of the *curse of the dimensionality* and by the robustness of the resulting algorithm (see Section 4.1.5.1). In this study, the reduction of the dimensionality of the input space has been based on a trial-error approach. From a minimum set of inputs (R_{rs} at some wavelengths), additional inputs (R_{rs} at wavelengths unused in the minimal input set) have been added one at a time, and that giving an effective improvement of the algorithm performance has been retained for the next step. This approach allows a direct use of prior knowledge on the relevance of the R_{rs} at some specific wavelengths with respect to the water components being modelled. It also provides information on the final set of relevant inputs. The latter outcome could not directly be obtained by applying methods for feature extraction based on projection (linear or non linear) of the initial input quantities on a low dimensional latent space. Nevertheless, among these projection methods, the Independent Component Analysis (ICA, see Hyvärinen et al. (2000)) seems to be well addressed to the problem of ocean colour. In fact, ICA was developed with the specific aim of separating the contribution

[2]Details on the COLORS data set can be found in the COLORS "Final Report", European Commission, contract MAS3-CT97-0087.

[3]Details on the COASTlOOK data set can be found in the COASTlOOK "Final Report", European Commission, contract ENV4-CT96-0310

of independent sources from the signal resulting from their combination. That is exactly what happens to the R_{rs} spectra due to the independent variability of the optically active components in Case 2 waters. As it has been already observed for the MLP, routine libraries[4] are available for the application of the ICA, simplifying their implementation.

7.3.4 Modelling of the algorithms accuracy

Further improvements in the determination of the algorithm applicability are expected by directly modelling the algorithm accuracy as a function of the input data. It has been observed that specific combinations of the individual seawater components are likely to constrain the algorithm performance, in spite of the statistical density of the training data. An approach to express the algorithm accuracy as a function of the input data is represented by the modelling of the distribution of the algorithms output in a Bayesian framework (Bishop, 1995). Alternatively, a MLP algorithm could be directly trained to learn the difference between measured and modelled values.

7.3.5 Modelling of the noise affecting input data

In this work, the development of the algorithms has been carried out without considering the effects of the additional uncertainty that affect actual remotely sensed data (i.e., mainly due to the atmospheric correction processing scheme). This additional uncertainty presumably affects both the accuracy of the algorithm products, as well as the effectiveness in identifying their range of validity. Also, the MLP is likely to be more sensitive to the uncertainty affecting the input quantities than regressional algorithms based on band ratios. In fact, the former use the R_{rs} at individual wavelengths, while the latter use band ratios, thus automatically reducing the effect of the uncertainty correlated at different wavelengths. Thus, a possible approach to improve the MLP effectiveness in operational applications would be to add to the training data additive noise having the same spectral distribution as that affecting actual remote sensing data.

[4]A MATLAB toolbox for the ICA is downloadable from http://www.cis.hut.fi/projects/ica

7.3.6 Range of applicability and algorithm blending

The modelling of the probability density of the input values of the training data set has been realized in the present work through a single multivariate Gaussian distribution. This is supported by the lognormal distribution of the bio-optical quantities (Campbell, 1995), and allows an immediate identification of the range of applicability of the algorithms through the Mahalanobis distance (see Chapter 6). The identification of the range of applicability could alternatively be obtained by modelling the *pdf* of the input data through a Gaussian Mixture Model (GMM) (Nabney, 2001). The advantage of this approach is that it does not need the assumption that the input data are Gaussian distributed. Nevertheless, the operational application of this approach is more costly in time, as it requires the evaluation of the probability density function for each pixel of the image to be classified.

Finally, the accuracy of the result obtained by the combination of different algorithms representative of specific water types could be improved through a Mixture of Expert (ME) approach (Bishop, 1995). The blending method developed in the present study as an extension of the novelty detection technique, is exclusively based on the statistical representativeness of the input R_{rs}. Conversely, the ME approach would define the mixing coefficient for the various algorithms through a minimization process. This should further reduce the difference between the result of the algorithm blending and the corresponding measured value.

