THE SeaBAM EVALUATION DATA SET

Stéphane Maritorena[1], John E. O'Reilly[2] and Brian D. Schieber[3]

[1] Universities Space Research Association, NASA/GSFC, SeaWiFS Project, Code 970.2, Greenbelt, MD 20 771

[2] NOAA/NMFS, Narragansett Laboratory, Tarzwell Drive, Narragansett, RI 02882-1199

[3] General Science Corporation, NASA/GSFC, SeaWiFS Project, Code 970.2, Greenbelt, MD 20 771

Abstract

In order to evaluate and compare the performance of various Ocean Color chlorophyll *a* or pigment (Chl *a* + phaeo) algorithms, a large data set containing coincident *in situ* pigment and remote sensing reflectance measurements at the SeaWiFS wavelengths has been assembled from various sources. Methods used to obtain and process the data varied among investigators but all radiometric data were unified as "above surface" remote sensing reflectances (Rrs = Lw/Ed(0+)). Graphical methods as well as results from selected bio-optical models were used to identify and remove outlier stations. Because the phaeophytin information was missing for almost one half of the data, an equation relating chlorophyll a and pigment, derived from the SEABASS database, was applied to estimate the Chl *a* + phaeo concentration at stations were only chl *a* concentration was available.

Some data in the global data set do not match exactly the SeaWiFS wavebands. The major differences occur for stations having data at 520 nm instead of 510 nm and stations having Rrs(565) instead of Rrs(555). Because Rrs(555) and Rrs(565) are in a relatively "flat" region of the spectrum and are highly correlated, Rrs(565) were converted to proxy-Rrs(555). Because these conditions are not true for Rrs510 and Rrs520, no corrections were attempted on the 520 nm data. (The influence of these wavelengths mismatches on algorithm performance are analyzed and discussed in the chapter by O'Reilly and Maritorena.).

The final SeaBAM global data set contains 919 stations and encompasses a wide range of chlorophyll concentration. The SeaBAM data set is mostly made up of Case I non-Polar waters with chl *a* concentrations range between 0.019 and 32.79 ug/l.

Introduction

Since the Coastal Zone Color Scanner (CZCS) mission a variety of bio-optical algorithms have been developed to relate measurements of ocean radiance to *in situ* concentrations of phytoplankton pigments. Most of these algorithms were derived from relatively small sets of *in situ* measurements. For example, the NET (Nimbus Experiment Team) data set contains less than 60 stations. However, and despite the limitations of the CZCS, some of these algorithms have been applied widely to obtain valuable results on phytoplankton distribution and chlorophyll *a* concentration in the World Ocean (see e.g. Feldman *et al.*, 1989; Yoder *et al.*, 1993).

Until recently, the limited number and geographic distribution of *in situ* data, and the even smaller number of *in situ* measurements coincident with satellite data (Gordon *et al.,* 1983; Balch *et al.*, 1992) hindered the evaluation of algorithms. These limitations in the amount of available data are critical also because algorithm evaluation and validation should be based on a different and larger data set than that used for algorithm development and calibration.

Now, with the arrival of new Ocean Color sensors (OCTS, POLDER, SeaWiFS, MODIS, MERIS, ...), a new generation of improved algorithms is expected. A systematic evaluation of their accuracy, precision, and utility, and comparisons with algorithms developed for CZCS, requires a suitable data set. This chapter describes the composition, processing, quality control measures, and evolution of such a data set which was used during and following the SeaBAM workshop to evaluate the performance of several semi-analytical and empirical algorithms.

The Preliminary Evaluation Data set (Clear Sky data set)

Prior to the SeaBAM workshop, a small data set was developed by the SeaWiFS project to evaluate independently the performance of several SeaWiFS-Chla or pigment algorithms. Ideally, the evaluation data set should meet several requirements. It should:

1) contain Remote-sensing reflectances (Rrs) and Normalized Water-leaving radiances (Lwn) at or close to the SeaWiFS visible wavelengths (a prerequisite for the semi-analytical algorithms);

2) have the *in situ* chlorophyll *a* concentrations associated with the stations from which Rrs were derived;

3) encompass the widest possible chlorophyll a concentration range;

4) not contain data used for the development of the algorithms under evaluation;

5) be the same for all algorithms under evaluation.

This preliminary evaluation data set was extracted from the SeaBASS (SeaWiFS Bio-optical Archive and Storage System) database (Hooker *et al.*, 1994). Only stations with calm seas and clear and stable sky conditions were selected to ensure a "best case" situation. Additionally, as the original purpose of this evaluation was to identify a general algorithm which performs well in various kinds of waters, some Case II waters data (Morel & Prieur, 1977) were also included.

When not directly available in the database, the remote-sensing reflectances were derived from in-water measurements. Each step of the data processing was carefully conducted on a station per station basis according to the SeaWiFS protocols (Mueller & Austin, 1995). A major step in data processing was a visual inspection of the radiometric profiles to determine the right depth interval for the extrapolations to the surface. Depending on the kind of water under consideration (oligo-, meso-, eutrophic, Case 2), depth intervals for the extrapolations may have differed between Lus and Eds and sometimes between short and long wavelengths. The remote-sensing reflectances were then calculated at each wavelength as:

$$Rrs = Lu(0-)*0.54/Ed(0+) \qquad \text{(Eq. 1)}$$

where Lu(0-) is the upwelling radiance just below the air/water interface, 0.54 is the reflection-refraction correction term to extrapolate Lu(0-) above the interface and Ed(0+) is the incident irradiance above the surface. A reasonable agreement between the extrapolated Ed(0-) value and the deck-cell measurement was checked as a quality control procedure. When available, the Ed(0+) was taken from the deck-cell measurements. Otherwise the Ed(0-) was extrapolated through the interface considering a 4% loss at the interface between Ed(0+) and Ed(0-). The Normalized Water-leaving radiances were derived by multiplying the Rrs by the mean extra-terrestrial incident solar irradiance, Fo, adjusted to the SeaWiFS wavebands.

