Metrics for Quantifying the Uncertainty in a Chlorophyll Algorithm: Explicit equations and examples using the OC4.v4 algorithm and NOMAD data

Janet W. Campbell & John E. O'Reilly
January 3, 2006

Summary

This document describes statistical procedures for quantifying the uncertainty associated with a single retrieval of a chlorophyll algorithm using a data base of coincident optical properties and chlorophyll-a concentration measured *in situ*. The procedures are actually more general than just chlorophyll and would apply to any property that varies over several orders of magnitude. Explicit equations are given, and the method is demonstrated by quantifying the uncertainty of the SeaWiFS chlorophyll algorithm, OC4.v4, using the newly published NASA bio-Optical Marine Algorithm Data set (NOMAD, Werdell *et al*. 2005).

Chlorophyll ranges over four orders of magnitude globally, from 0.01 to 100 mg m^{-3}, and is approximately lognormally distributed within discrete water masses (Campbell, 1995). The chlorophyll algorithm should perform well over the range of concentrations found globally. Thus, the *in situ* chlorophyll measurements are log-transformed, and the algorithm retrieval error is defined as the difference between the logarithm of the algorithm-derived chlorophyll, $\hat{C}$, and the log-transformed *in situ* chlorophyll, C. Uncertainty is then characterized by the mean error (bias) and root-mean-square error (RMSE). While these statistics are useful in comparing different algorithms, they are difficult to interpret because the units are decades of log. It is desirable to express uncertainty in the algorithm as a relative or percentage error. For each retrieval, the relative error can be derived from the log error with a simple inverse log transform, but the same relationship does not hold for the error statistics. Statistics on relative errors can be derived empirically from the data or, if the log errors are normally distributed, from the log error statistics. A lognormal assumption is reasonable for the chlorophyll algorithm, and thus the mean, median and standard deviation of the relative or percentage error can be derived from the statistics of the log errors. Equations are provided for converting log-based error statistics to percentage error statistics.

Issues related to the distribution of the *in situ* data vis-à-vis the global chlorophyll distribution are addressed. *In situ* databases tend to over-represent high-chlorophyll waters because more data are collected in coastal regions than in the open ocean. We derived weighted error statistics that compensate for differences between the data distribution and the distribution of chlorophyll found globally. In addition to prescribing a method for quantifying uncertainty, we also describe methods found in the literature which are incorrect or misleading.

All procedures described in this paper are illustrated with the SeaWiFS chlorophyll algorithm, OC4.v4 (O'Reilly *et al*. 2000) and evaluated using a subset of the newly published NOMAD data set (Werdell *et al*. 2005). The subset contains data from 2208 stations for which there were measurements of the remote-sensing reflectance, $R_{rs}(\lambda)$, at wavelengths λ = 411, 443, 489, 510, and 555, and a coincident measurement of the upper-layer chlorophyll concentration. These wavelengths correspond to the first 5 bands of SeaWiFS.

The chlorophyll algorithm

The SeaWiFS project currently uses the OC4.v4 chlorophyll algorithm (O'Reilly *et al*. 2000):

$$\log[\mathrm{Chl}] = a_0 + a_1 X + a_2 X^2 + a_3 X^3 + a_4 X^4 \qquad (1)$$

where

$$X = \log\left[\frac{\max(R_{rs}(443), R_{rs}(489), R_{rs}(510))}{R_{rs}(555)}\right] \qquad (2)$$

and the coefficients a_0, a_1, a_2, a_3, a_4 are 0.366, -3.067, 1.930, 0.649, and -1.532, respectively. A new set of coefficients was obtained by fitting the polynomial to the NOMAD data. Herein the algorithm using the fitted polynomial is called OC4.fit.