References

Aarup, T., Holt, N., and Hjerslev, N. (1996). Optical measurements in the North Sea - Baltic transition zone. Statistical analysis of bio-optical data from the Eastern North Sea and the Skagerrak and the Kattegat. *Cont. Shelf Res.*, 16:1355–1377.

Aiken, J., Moore, G., Trees, C., Hooker, S. B., and Clark, D. K. (1995). *The SeaWiFS-CZCS type pigment algorithm.* SeaWiFS Technical Report Series. NASA Goddard Space Flight Center, TM-1995-104566, Greenbelt, MD.

Antoine, D., Morel, A., and André, J.-M. (1995). Algal pigment distribution and primary production in the Weastern Mediterranean as derived from coastal zone color scanner observation. *J. Geophys. Res.*, 100:16193–16209.

Berthon, J.-F., Zibordi, G., Doyle, J., Grossi, S., van der Linde, D., and Targa, C. (2002). *Coastal Atmosphere and Sea Time Series (CoASTS): Data analysis*, volume 20 of *SeaWiFS postlaunch Technical Report Series*, pages 1–25. NASA Goddard Space Flight Center, TM-2002-206892, Greenbelt, MD.

Bishop, C. M. (1994). Novelty detection and neural network validation. *IEEE Proc. Vision and Image & Sig. Proc.*, 141:217–222.

Bishop, C. M. (1995). *Neural networks for pattern recognition.* Clarendon.

Bricaud, A., Babin, M., Morel, A., and Claustre, H. (1995). Variability in the chlorophyll-specific absorption coefficient of natural phytoplankton : analysis and parametrization. *J. Geophys. Res.*, 100, C7:13321–13332.

Bricaud, A., Bosc, E., and Antoine, D. (2002). Algal biomass and sea surface temperature in the Mediterranean Basin. Intercomparison of data from various satellite sensors, and implications for primary production estimates. *Remote Sens. Environ.*, 81:163–178.

Bricaud, A., Morel, A., and Prieur, L. (1981a). Absorption by dissolved organic matter of the sea (yellow substance) in the UV and visible domain. *Limnol. Oceanogr.*, 26:816–832.

Bricaud, A., Morel, A., and Prieur, L. (1981b). Absorption by dissolved organic matter of the sea (yellow substance) in the UV and visible domain. *Limnol. Oceanogr.*, 26:816–832.

Buckton, D., O'Mongane, E., and Danaher, S. (1999). The use of neural networks for the estimation of oceanic constituents based on the MERIS instrument. *Int. J. Remote Sensing*, 20(9):1841–1851.

Bulgarelli, B. and Zibordi, G. (2002). Remote sensing of ocean color: accuracy assessment of an approximate atmospheric correction method. *Int. J. Remote Sens.* (Accepted for pubblication).

Campbell, J. W. (1995). The lognormal distribution as a model for the bio-optical variability in the sea. *J. Geophys. Res.*, 100:237–254.

Carder, K. L., Chen, F. R., and Cannizzaro, J. P. (2002). Case 2 chlorophyll a algorithm and Case 2 absorption coefficient algorithm. Technical report, MODIS Algorithm Theoretical Basis Document. Version 6.0.

Carder, K. L., Chen, F. R., Lee, Z. P., and Hawes, S. K. (1999). Semianalytic Moderate-Resolution Imaging Spectrometer algorithms for chlorophyll a and absorption with bio-optical domains based on nitrate-depletion temperatures. *J. Geophys. Res.*, 104:5403–5421.

Cherkassky, V. and Muller, F. (1998). *Learning from data.* Wiley-Interscience Publication.

Cipollini, P., Corsini, G., Diani, M., and Grasso, R. (2001). Retrieval of sea water optically active parameters from hyperspectral data by means of generalized radial basis function neural network. *IEEE Trans. Geosci. Remote Sensing*, 41:1508–1524.

Clark, D. K. (1999). Bio-optical algorithms: Case 1 waters. MODIS algorithm Theoretical Basis Document. Version 1.2:pp. 1–23.