The resulting initial evaluation data set, also known as the "Clear Sky" data set, contained 98 stations with chlorophyll *a* concentrations varying between 0.05 and 22 ug/l. This data set was used for a preliminary evaluation of three algorithms which was presented at the 7th SeaWiFS Bio-optical Algorithm and Optical Protocols Workshop in Halifax (Nov. 1996). The SeaBAM workshop was initiated right after this workshop, during the Ocean Optics XIII meeting.

Despite the usefulness of the "Clear Sky" data set it was clear that, for the purpose of algorithm evaluation, a much larger data set was highly desirable. Such a larger data set obviously needed to have the same above-mentioned characteristics as the initial evaluation data set. Unfortunately, because of the limited number of recent bio-optical data available and because of requirements 1) and 4) the evaluation data set could not be substantially augmented from completely independent data. In order to circumvent this problem it was decided that a global data set would be created by merging the data sets used by the various participants of the SeaBAM workshop. This choice does not satisfy requirement 4) however the rationale for doing this was that even if some of the algorithms under evaluation would have been tuned with some of the data present in the global data set, these data would only represent a fraction of the global data set.

The SeaBAM Global Evaluation Data Set (SBAM919)

1 - Data processing in SBAM919

The Rrs values used to create the global SeaBAM data set come from various sources and were derived in different ways depending on who processed the data. A summary of these sources, and the methodological differences among subsets comprising the SeaBAM global dataset is provided in **Table** 1. The BBOP data and data from JGOFS (JGOFS-WOCE, EQPAC and NABE) were assembled and processed by S. Garver (Garver & Siegel, this volume) and represent almost one half of the global dataset. The Carder data set is made up of above-water measurements (Carder & Steward, 1985; Lee *et al.*, 1994) collected during various cruises (Carder *et al.*, this volume). The CalCOFI data set contains more than 300 stations which were processed as described in the chapter by Mitchell & Kahru. North Sea and AMT data (Atlantic Meridional Transect) were directly provided to us as remote sensing reflectances by the related investigator, R. Doerffer and G. Moore, respectively. Chesapeake Bay (data from L. Harding) and Canadian Arctic (data from G. Cota) measurements were taken from the SeaBASS archive and were processed as for the "Clear Sky" data set. The MOCE data (from D. Clark and C. Trees) were also extracted from the SeaBASS archive. Because these are hyperspectral measurements at discrete depths, extrapolations to the surface were performed by calculating the diffuse attenuation coefficient between depths of measurement and then using these attenuation values to extrapolate Lu(z) to the surface.

Table 1. Data sets in the global SeaBAM data set

Data set	Provider/PI	Measurements	Data processing
BBOP	D. Siegel	in water	Rrs = Lu(z)/Ed(z) (z < 5 m)
WOCE	J. Marra	in water	Rrs = Lu(z)/Ed(z) (z < 5 m)
EQPAC	C. Davis	in water	Rrs = Lu(z)/Ed(z) (z < 5 m)
NABE	C. Davis - C. Trees	in water	Rrs = Lu(z)/Ed(z) (z < 5 m)
CARDER	K. Carder	above water	Lsky corrected Lu/Ed(0+)
CALCOFI	G. Mitchell	in water	Rrs = Lw/Ed(0-) (extrap.)
MOCE	D. Clark	in water	Rrs = Lw/Ed(0+) (extrap.)
NORTH SEA	R. Doerffer	above water	Lsky corrected Lu/Ed(0+) CaseII
CHESAP. BAY	L. Harding	in water	Rrs = Lw/Ed(0+) (extrap.) CaseII
CANADIAN ARCTIC	G. Cota	in water	Rrs = Lw/Ed(0+) (extrap.)
AMT	G. Moore	in water	Rrs = Lw/Ed(0+) (extrap.)

The major differences in the processing of data contained in the SeaBAM evaluation data set involve the way radiometric data were measured (above-water measurements for North Sea and Carder data, under-water measurements for all other data sets) and the way Rrs were calculated (with extrapolations to the surface for all under-water data sets except BBOP, WOCE, EQPAC and NABE). The various data processing methodologies resulted in Rrs which were either "under-water" Rrs, i.e. Lu(z,~0)/Ed(z, ~0), or "above-water", Lw/Ed(0+). For consistency, all data were converted to "above-water" Rrs.

Some bio-optical algorithms require normalized water-leaving radiance measurements (Lwn) instead of Rrs. For these algorithms the Remote sensing reflectances were converted to Lwn by multiplying the Rrs by the mean extraterrestrial Solar Irradiance corrected for the spectral sensitivity of the relevant sensor (SeaWIFS, POLDER, OCTS).