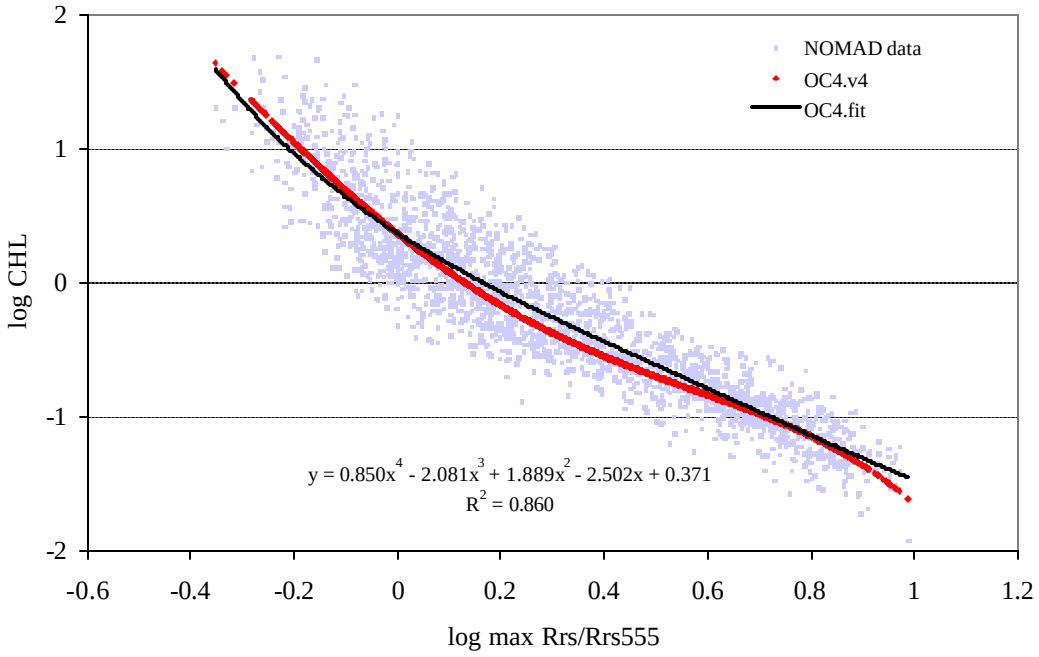

Fig. 1. NOMAD chlorophyll vs. max R_{rs} ratio on log-log scale. Also shown are the OC4.v4 algorithm (red line) and a 4th-order polynomial fitted to the data (OC4.fit).

Errors associated with log-log regressions

In fitting a polynomial to log-log data, the resulting curve minimizes the mean squared error (MSE) between the logarithm of the predicted chlorophyll and the logarithm of the measured chlorophyll. The R^2 associated with the fitted polynomial is a measure of goodness of fit between the log-transformed data and the curve. In regressions of predicted vs. measured chlorophyll, however, the R^2 is not a measure of the algorithm performance. This and other misleading methods are discussed later.

Definition of errors

The error associated with the i^{th} data point is defined as:

$$\delta_i = \log \hat{C}_i - \log C_i = \log\left(\frac{\hat{C}_i}{C_i}\right) \tag{3}$$

The δ_i distributions associated with the two OC4 algorithms are approximately normally distributed (Fig. 2), and thus the ratio $\hat{C}_i / C_i$ is approximately lognormally distributed. In figure 2, the horizontal axis is linear in δ_i but labeled in terms of the ratio $\hat{C}_i / C_i$ on a log scale.