D'Alimonte and Zibordi, G. (2001). *The JRC Data Processing System*, volume 15 of *SeaWiFS postlaunch Technical Report Series*, pages 52–56. NASA Goddard Space Flight Center, TM-2001-206892, Greenbelt, MD.

D'Alimonte, D., Mélín, F., Zibordi, G., and Berthon, J.-F. (2003). Use of the novelty detection technique to identify the range of applicability of empirical ocean colour algorithms. *IEEE Trans. Geosc. Rem. Sens.*, 41:2833–2843.

D'Alimonte, D. and Zibordi, G. (2003). Determination of CDOM and NPPM absorption coefficient spectra from coastal water remote sensing reflectances. *IEEE Trans. Geosc. Rem. Sens.* (Submitted).

D.D'Alimonte and G.Zibordi (2003). Phytoplankton determination in an optically complex coastal region using a multi layer perceptron neural network. *IEEE Trans. Geosc. Rem. Sens.*, 41:2861–2868.

Doerffer, R. and Fischer, J. (1994). Concentrations of chlorophyll and suspended matter and and gelbstoff in Case 2 waters derived from satellite coastal zone color scanner data with inverse modeling methods. *J. Geophys Res.*, 99:7457–7466.

Doyle, J. P. and Zibordi, G. (2002). Monte Carlo modeling of optical trasmission within 3-d shadowed field: Application to large deployment structure. *Appl. Opt.*, 41:4283–4306.

Dransfeld, S. (2003). *Comparison of neural networks with other ocean colour algorithms and effects of noise on reflectance band weighting.* PhD thesis, University Of Southampton.

Duda, O. R., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification.* John Wiley & Sons Inc.

Ferrari, G. (2000). The relationship between chromophoric dissolved organic matter and dissolved organic carbon in the European Atlantic area and in the West Mediterranean Sea (Gulf of Lions). *Marine Chemistry*, 106:339–357.

Ferrari, G. and Tassan, S. (1999). A method using chemical oxidation to remove light absorption by phytoplankton pigments. *J. Phycol.*, 35:1090–1098.

Ferrari, G. M., Dowell, M. D., Grossi, S., and Targa, C. (1996). Relationship between the optical properties of chromophoric dissolved organic matter and total concentration of dissolved organic carbon in the Southern Baltic Sea region. *Mar. Chemistry*, pages 299–316.

Garver, S. A. and Siegel, D. A. (1999). Inherent optical property inversion of ocean color spectra and its biogeochemical interpretation 1. time series from the Sargasso Sea. *J. Geophys. Res..*, 102:18,607–18,625.

Gordon, H. R. (1976). Radiative transfer: a technique for simulating the ocean in satellite remote sensing calculations. *Appl. Opt.*, 15:974–979.

Gordon, H. R., Brown, O. B., Evans, R. H., Brown, J. W., Smith, R. C., Baker, K. S., and Clark, D. K. (1988). A semianalytic radiance model of ocean color. *J. Geophys. Res.*, 93(10):909–924.

Gordon, H. R., Clark, D. K., Brown, J. W., Brown, O. B., Evans, R. H., and Broenkow, W. W. (1983). Phytoplankton pigment concentrations in the Middle Atlantic Bight: comparison of ship determinations and CZCS estimates. *Appl. Opt.*, 22:20–36.

Gross, L., Thiria, S., Frouin, R., and Mitchell, B. G. (2000). Artificial neural network for modeling the transfer function between marine reflectance and phytoplankton pigment concentration. *J. Geophys. Res.*, 105:3483–3495.

Halliday, D., Resnick, R., and Krane, K. S. (2001). *Physics*, volume 2. John Wiley & Sons Inc.

Haykin, S. (1998). *Neural Networks A comprehensive foundation.* Prentice-Hall.

Hoge, F. E. and Lyon, P. E. (1996). Satellite retrieval of inherent optical properties by linear matrix inversion of oceanic radiance models: an analysis of models and radiance measurment error. *J. Geophys. Res.*, 101:631–648.