2 - Verification and selection of the data

For round #1 of the algorithm evaluation, the global data set was used "as is" that is without major quality control procedures performed on the data. At that time, the global data set had 941 stations. During the workshop, several problems arose which were associated with some of the data. These problems were quickly identified as either errors in data processing, wrong unit conversions or calibration problems. Detecting bad or wrong radiometric data is not always straightforward. For instance, the values of band ratios may appear "reasonable" yet the ratio components may contain errors which cancel in the ratios. In order to detect this type of problem and as a quality control procedure of the data, band ratios vs C and single bands vs C were plotted for the various data sets in order to identify outliers. In **figure 1.edc**., the upper panel shows the Rrs at 412nm as a function of the *in situ* C concentration whereas the lower panel is a plot of the ratio of Rrs(412) to Rrs(555) for the same stations. In the upper panel, at least three subsets show questionable values at 412 nm. The first one (white diamonds) corresponds to one of the three legs contained in the WOCE data set. These data show very high Rrs compared with all other data sets at equivalent Chl *a* concentrations. These high values were caused by incorrect calibration of this channel which was corrected on subsequent cruises. The very high values corresponding to the white circles in **figure 1.edc** resulted from an error in the conversion of the original data. As the lower panel in **figure 1.edc** shows, this kind of error cannot be uncovered from bands ratio plots. A third set of data (white triangles) also appeared anomalous on the single wavelength plot (**figure 1.edc**, upper panel) whereas band ratios for this set do not depart dramatically from the general trend (figure 1.edc, lower panel). In this case, an error in data processing was responsible for these high values. On the upper panel of **figure 1.edc**, stations marked with a star are above-water measurements in Case 2 waters. These plots were very useful in revealing data with errors, and useful in our determinations of which data could be corrected and which should be removed from the global data set.

Other methods were also used to identify additional probable errors in the radiance data. Eight bio-optical chlorophyll/pigment algorithms (CalCOFI 2-band Linear; CalCOFI 4-band; Global processing; Morel1; Morel3; OCTS_C; Carder; Garver/Siegel, (see Table **1** in the algorithm evaluation chapter) were used to estimate chlorophyll or pigment from radiance data. These models were chosen because their outputs showed a good linearity in log-log space (minimal curvilinearity) and/or they use three or more wavelengths. A station was considered an outlier when the ratio of the modeled Chlorophyll a concentration to the *in situ* concentration exceeded 5:1 or was less than 1:5, for two or more models. This screening method allowed the identification of most of the stations with one or more "anomalous" radiance value. We deliberately choose rejection criteria which were not too restrictive (1:5 and 5:1) in order to eliminate only the extreme stations. The results of this part of the data evaluation process are presented in **Table 2 and** Appendix I.

After the elimination of suspect stations, 876 stations remained in the global data set. Following the workshop, 43 new stations from CalCOFI were later tested and added, and the current version of the global data set has 919 stations (Table 3). The SeaBAM participants may have used either of these two versions (n=876 or n=919).

Table 2. Frequency of outliers for each data set and algorithm (See text for criteria used to detect outliers).

DATA SET

Algorithm	BBO1	BBO2	WOCE	EQPA	NABE	CARD	COFI	UMD	MOCE	DOER	AMT	COTA	Total
Calco_2b													0
Calco_4b												3	3
GPs	1	5										2	8
Morel1		1				1						2	4
Morel2	1	3				1		5				2	12
OCTS_c						5							5
Carder		1						4				4	9
Garver/Siegel			3			5	1	2				2	13
Total	2	10	3	0	0	12	1	11	0	0	0	15	54

3 - Description of the final Global Evaluation Data set

Beside the differences in methods used to process radiance data described above, there are other important differences among subsets included in the global data set. Most notable differences concern the methods used to measure chlorophyll *a* (and phaeophytin), and the actual wavelengths available in each individual data set.

a) Chlorophyll *a*

The global evaluation data set is comprised of fluorometric and/or HPLC measurements of chlorophyll *a* (hereafter C). **Table** 3 indicates the type of pigment data available in each subset. The consequences of using either fluorometric or HPLC chlorophyll cannot be fully assessed as only three subsets contain both kinds of measurements (BBOP, MOCE, AMT). Because of some uncertainties in the HPLC data in AMT-1 (according to the data provider) only BBOP and MOCE data could be used for the comparison of both methods. Once the duplicates are removed from the BBOP data set, 65 stations remain. A linear regression between the two sets of measurements indicates a very good agreement, with fluorometric C being slightly lower than HPLC C (HPLC = 0.96*Fluor + 0.005, $R^2 = 0.96$). These results are consistent with earlier observations (Trees *et al.*, 1985; Bricaud *et al.*, 1995; see also Mitchell & Kahru this volume). Furthermore, when all chlorophyll data (HPLC and fluorometric) are plotted against radiance band ratios, no particular trend differentiates the two types of chlorophyll measurements.

Therefore, the following strategy was used to generate a unified set (n=919) of chlorophyll measurements which may be used to evaluate the performance of chlorophyll algorithms. When available (except for AMT), HPLC C was selected because HPLC is considered a more precise method. For the subsets which do not contain HPLC data, fluorometric C was selected.

The statistical characteristics of the global SeaBAM pigment data and its subsets are illustrated in figure 2.edc (see also Garver & Siegel, this volume). The *in situ* C concentration available in the global data set ranges between 0.019 and 32.79 ug/l with a mean value of 0.915 ug/l. The BBOP, WOCE and EQPAC data sets represent most of the data at the low end of the concentration range. The Carder, CalCOFI and AMT data cover a relatively wide range of concentration from oligo- to eutrophic waters. NABE, North Sea, Chesapeake Bay and Canadian Arctic data are mostly at concentrations over 1ug/l.

Table 3. Characteristics of the data in the final version of the SeaBAM global data set.