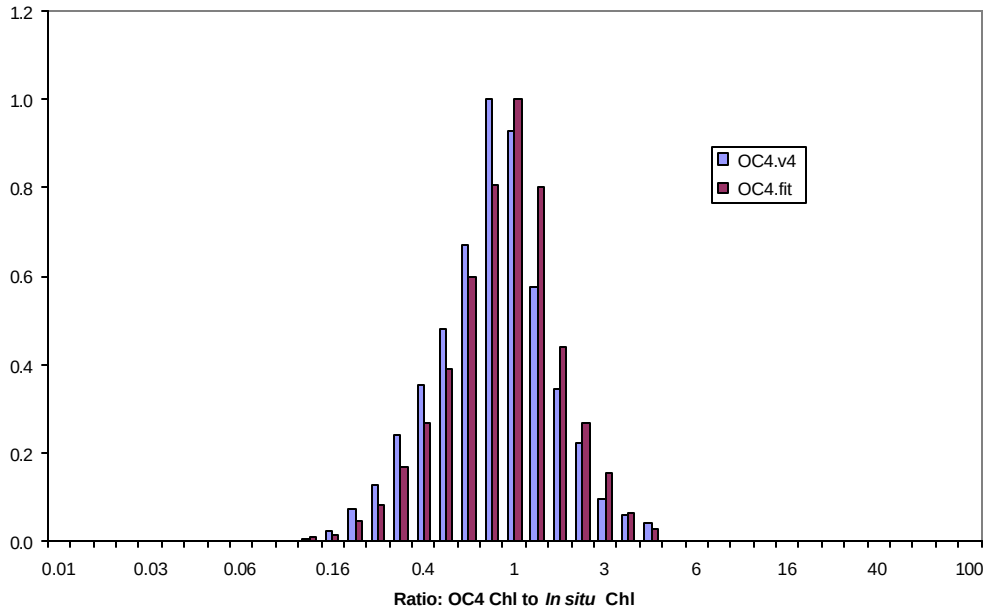

Fig. 2 - Histograms of the δ_i errors as defined in equation (3) for the two algorithms.

Two metrics for comparing algorithms derived from the δ_i sample are the mean error:

$$\text{bias} = \frac{1}{N}\sum_{i=1}^{N}(\log \hat{C}_i - \log C_i) \tag{4}$$

and the square root of the mean squared error (RMSE) where the mean square error is:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\log \hat{C}_i - \log C_i)^2 \qquad (5)$$

The MSE can also be derived from the sample mean (m = bias) and variance (s^2) as:

$$\mathrm{MSE} = m^2 + \frac{N-1}{N}s^2 \qquad (6)^1$$

The RMSE can serve as a single metric because it combines both the mean and variance of the error distribution. However, the bias should be considered as well because it represents an error that cannot be reduced by spatial averaging.

Table 1 lists the error statistics associated with the error histograms shown in figure 2. Note that by fitting the algorithm to this data set the average error (bias) was eliminated and RMSE was minimized.

Table 1. Error statistics for the OC4 chlorophyll algorithms.				
algorithm	N	bias	RMSE	R^2
OC4.v4	2208	-0.047	0.256	0.85
OC4.fit	2208	0.000	0.245	0.86

Table 1 gives estimates of the bias and RMSE for the δ_i sample, but what do these tell us about the relative error? What about the mean and "one standard deviation" range of relative errors?

Relationship between log errors and relative errors

The relative error is defined as:

$$\mathrm{relerr}_i = \frac{\hat{C}_i - C_i}{C_i} \qquad (7)$$

and the percentage error as $100\% \cdot \mathrm{relerr}_i$. For each data point, there is a unique relationship between δ_i and relerr_i:

[1] This equation is provided for those who use the formuli for sample mean and variance in spreadsheets or statistical software. The variance must be multiplied by (N-1)/N to get exactly the same result as that based on equation (5).

$$\delta_i = \log(1 + \text{relerr}_i) \tag{8}$$

Statistics of the relative errors can be calculated empirically from the data, or they can be estimated from the statistics in Table 1 given the reasonable assumption that the δ_i are normally distributed (Fig. 2).