Hoge, F. E. and Lyon, P. E. (1999). Spectral parameters of inherent optical property models: method for satellite retrieval by matrix inversion of an oceanic radiance model. *Appl. Optics*, 38:1657–1662.

Højerslev, N. K. (1998). *Spectral light absorption by gekbstoff in coastal waters displaying highly different concentrations.* Ocean Optics XIV, Office of Naval Research, Washington, DC.

Højerslev, N. K., Holt, N., and Aarup, T. (1996). Optic measurements in the North Sea-Baltic Sea transition zone. On the origin of the deep water in the Kattegat. *Cont. Shelf. Res.*, 16:1329–1342.

Hooker, S. B. and Esaias, W. S. (1993). An overview of the SeaWiFS mission. *EOS, Trans., Amer. Geophys. Union*, 74:241–246.

Hooker, S. B. and McClain, C. R. (2000). The calibration and validation of SeaWiFS data. *Prog. Oceanogr.*, 45:427–465.

Hyvärinen, A., Karhunen, J., and Oja, E. (2000). *Independent Component Analysis.* 2001 John Wiley & Sons.

Jago, C. F., Bale, A. J., Green, M. O., Howarth, M. J., Jones, S. E., McCave, I. N., Millward, G. E., Morris, A., Rowden, A. A., and Williams, J. J. (1994). *Resuspension processes and seston dynamics, southern North Sea.* Understanding the north Sea system. Chapman & Hall, London.

Jeffrey, S. W., Montoura, R. F. C., and Wright, S. (1997). Phytoplankton pigments in oceanography: Guidelines to modern methods. *UNESCO Publishing.*

Jerlov, N. G. (1976). *Marine Optics*, volume 14 of *Ocheanography.* Elsevier.

JGOFS (1994). Joint Global Flux Study, protocols for the JGOFS core measurements. *UNESCO Publishing*, 29:91–96.

Kahru, M. and Mitchell, B. G. (1999). Empirical algorithm and preliminary SeaWiFS validation for the California current. *Int. J. Remote Sens.*, 20:3423–3429.

Kahru, M. and Mitchell, B. G. (2001). Seasonal and nonseasonal variability of satellite derived chlorophyll and coloured dissolved organic matter concentration in the California current. *J. Geophys. Res.*, 106:2517–2529.

Keiner, L. E. and Brown, C. W. (1999). Estimating oceanic chlorophyll concentrations with neural networks. *Int. J. Remote Sensing*, 20:189–194.

Kirk, J. T. O. (1994). *Light & Photosynthesis in aquatic Ecosystems.* Cambridge University Press.

Kishino, M., Ishizaka, J., Saitoh, S., Senga, Y., and Utashima, M. (1997). Verification plan of ocean color and temperature scanner atmospheric correction and phytoplankton pigment by moored optical buoy system. *J. Geophys. Res.*, 102:17197–17207.

Kowalczuk, P. (1999). Seasonal variability of yellow substance absorption in the surface layer of the Baltic Sea. *J. Geophys. Res.*, 104:30047–30058.

Lee, Z. P., Carder, K. L., Marra, J., Steward, R. G., and Perry, M. J. (1996). Estimating primary production at depth from remote sensing. *App. Opt.*, 35:463–474.

Lee, Z. P., Carder, K. L., Mobley, C. D., Steward, R. G., and Patch, J. S. (1999). Hyperspectral remote sensing for shallow waters: 2. deriving bottom depths and water properties by optimization. *App. Opt.*, 38:3831–3843.

Loisel, H., Bosc, E., Stramski, D., Oubelkheir, K., and Deschamps, P.-Y. (2001). Seasonal variability of the backscattering coefficient in the Mediterranean Sea based on satellite SeaWiFS imagery. *Geophys. Res. Lett.*, 28:4203–4206.

Loisel, H. and Morel, A. (1998). Light scattering and chlorophyll concentration in Case 1 waters: a reexamination. *Limnol. Oceanogr.*, 43:847–858.