Data set	n	f_chla	f_phaeo	h_chla	h_phaeo	Wavelengths					
BBOP92-93	72	72	72	72		410	441	488	520	565	665
BBOP94-95	67	61	61	67		410	441	488	510	555	665
WOCE	70	70				410	441	488	520	565	665
EQPAC	126			126		410	441	488	520	550	683
NABE (CHORS)	72			72		412	441	488	521	550	-
NABE (NRL)	40			40		410	441	488	520	550	683
CARDER	87	87				412	443	490	510	555	670
CALCOFI	303	303	303			412	443	490	510	555	665
MOCE1	8	8	0	8		412	443	490	510	555	-
MOCE2	5	5	5	5		412	443	490	510	555	-
North Sea	10			10		412	443	490	510	555	670
Ches. Bay	9			9	9	412	443	490	510	555	671
Can. Arctic 96	8	8	7			412	443	490	509	555	665
AMT	42	42		33		412	443	490	510	555	-
TOTAL	919	656	448	442	9						

If 0.1 and 1 ug/l are taken as limits between oligo- and mesotrophic waters and between meso- and eutrophic waters (see e.g. Antoine *et al.*, 1996), the global data set is approximately comprised of 23, 59 and 18 % of Oligo-, Meso- and Eutrophic stations, respectively. Compared with the World Ocean distribution, the Global data set tends to be over-represented by mesotrophic and eutrophic data. Alternatively, for algorithm evaluation purposes, where a more uniform distribution over the concentration range may be desirable, the global data set has a relative insufficiency of concentrations > 7-8 ug/l. Almost all Case II water data (i.e. North Sea and Ches. Bay) are at concentrations > ~3 ug/l.

b) Pigment

Most recent algorithms estimate C, but several empirical algorithms were designed to estimate pigment, [C+P]. The term “Pigment” is actually related to the product that was derived from the CZCS data and it represents the sum of Chl *a* and phaeophytin concentrations as measured by the fluorometric method (Yentsch & Menzel, 1963). The aim and rationale of pigment algorithms are discussed in the evaluation chapter (O’Reilly & Maritorena, this volume).

In the SeaBAM dataset only 448 of 919 records (49%) had fluorometric measurements of both chlorophyll *a* and phaeophytin *a* (Table 3). This included observations from 4 of the 11 subsets in the global dataset [BBOP(n=133), COFI(n=303), MOCE(n=5), and Canadian Arctic (n=7)]. Consequently, the design of the SeaBAM global dataset is less suited for the evaluation or the construction of pigment algorithms. Instead of only using these stations we maintain that it is preferable that all algorithms be evaluated by data sets with similar characteristics and with the same number of stations. This would allow fairer statistical comparisons. We therefore investigated a statistical relationship between Chl *a* and pigment. A tight, coherent relationship would mean that reasonable estimates of pigment could be made from chlorophyll for those stations where pigment (phaeophytin *a*) was missing. A set of 2329 stations (1978-1996) with both Chl *a* and phaeo fluorometric data, all in the 0-10m depth range, was extracted from the "Historical Pigment Database" available in the SeaBASS archive (Firestone & McClain, 1994). This set includes 336 stations with fluorometric data available in the SeaBAM data set. The data come from various investigators and are mostly located in the US coasts (East and West coasts, Bermuda, Sargasso Sea, Gulf of Mexico). A few other data come from the Atlantic and Pacific Oceans. Stations with [phaeo] > [Chl *a*] (n = 67) were considered extreme cases and removed. A type II (RMA) regression on log-transformed data was then performed between [Chl *a*] and pigment ([Chl *a* + phaeo]) on the 2262 remaining stations (**figure 3.edc**) and yields the following equation :

$$[C+P] = 1.34*C^{0.983} \quad (n = 2262, R^2 = 0.993) \qquad \text{(Eq. 2)}$$

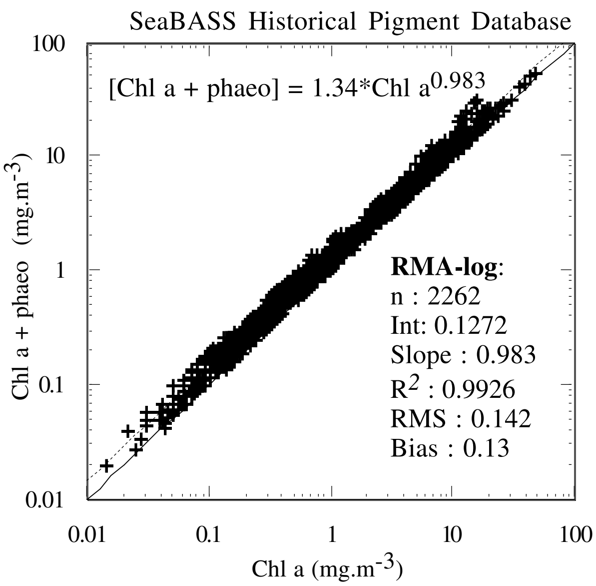

Figure 3. Plot and statistics of the regression (Type II - RMA) between pigment and Chl *a* for the data set extracted from the SeaBASS "Historical Pigment Database" (n = 2262).