Using the lognormal assumption

To estimate relative error statistics from the log error statistics, we assume that the log error δ is normally distributed with mean *m* and standard deviation *s*:

$$s = \sqrt{\frac{N(RMSE^2 - m^2)}{N-1}} \tag{9}$$

where *m* = bias. Under this assumption, the ratio $\hat{C}/C$ is lognormally distributed. Its mean, median, and standard deviation are:

$$\text{mean} = \exp\left(M + \frac{S^2}{2}\right) \tag{10}$$

$$\text{median} = \exp(M) \tag{11}$$

$$\text{std dev} = \text{mean}\ \sqrt{\exp(S^2) - 1} \tag{12}$$

where M = *m* ln(10) and S = *s* ln(10) are the mean and standard deviation of $\ln(\hat{C}/C)$. The "one sigma" range for the $\hat{C}/C$ ratio about its median value is:

$$\text{lower} = \exp(M - S) \tag{13}$$

$$\text{upper} = \exp(M + S) \tag{14}$$

Under the lognormal assumption, 68% of all ratios will lie within this range.

Statistics for percentage errors calculated empirically and based on the lognormal assumption (eqs. 9-14) are listed in Table 2. The mean and median percentage errors are derived by subtracting 1 from the mean (eq. 10) or median (eq. 11) and then multiplying by 100%. The standard deviation is the standard deviation of $\hat{C}/C$ (eq. 12) multiplied by 100%. The lower rows of Table 2 provide the median (50%) and "one standard deviation" range of the $\hat{C}/C$ ratio corresponding to the lower (16%) and upper (84%) percentiles associated with ±1 standard deviation. The values given under the empirical column were derived from the actual percentiles in the sample rather than using the lognormal assumption.

Table 2. Statistics of the percentage errors and $\hat{C}/C$ ratios associated with the chlorophyll algorithms based on the NOMAD subset.

	Empirical		Lognormal	
relerr (%)	OC4.v4	OC4.fit	OC4.v4	OC4.fit
mean	6%	17%	6%	17%
median	-8%	4%	-10%	0%
std dev	66%	68%	67%	72%
$\hat{C}/C$	based on percentiles		from eqns. (11) - (14)	
lower	0.50	0.58	0.50	0.57
median	0.92	1.04	0.90	1.00
upper	1.54	1.70	1.60	1.76

The first thing to notice is that the errors are no longer unbiased; the mean relative error is positive. If the log-error bias is zero, as is the case with the OC4.fit, then the median relative error is also zero under the lognormal assumption. The second point is that the standard deviations are quite large; it doesn't make sense to think of relative errors as having a range of ±3 standard deviations as would be the case if they were normally distributed.

Expressing uncertainty in terms of ±1 standard deviation has meaning if the distribution is symmetric about the mean, but in the case of large relative errors, the distribution is skewed. Negative errors cannot be larger than -100% whereas positive errors can be arbitrarily large. Use of the log error δ helps to alleviate this problem, but then the units are decades of log which are not easily interpreted. Error histograms (Fig. 2) are a good way to express errors, where the horizontal axis is on a log scale. The axis can be labeled to express the log error (δ) as percentage errors or ratios. The symmetry of the log error (δ) distribution about its mean makes it clear that +100% is equivalent to -50%.

Predicted vs. measured chlorophyll

A common way to evaluate an algorithm is to plot the algorithm-predicted value against the measured value (Fig. 3), and to show both the one-to-one line and a regression of the two. The regression should have a slope of 1 and an intercept of 0, but often this is not the case. The R^2 of such a regression is not a measure of the accuracy of the algorithm as is clearly evident in figure 3. However, such plots are useful for showing systematic trends in the errors. The algorithm used in figure 3 was neither of the OC4 algorithms described earlier.