MacKay, D. J. C. (1994). *Bayesian Methods for backpropagation networks*, volume Models of Neural Networks, chapter 6. New York: Springer-Verlag.

Maritorena, S., Morel, A., and Gentili, B. (1994). Instrumental self-shading in underwater optical measurements: experimental data. *Limnol. Oceanogr.*, 39:689–703.

Maritorena, S. and O'Reilly, J. E. (2000). *OC2v2: update on the initial operational SeaWiFS chlorophyll a algorithm*, volume 11 of *SeaWiFS postlaunch Technical Report Series*, pages 3–8. NASA Goddard Space Flight Center, TM-2000-206892, Greenbelt, MD.

Mélín, F., C. Steinich, N. G., Pinty, B., and Verstraete, M. M. (2002). Optimal merging of LAC and GAC data from SeaWiFS. *Int. J. Remote Sens.*, 23:801–807.

Mélín, F., Zibordi, G., and Berthon, J.-F. (2003). Assessment of SeaWiFS atmospheric and marine products for the Northern Adriatic Sea. *IEEE Trans. Geosc. Rem. Sens.*, 41:548– 558.

Mobley, C. D. (1994). *Light and Water. Radiative Transfer in Natural Waters.* Academic Press.

Moore, S. T., Campbell, J. S., and Feng, H. (2001). A fuzzy logic classification scheme for selecting and blending satellite ocean colour algorithms. *to IEEE Trans. Geosc. Rem. Sens.*, 39:1764–1776.

Morel, A. (1973). *Diffusion de la lumiere par les eaux de mer. Résultats éxperimentaux et approche théorique*, pages 3.1.1–3.1.76. AGAR Lecture Series No. 61. Optics of the Sea (Interface and In-water Trasmission and Imaging). North Atlantic Treaty Organozation. Academic Press Inc., London.

Morel, A. (1974). *Optical properties of pure seawater*, pages 1–24. Optical Aspects of Oceanography, N. G. Jerlov and E. Steemann Nielsen (eds.). Academic Press Inc., New York.

Morel, A. (1988). Optical modeling of the upper ocean in relation to its biogenous matter content (Case 1 waters). *J. Geophys. Res.*, 93:10749–10768.

Morel, A. (1991). Light and marine photosynthesis : a spectral model with geochemical and climatological implications. *Prog. Oceanogr.*, 26:263–306.

Morel, A. and André, J. M. (1991). Pigment distribution and primary production in the Western Mediterranean as derived and modeled from coastal zone color scanner. *J. Geophys. Res.*, 96:12685–12698.

Morel, A. and Prieur, L. (1977). Analysis of variation in ocean colour. *Limnol. Oceanogr.*, 22:709–722.

Muller, J. L. and Austin, R. W. (1995). *Ocean optics protocols SeaWiFS for validation, Revision 1*, volume 25 of *SeaWiFS Technical Report Series*, page 66 pp. NASA Goddard Space Flight Center, TM-1995-104566, Greenbelt, MD.

Nabney, I. T. (2001). *Netlab: Algorithms for pattern recognition.* Springer Edition.

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver,

S. A., Kahru, M., and McClain, C. R. (1998). Ocean color chlorophyll algorithms for SeaWiFS. *J. Geophys. Res.*, 103:24937–24953.

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., Kahru, M., and McClain, C. R. (2000). Atmospheric correction for SeaWiFS imagery for turbid coastal and inland waters. *Appl. Opt.*, 39:897–912.

Pope, R. M. and Fry, E. S. (1997). Absorptium spectrum (380-700 nm) of pure water: II. integrating cavity measurements. *Appl. Opt.*, 36:8710–8723.

Press, W. H., Teukolsky, S., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C: the art of scientific computing.* Cambridge University Press.

Ripley, D. B. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

Robinson, I. S. (1994). *Satellite Oceanography.* Jhon Wiley & Sons.