It is acknowledged that C and C+P cannot be considered as completely independent variables in the above regression. The major benefit of this approach is to allow the derivation of a chlorophyll/pigment relationship from a more stable basis than one based on chlorophyll vs phaeopigment. We should note here that Balch *et al.* (1992) reported large variability (R^2=0.68) in the relationship between phaeopigment and chlorophyll for a large dataset (mostly from northern hemisphere; n=23,000). Based on their results, the C/[C+P] ratio, ρ, varies from 0.42 to 0.79 when C increases from 0.01 to 10ug/l. For the same concentration range, our chlorophyll to pigment relationship generates C/[C+P] ratios from 0.69 to 0.78, which is close to the results presented by Morel & Berthon (1989). Smith and Baker (1978), using a smaller data set (n = 136), reported an equation which yields ρ between 0.59 and 0.90 for C at 0.01 and 10.0ug/l, respectively. Part of the observed variability in these chlorophyll to pigment relationships results from methodological differences among investigators (Welschmeyer 1994), but the major source of variability probably comes from natural sources associated with which waters are considered. This makes the derivation of a "global" Chlorophyll to pigment relationship very difficult. However, we consider that our equation yields reasonable C/[C+P] ratios and that it can thus be used to estimate the pigment concentration for the stations where the complete fluorometric information is missing (n= 471).

Therefore, the following strategy was used to build a complete (n=919), unified "Pigment" field in the SeaBAM global dataset (Column 9) : The pigment field contains the sum of fluorometric Chl *a* + pheo concentrations when both were available. Otherwise, pigment concentration was estimated from Chl *a* according to equation 2. The calculation used fluorometric Chl *a* when available otherwise HPLC Chl *a* was used.

c) Radiometric data

As can be seen in **table 3**, some subsets contain data which do not exactly match the SeaWiFS wavebands. The biggest differences from the required SeaWiFS wavelengths occurred for the 670 nm channel. However, among all algorithms tested, the only one which uses the 670 nm channel is the Garver/Siegel model (see the chapter by Garver & Siegel). This model is not very sensitive to the 670 nm value (as far as it is "reasonable") and the parameters of the model can be adapted to the available wavelengths in the data sets (See the algorithm evaluation chapter). The 2 nm differences in data sets with 410, 441 and 488 nm was considered negligible. The other major departures from the SeaWiFS wavebands involve the data sets which have 520 nm instead of 510 nm (BBOP 92-93, WOCE, EQPAC and NABE) and those which contain 565 nm instead of 555 nm (BBOP 92-93 and WOCE).

The BBOP94-95 data have radiometric measurements at both 555 and 565 nm. A linear regression between the Rrs at these two wavelengths yields the following equations (n = 78) :

$$Rrs(555) = 1.0628*Rrs(565) + 0.0002 \qquad R^2 = 0.975 \qquad \text{(Eq. 3a)}$$

Because Rrs(555) and Rrs(565) nm are very highly correlated in the BBOP94-95 data set and as the chlorophyll *a* concentration range of the BBOP data is relatively narrow, the BBOP92-93 Rrs(565) data, which were obtained in the same waters as BBOP94095, were converted into "proxy-Rrs(555)" according to equation 3a. As the WOCE data set comes from Case I waters with Chl *a* concentrations very similar to BBOP, equation 3a was also applied to the WOCE Rrs(565)nm data. These convertions allow a better match of the radiometric data with the SeaWiFS bands.

It must be kept in mind that in the 550-565 nm region, the Rrs spectrum is relatively stable and relatively unresponsive (flat) to changes in C concentration as chlorophyll absorbs only weakly at these wavelengths. Rrs(555) and Rrs(565) have very close values and their ratio never departs much from 1 (see O'Reilly and Maritorena, this volume). In the SeaBAM data set, column 18 (r555_565) is a unified column which contains mixed 555 and 565 nm data depending on the available wavelength in the original subset. Column 19, r555_565b, contains Rrs(555) from the subsets which actually contained measurements at this wavelength whereas it contains "proxy-Rrs(555) as derived according to equation 3a for the stations which have Rrs data at 565 nm.

Contrary to the 555-565 nm region, the 510-520 nm domain is highly influenced by pigment absorption (mostly by carotenoids) and the Rrs(510)/Rrs(520) ratio is much more variable along the C range (from ~1.33 at C = 0.01 to ~0.9 at C = 10 ug/l, see O'Reilly and Maritorena, this volume). The BBOP94-95 data set also contains both 510 and 520 nm data. The linear regression equation relating these two wavelengths is:

$$Rrs(510) = 1.3239*Rrs(520) - 0.0004 \qquad R2 = 0.8784 \qquad \text{(Eq. 3b)}$$

Because of a much higher systematic difference between Rrs(510) vs Rrs(520) (regression coefficient =1.32) and a weaker statistical relationship between the two wavelengths, we decided not to make estimates of Rrs(510) from Rrs(520) nm data. Instead, we assembled a complete (n=919), unified set of data by using Rrs(510) when available, and Rrs(520) when Rrs(510) was not available. (This field is data column 15 (r510_520) in the SeaBAM dataset). The consequences of using data at wavelengths which do not exactly match wavelengths required by algorithms are analyzed and discussed in O'Reilly & Maritorena (this volume). A complete description of the 20 fields in the SeaBAM global data set is provided in appendix 2.