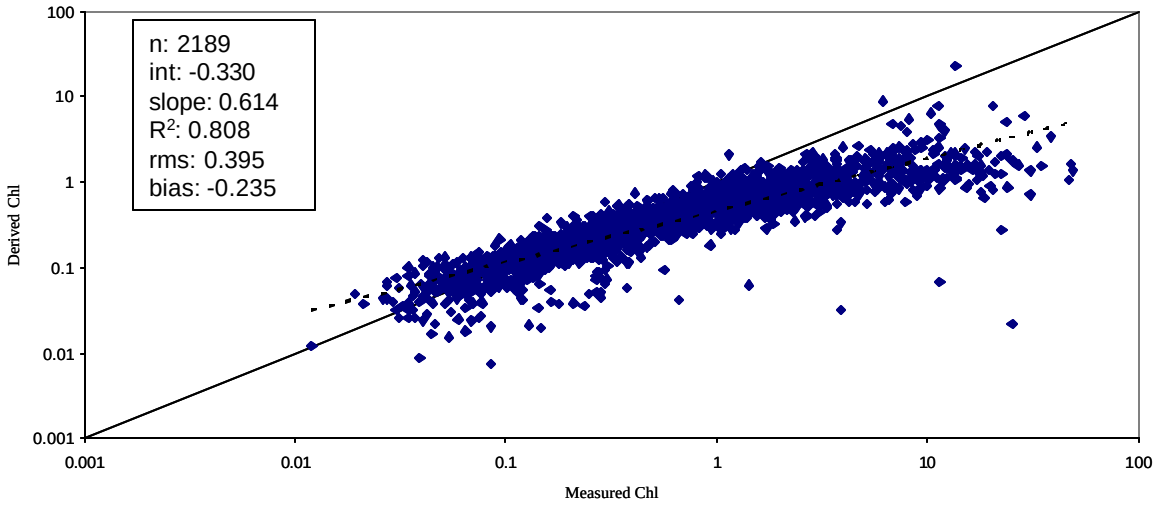

Fig. 3 – Plot of predicted vs. measured chlorophyll (log-log) for an algorithm with systematic errors. The relatively high R^2 associated with the regression (dashed line) is not sufficient evidence of a good algorithm since the slope and intercept are not 1 and 0, respectively.

Metrics used by the SeaBAM workshop

The OC4.v4 algorithm was selected based on metrics that were prescribed by participants at the SeaWiFS Bio-optical Algorithm Mini (SeaBAM) Workshop held in January 1997. The metrics called for 4 diagnostic plots (Fig. 4).

The plots include a log-log plot of the algorithm *vs. in situ* chlorophyll, and a derived regression. The goals for an acceptible algorithm were that the regression meet the following criteria: slope = 1 ± 0.01; intercept = 0. ± 0.01; bias = 0 ± 0.01; RMSE < 0.185; R^2 > 0.9; no negatives; and few outliers ($\hat{C}/C > 5$ or < 0.2) .

Three additional diagnostic plots and criteria were: (1) the histogram of log-based errors (e.g., Figs. 2 and 4c) should be symmetric about zero; (2) histograms of the derived and measured chlorophyll (4d) should be congruent; and (3) quantile-quantile plots (4b) should be linear, overlap the 1:1 line, and have no discontinuities.

To derive the quantile-quantile plots, the derived and measured chlorophyll values are sorted separately, and the sorted arrays plotted against one another. This plot reveals any systematic differences in the frequency distribution of the two data sets (measured vs. predicted).