Roesler, C. S. (1989). Theoretical and experimental approaches to improve the accuarcy of particulate absorption coeffcients derived from the quantitative filter technique. *Limnol. Oceanogr.*, 43:1649–1660.

Ruddick, K. G., Ovidio, F., and Rijkeboer, M. (2000). Atmospheric correction for SeaWiFS imagery for turbid coastal and inland waters. *Appl. Opt.*, 39:897–912.

Sathyendranath, S. (1989). A three component model of ocean colour and its application to remote sensing of phytiokankton pigments in coastal waters. *Int. J. Remote Sensing*, 10:1373–1394.

Sathyendranath, S. (2000). Remote sensing of Ocean Colour in coastal, and other optically-complex waters. International Ocean-Colour Coordinating Group, IOCCG Report Number 3.

Shiller, H. and Doerffer, R. (1997). Neural network for the evaluation of the inverse model - operational derivation of Case 2 water properties from MERIS data. *Int. J. Remote Sensing*, 20:1735–1746.

Siegel, H. (1991). Empirical algorithms for the determination of chlorophyll by remote sensing methods. *Beitr. Meereskd.*, 62:69–78.

Siegel, H., Gerth, M., and Beckert, M. (1994). *The variation of optical properties in the Baltic Sea and algorithms for the application of remote sensing data*, volume 2258, pages 894–905. Ocean Optics XII.

Smith, R. C. and Baker, K. S. (1981). Optical properties of the clearest natural waters (200-800 nm). *Appl. Opt.*, 20:177–184.

Stramsky, D. and Kiefer, D. A. (1991). Light scattering by microorganisms in open ocean. *Prog. Oceanogr.*, 28:343–383.

Strickland, J. D. H. and Parsons, T. R. (1972). *A practical handbook of sea water analysis*. Fish. Res. Board. Canada.

Sturm, B. and Zibordi, G. (2002). SeaWiFS atmospheric correction by an approximate model and vicarious calibration. *Int. J. Remote Sens.*, 23:489–501.

Targa, C., van der Linde, D., and Berthon, J.-F. (2000). *The first SeaWiFS HPLC Analysis Round-Robin Experiment (SeaHARRE-1) - The JRC method*, volume 14 of *SeaWiFS postlaunch Technical Report Series*, pages 27–29. NASA Goddard Space Flight Center, TM-2000-206892, Greenbelt, MD.

Tassan, S. (1994). Local algorithms using SeaWiFS data for the retrieval of phytoplankton and pigments, suspended sediment and yellow substance in coastal waters. *Appl. Opt.*, 33:2369–2378.

Tassan, S. and Ferrari, G. (1995). An alternative approach to absorption measurements of aquatic particles retained on filters. *Limnol. Oceanogr.*, 40:1358–1368.

Tassan, S. and Ferrari, M. (2004). Variability of light absorption by acquatic particles in the near infrared spectral region. *Appl. Opt.*, 42:4802–4810.

Zibordi, G., Berthon, J.-F., Doyle, J. P., Grossi, S., van der Linde, D., Targa, C., and Alberotanza, L. (2002a). *Coastal Atmosphere and Sea Time Series (CoASTS): A long-term measurement program*, volume 19 of *SeaWiFS postlaunch Technical Report Series*, pages 1–29. NASA Goddard Space Flight Center, TM-2001-206892, Greenbelt, MD.

Zibordi, G., D'Alimonte, D., and Berthon, J.-F. (2004). An evaluation of depth resolution requirements for optical profiling in coastal water. *Journal of Atm. and Ocean. Tech.* (Accepted for pubblication).

Zibordi, G. and Ferrari, G. (1995). Instrumental self-shading in underwater optical measurements: experimental data. *Appl. Opt.*, 34:767–779.

Zibordi, G., Hooker, S. B., Berthon, J.-F., and D'Alimonte, D. (2002b). Autonomous above-water radiance measurement from an offshore platform: A field assessment experiment. *Journal of Atm. and Ocean. Tech.*, 19:808–819.