The Rrs(443)/Rrs(555) ratio as a function of in situ C for the SeaBAM global data set and for its subsets is shown in figure 4.edc. These plots illustrate the dynamic range associated with each subset and in the Global SeaBAM dataset. The dispersion of the data is another informative feature of these plots. The CalCOFI data set, for instance, appears very coherent whereas some other data sets are slightly noisier. Despite the fact that the data in the Global data set come from various investigators and that they were processed differently by different people, the variability in the radiometric data is reasonably limited. At high C concentration, the dispersion of radiance ratios in the SeaBAM global data set increases, mostly because of the low ratio values observed for the Case II waters from the Chesapeake Bay. The shape of the scatter plot for the SeaBAM global data set is clearly sigmoid (in log-log space). This trend is also observed, although it is less marked, for the CalCOFI data. At the low end of the concentration range, all stations have Rrs(443)/Rrs(555) ratios less than 10, much less than the theoretical value for clear waters (~16.53) as predicted by the latest a_w values (Pope & Fry, 1997) and b_{bw} from Morel (1974; see also Smith & Baker, 1981). Information about regression between the Rrs ratio and the in situ [Chl a} is also indicated. It must be kept in mind however that these regressions only illustrate the (relative) linear relationship between the 2 quantities in log-log space and that the parameters of the regression (slope, intercept) are really meaningful only for data sets which have significant dynamic range.

The same sigmoid shape also appears, with a less steep slope, in the Rrs(490)/Rrs(555) vs in situ [Chl a] plot (figure 5.edc). As evidenced by the plots for the SeaBAM global dataset and the CalCOFI subset, a better linearity (in log-log space) is achieved with this ratio in the mid-range of C concentrations. In all cases (subsets and global data set), the dispersion of the band ratio data using the 490/555 is relatively less than using the 443/555 ratio. At low concentration, the highest values of the Rrs(490)/Rrs(555) ratio are close to the theoretical limit (~6.2) for clear natural waters (see O'Reilly and Maritorena, this volume).

The distributional shapes and magnitudes of data shown on figures 4.edc and 5.edc are consistent with the "historical" NET (Clark, 1981) data (except for 3 or 4 stations). They are also consistent with the distributional patterns generated by semi-analytical models from Morel (1988) and Gordon et al. (1988), except over 8 ug/l (see O'Reilly & Maritorena, this volume).

Conclusions and Recommendations

Differences in data processing methodologies is probably responsible for part of the observed dispersion in the SeaBAM global data set. If data were reprocessed in a more consistent manner, the variability within some of the subsets might be reduced but it's not certain that the variability in the global dataset would actually decrease. We expect that there would still be important variability in the data which would reflect the real inherent variability of the bio-optical properties among the waters represented by the various subsets.

It must be kept in mind that one of the factors limiting the size (number of stations) of the Global data set is the "all SeaWiFS wavelengths" requirement. This requirement would be partially canceled if only empirical algorithms were taken into account, allowing the global data set to be much bigger. However, the expansion of the SeaBAM global dataset should not be considered as a simple increase in the number of stations available but rather as an increase in the diversity of bio-optical conditions it encompasses. Several major bio-optical provinces occuring in the World Ocean are not or poorly represented in the SeaBAM data set. For example, as pointed out by Mitchell & Kahru (this volume), the current version of the SeaBAM dataset includes very little data from the polar regions (none from Antarctica). High C concentration (Case I or Case II) data are strongly needed as well as more data with low C concentrations. Spatial and temporal variability in the *in situ* data should also be taken into account in the future developments of the SeaBAM data set.

Also, the increasing need for intercomparison and evaluation of data or algorithms related to various sensors (CZCS, OCTS, POLDER, SeaWIFS, MODIS,...) emphasizes the importance of acquiring data sets which include all wavelengths involved. For the above satellite ocean color sensors, this means that in situ radiometric data should be collected at a minimum of ten wavelengths (412, 443, 490, 510, 520, 530, 550, 555, 565 and ~670 nm) or by hyperspectral devices. **We therefore strongly recommend that more emphasis be placed on acquiring *in situ* radiometric data at wavelengths would allow intercomparisons among ocean color sensors and algorithms.**

The Global data set has its own limitations. The greatest limitations involve wavelength differences and the types of pigment data available in the various subsets. We attempted to quality control the data, excluding obvious outliers, without being too restrictive in the selection of the stations to be included in the Global data set. Because the main objective of the SeaBAM workshop was to define the best generalistic algorithm, we considered that such algorithm should be able to manage a great variety of situations and thus should be able to deal with a certain amount of noise in the input data. For more spatial and temporal targeted objectives more severe selection of the data may be needed. To allow users to manage the data in the way they want, besides the information available in this TM volume, the Global data set is provided to the community in its more detailed format (Appendix II).

References

Antoine, D., J.-M. Andre and A. Morel. 1996. Oceanic Primary Production 2. Estimation at global scale from satellite (CZCS) chlorophyll. Glob. Biochem. Cycles, 10(1) : 57-69.

Balch, W., R. Evans, J. Brown, G. Feldman, C. McClain & W. Esaias. 1992. The Remote Sensing of Ocean Primary Productivity : Use of a New Data Compilation to Test Satellite Algorithms. J. Geophys. Res., 97 (C2) : 2279-2293.

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

Carder, K.L. & R.G. Steward. 1985. A remote-sensing reflectance model of a red tide dinoflagellate off West Florida. Limnol. Oceanogr., 30 : 286-298.

Campbell, J.W. 1995. The lognormal distribution as a model for bio-optical variability in the sea. J. Geophys. Res., Vol.100, C7 : 13237-13254.

Clark, D.W. 1981. Phytoplankton pigment algorithms for the Nimbus-7 CZCS. Oceanography from Space, J.F.R. Gower Ed., Plenum Press, 227-237.

Evans, R.H. & H.R. Gordon. 1994. Coastal zone color scanner “system calibration”: A retrospective examination. Journal of Geophysical Research, 99 : 7293-7307.

Gordon, H.R., O.B. Brown., R.H. Evans, J.W. Brown, R.C. Smith, K.S. Baker and D.K. Clark. 1988. A semianalytic radiance model of ocean color. J. Geophysic. Res., 93(D9) : 10909-10924.