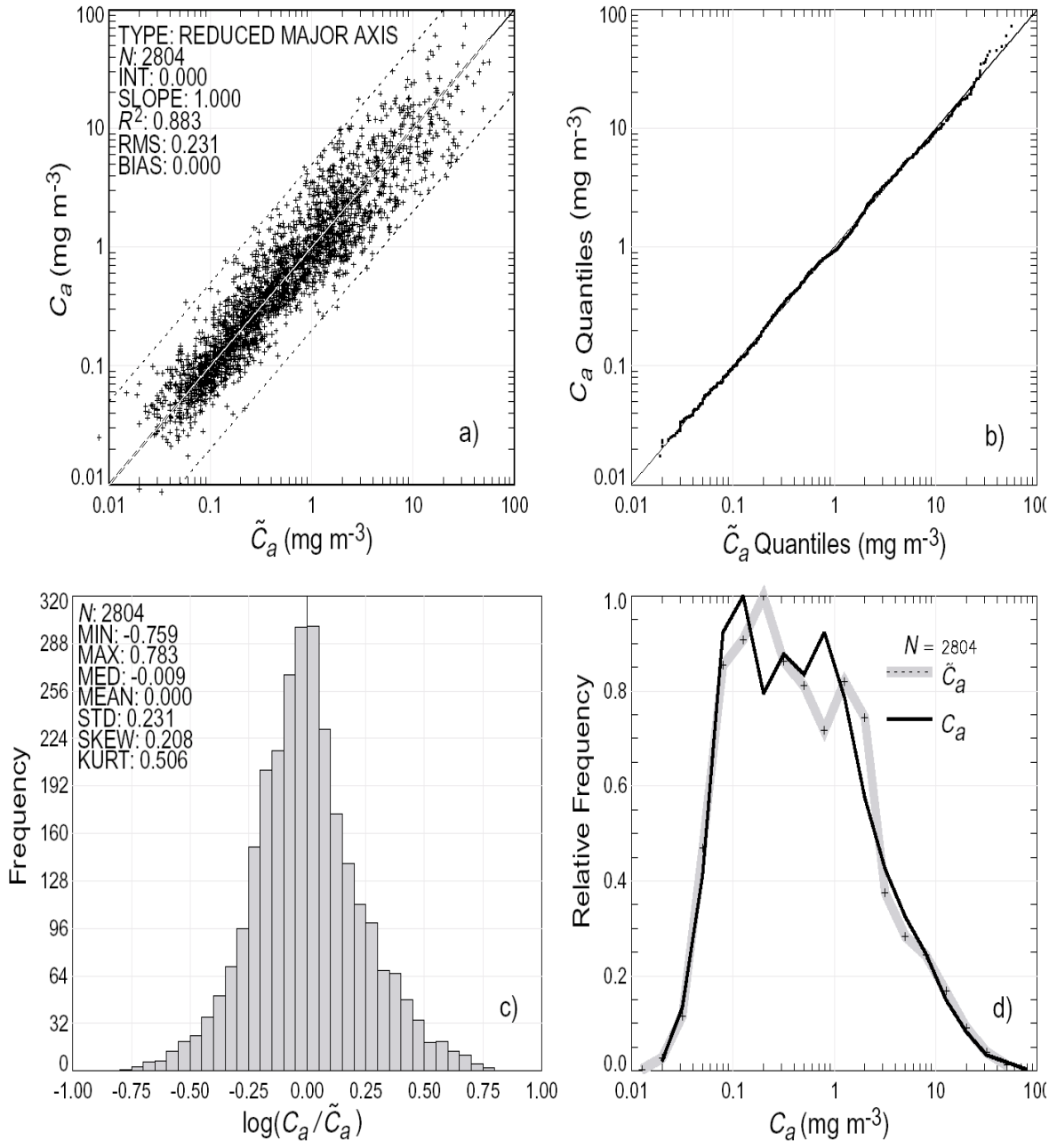

Fig. 4. Diagnostic plots recommended by participants at the SeaBAM workshop, and subsequently used by the SeaWiFS project to select the OC4.v4 algorithm. These plots and related statistics were used to evaluate candidates for the at-launch chlorophyll algorithm for SeaWiFS. Shown here are statistics for the OC4.v4 algorithm compared with an *in situ* data set derived from the SeaWiFS Bio-optical Archive (SeaBASS). Details about this algorithm and the *in situ* data set may be found in the NASA Technical Memorandum 2000-206892, Volume 11. This figure appears in figure 9 (pg. 18) of that publication and the notation is unchanged. C_a is the predicted chlorophyll and $\tilde{C}_a$ is the *in situ* chlorophyll.

Mistakes found in the literature

A number of erroneous estimates of the relative uncertainty in chlorophyll algorithms have appeared in the literature. One is to define relative uncertainty as $100\% \cdot \delta_i$ (Fig. 5). If errors are small, and the natural logarithm is used instead of the base-10 logarithm, then one could approximate relative errors by $100\% \cdot (\ln \hat{C}_i - \ln C_i)$. This stems from the fact that

$$\partial \ln C = \frac{\partial C}{C} \tag{15}$$

This estimate of the relative error is also plotted on figure 5 (red line). It lies closer to the 1:1 line near the origin, but still deviates for errors greater than ± 10%.

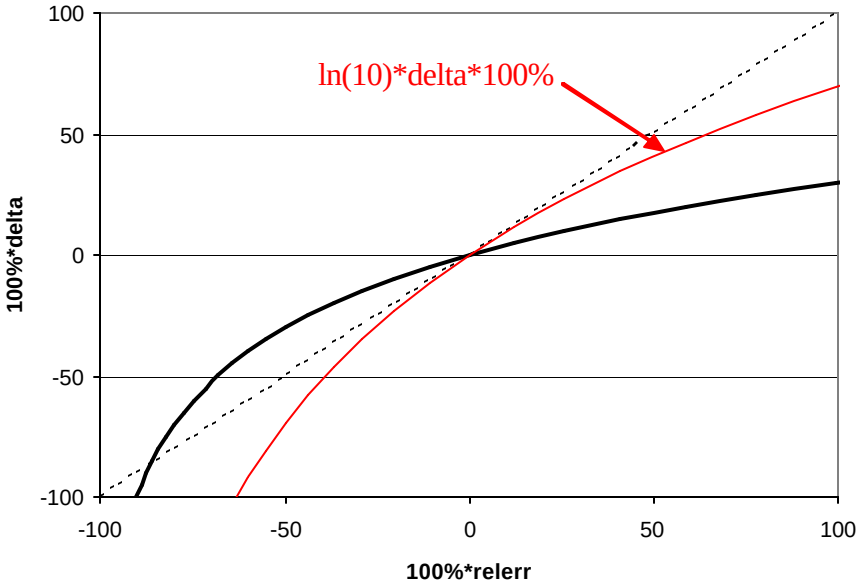

Fig. 5. Plot showing that 100% δi is not the relative error.

Another mistake is to claim that 1- R^2 is the percentage error. One reasons that if R^2 is the percentage of the variance explained by the algorithm, then 1- R^2 is the percentage unexplained, hence the percentage error. If the R^2 used is that associated with the regression of predicted *vs.* measured, then this is definitely wrong since, as already explained, the R^2 in this type of regression is not a measure of the accuracy of the algorithm. However, the R^2 associated with a polynomial fit (e.g., that shown in figure 1) is a measure of the fit, and does relate to the percentage of the variance explained by the fitted curve. However, it is a percentage of the variance explained in log Chl, not Chl. In the case of the OC4.fit algorithm, the total variance in log Chl was 0.4285 and the R^2 was 0.86. Therefore, the unexplained variance was 0.14 x 0.4285 = 0.0600, and the standard error (square root of the error variance) is 0.245. This is the same as the RMSE.

Adjusting for the distribution of chlorophyll

One problem with using a dataset such as NOMAD is that the distribution of the data is not representative of the distribution of chlorophyll globally (Fig. 6).