Gordon, H.R., D.K. Clark, J.W. Brown, O.B. Brown, R.H. Evans & W.W. Broenkow. 1983. Phytoplankton pigment concentrations in the Middle Atlantic Bight : comparison of ship determinations and CZCS estimates. Appl. Opt., 22(1) : 20-36.

Feldman, G., N. Kuring, C. Ng, W. Esaias, C. McClain, J. Elrod, N. Maynard, D. Endres, R. Evans, J. Brown, S. Walsh, M. Carle and G. Podesta (1989) Ocean Color: Availability of the global data set. *Eos*, 70:634-641.

Firestone J.K. & C.R. McClain, 1994. The Historical Pigment Database. *In* The SeaWiFS Bio-Optical Archive and Storage System (SeaBASS), Part 1. NASA Tech. Memo. 104566, Vol. 20, NASA Goddard Space Flight Center, Greenbelt, Maryland, 40pp.

Hooker, S.B., C.R. McClain, J.K. Firestone, T.L. Westphal, E.-N. Yeh & Y. Ge. 1994. The SeaWiFS Bio-Optical Archive and Storage System (SeaBASS), Part 1. NASA Tech. Memo. 104566, Vol. 20, NASA Goddard Space Flight Center, Greenbelt, Maryland, 40pp.

Lee, Z., K.L. Carder, S.K. Hawes, R.G. Steward, T.G. Peacock & C.O. Davis. 1994. Model for the interpretation of hyperspectral remote-sensing reflectance. Appl. Opt., 33(24) : 5721-5732.

Morel, A. 1974. Optical properties of pure water and pure sea water, in Optical aspects of Oceanography, edited by N.G. Jerlov and E. Steemann Nielsen, pp 1-24, Academic Press, San Diego, Calif.

Morel, A. 1988. Optical modeling of the upper ocean in relation to its biogenous matter content (Case I waters). J. Geophys. Res., 93 (C9) : 10 749-10 768.

Morel, A. & Berthon, J.F. 1989. Surface pigments, algal biomass profiles, and potential production of the euphotic layer: Relationships reinvestigated in view of remote sensing applications. Limnol. Oceanogr., 34(8) : 1545-1562.

Morel, A. & L. Prieur. 1977. Analysis of variations in ocean color. Limnol. Oceanogr., 22(4) : 709-722.

Mueller, J. L. & R.W. Austin. 1995. Ocean Optics Protocols for SeaWiFS Validation, Revision 1. NASA Tech. Memo. 104566, Vol. 25, NASA Goddard Space Flight Center, Greenbelt, Maryland, 67pp.

Pope, R.M. & E.S. Fry. 1997. Absorption spectrum (380-700 nm) of pure water: II. Integrating cavity measurements, submitted to Appl. Optics.

Smith, R.C. & K.S. Baker. 1978. The bio-optical state of ocean waters and remote sensing. Limnol. Oceanogr., 23(2) : 247-259.

Smith, R.C. & K.S. Baker. 1981. Optical properties of the clearest natural waters (200-800 nm). Appl. Opt., 20(2) : 177-184.

Trees, C.C., M.C. Kennicutt and J.M. Brooks. 1985. Errors associated with the fluorometric determination of chlorophylls and phaeopigments, Mar. Chem., 16 : 1-12.

Welschmeyer, N.A. 1994. Fluorometric analysis of chlorophyll a in the presence of chlorophyll b and phaeopigments. Limnol. Oceanogr. 39(8):1985:1992.

Yentsch, C.S. & D.W. Menzel. 1963. A method for the determination of phytoplankton, chlorophyll, and phaeophytin by fluorescence. Deep-Sea Res., 10 : 221-231.

Yoder, A.Y., C.R. McClain, G.C. Feldman and W.E. Esaias. 1993. Annual cycles of phytoplankton chlorophyll concentrations in the global ocean: A satellite view. Global Biochem. Cycles, 7 : 181-193.

List of figures

edc_figure 1. Example of quality control procedure used to identify outliers in the SeaBAM data set. Upper panel: Scatter plot of Rrs(412) vs *in situ* [Chl *a*]. Problems with stations marked with diamonds, triangles and circles were identified as related to calibration issues or errors in data processing. Lower panel: Scatter plot of Rrs(412)/Rrs(555) vs *in situ* [Chl *a*] for the same stations (see text).

edc_figure 2. Frequency distribution, cumulative frequency and statistics associated with the concentration of Chl *a* in the SeaBAM global dataset (upper left panel) and in the 11 subsets.

edc_figure 3. Plot and statistics of the regression (Type II - RMA) between pigment and Chl *a* for the data set extracted from the SeaBASS "Historical Pigment Database" (n = 2262).

edc_figure 4. Scatter plots of Rrs(443)/Rrs(555) vs *in situ* [Chl *a*] for the SeaBAM global dataset (upper left panel) and in the 11 subsets. Statistics of the regression (Type II - RMA in log units) between ratio and *in situ* [Chl *a*] are also indicated.

edc_figure 5. Same as figure 4 but for the Rrs(490)/Rrs(555) ratio.

Appendix 1. Stations identified as outliers by 1 or more algorithms

	Sta./Mod.	Calco_2b	Calco_4b	GPs	Morel1	Morel2	OCTS_c	Carder	Garver & Siegel	TOTAL
1	cota96.9		1	1	1	1		1	1	6
2	bbop9495.27			1	1	1		1		4
3	cota96.1		1		1			1	1	4
4	cota96.11		1	1		1		1		4
5	carder.26				1	1	1			3
6	bbop9293.65			1		1				2
7	bbop9495.24			1		1				2
8	bbop9495.25			1		1				2
9	carder.19						1		1	2
10	umd.1					1			1	2
11	umd.10					1		1		2
12	bbop9495.26			1						1
13	bbop9495.28			1						1
14	calco.227								1	1
15	carder.25						1			1
16	carder.29						1			1
17	carder.30						1			1
18	carder.33								1	1
19	carder.43								1	1
20	carder.51								1	1
21	carder.72								1	1
22	cota96.7							1		1
23	umd.11					1				1
24	umd.2								1	1
25	umd.3					1				1
26	umd.5							1		1
27	umd.6							1		1
28	umd.7							1		1
29	umd.9					1				1
30	woce.16								1	1
31	woce.2								1	1
32	woce.7								1	1
	TOTAL	0	3	8	4	12	5	9	13	54

Appendix 2. Description of the various fields (= columns) in the SeaBAM global data set.

The SeaBAM global data set is made available to the community in two formats, a Microsoft Excel table (SBAM919.xls) and an ascii text file (tab. separated, SBAM919.tab). The Excel table contains headers in the first line, the ascii file does not have headers.

There are no empty fields. When a data is missing, it is replaced by -1.

Column 1 : Station ID.

An ID has been given for each station of the various subsets. This ID is comprised of two parts which are separated by a dot (.) character. The left side of the ID refers to the subset the station comes from and the right side is the number of that station in that particular subset. The numbers were given prior to the quality control procedures so the station IDs in the data sets which contained outliers don't follow a continuous sequence.

Column 2 : Subset ID

For each subset, a 4 characters ID has been defined. These ID are as follows : BBO1 (BBOP 1992-1993), BBO2 (BBOP 1994-1995), WOCE, EQPA, CARD (Carder), COFI (CalCOFI), MOCE (MOCE1 & MOCE2), DOER (R. Doerffer, North Sea data), UMD_ (University of Maryland, Chesapeake Bay data), COTA (G. Cota, Canadian Arctic data, 1996), AMT_.

Column 3 : Original ID

This field contains the original ID of the station as defined by the data provider.

Column 4 : Chlorophyll *a* concentration

This column contains a "unified" (i.e. n = 919) field of [Chl a] for each station. This unified field contains the HPLC Chl *a* concentration when HPLC measurement is available otherwise it contains fluorometric Chlorophyll *a* .

Column 5 : Fluorometric Chlorophyll *a* concentration.

If available in the original data set, the fluorometric Chlorophyll *a* concentration is given in column 5 (otherwise -1 is indicated).

Column 6 : Fluorometric phaeophytin concentration.

If available in the original data set, the fluorometric phaeophytin concentration is given in column 6 (otherwise -1 is indicated).

Column 7 : HPLC Chlorophyll *a* concentration.

If available in the original data set, the HPLC Chlorophyll *a* concentration is given in column 7 (otherwise -1 is indicated).

Column 8 : HPLC phaephytin concentration.

If available in the original data set, the HPLC phaeophytin concentration is given in column 8 (otherwise -1 is indicated).

Column 9 : Pigment

Column 9 is a unified field which contains the pigment concentration defined as the sum of Chl a and phaephytin. This field contains the sum of fluorometric Chl *a* + pheo concentrations when both were available in the original data sets. Otherwise, pigment concentration was estimated from Chl *a* according to equation 2 ($[C+P] = 1.34*C^{0.983}$). The calculation used fluorometric Chl *a* when available otherwise HPLC Chl *a* was used.

Column 10 : Rrs(412)

This field contains the Remote sensing reflectance (= Lw/Ed(0+)) at 412 nm.

Column 11 : Rrs(443)

This field contains the Remote sensing reflectance (= Lw/Ed(0+)) at 443 nm.

Column 12 : Rrs(490)

This field contains the Remote sensing reflectance (= Lw/Ed(0+)) at 490 nm.

Column 13 : Rrs(510)

This field contains the Remote sensing reflectance (= Lw/Ed(0+)) at 510 nm (if available otherwise -1 is indicated).

Column 14 : Rrs(520)

This field contains the Remote sensing reflectance (= Lw/Ed(0+)) at 520 nm (if available otherwise -1 is indicated).

Column 15 : Mixed Rrs(510) and Rrs(520) (r510_520)

This column is a unified field from column 13 and 14 and contains mixed Rrs(510) and Rrs(520) data depending on the wavelength available in the original data set.

Column 16 : Rrs(555)

This field contains the Remote sensing reflectance (= Lw/Ed(0+)) at 555 nm (if available otherwise -1 is indicated).

Column 17 : Rrs(565)

This field contains the Remote sensing reflectance (= Lw/Ed(0+)) at 565 nm (if available otherwise -1 is indicated).

Column 18 : Mixed Rrs(510) and Rrs(520) (r555_565)

This column is a unified field from column 16 and 17 and contains mixed Rrs(555) and Rrs(565) data depending on the wavelength available in the original data set.

Column 19 : Mixed Rrs(555) and proxy-Rrs(555) (r555_565b)

This column is a unified field from column 16 and data from column 17 converted to "proxy-Rrs(555)" according to equation 3a (Rrs(555) = 1.0628*Rrs(565) + 0.0002; see text). The converted data are from BBOP and WOCE.

Column 20: Rrs(520)

This field contains the Remote sensing reflectance (= Lw/Ed(0+)) at 670 nm (if available otherwise -1 is indicated).