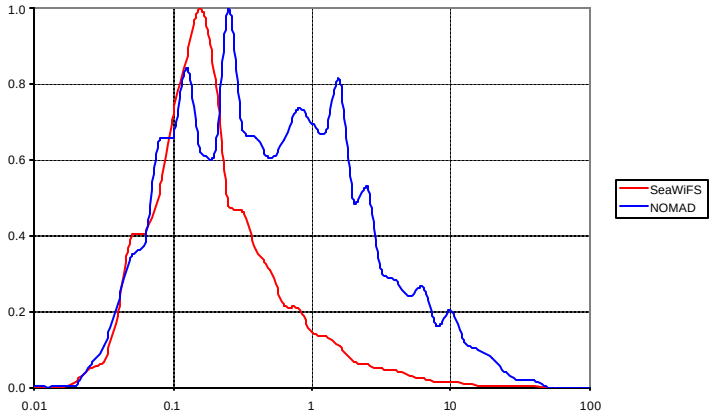

Fig. 6 – Histograms of chlorophyll data. NOMAD data distribution (blue) is compared with global distribution in the SeaWiFS climatology (1997-2005) (red).

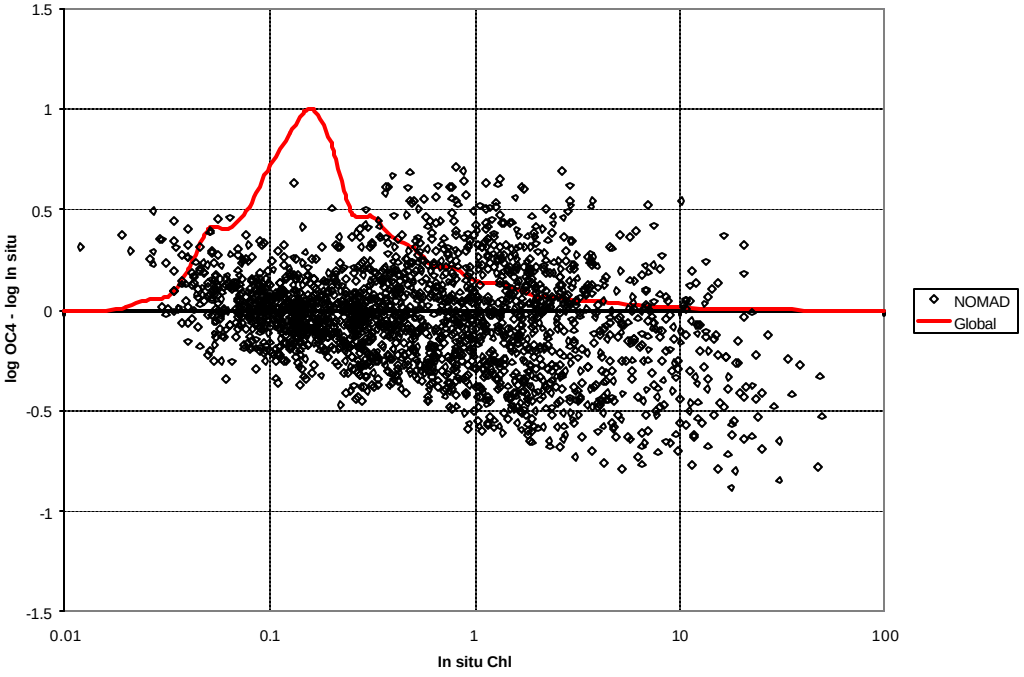

Fig. 7. Errors in the OC4 algorithm plotted against *in situ* Chl. The smallest errors are near the mode of the global Chl distribution shown in red.

Ideally, an algorithm should have the same uncertainty everywhere, that is, in high as well as low chlorophyll waters. To assess this, one should plot algorithm errors against chlorophyll (Fig. 7). It is immediately clear that the algorithm errors increase with increasing chlorophyll, and that the smallest errors are in the most probable areas (i.e., near the mode). By using the NOMAD data to assess the performance of an algorithm, results are biased by its relative poor performance in high chlorophyll waters.

To explore how the distribution of the NOMAD chlorophyll data affects our assessment of the RMSE, we computed the cumulative RMSE for all stations with chlorophyll below each measured Chl. These are plotted in figure 8 for the two algorithms.

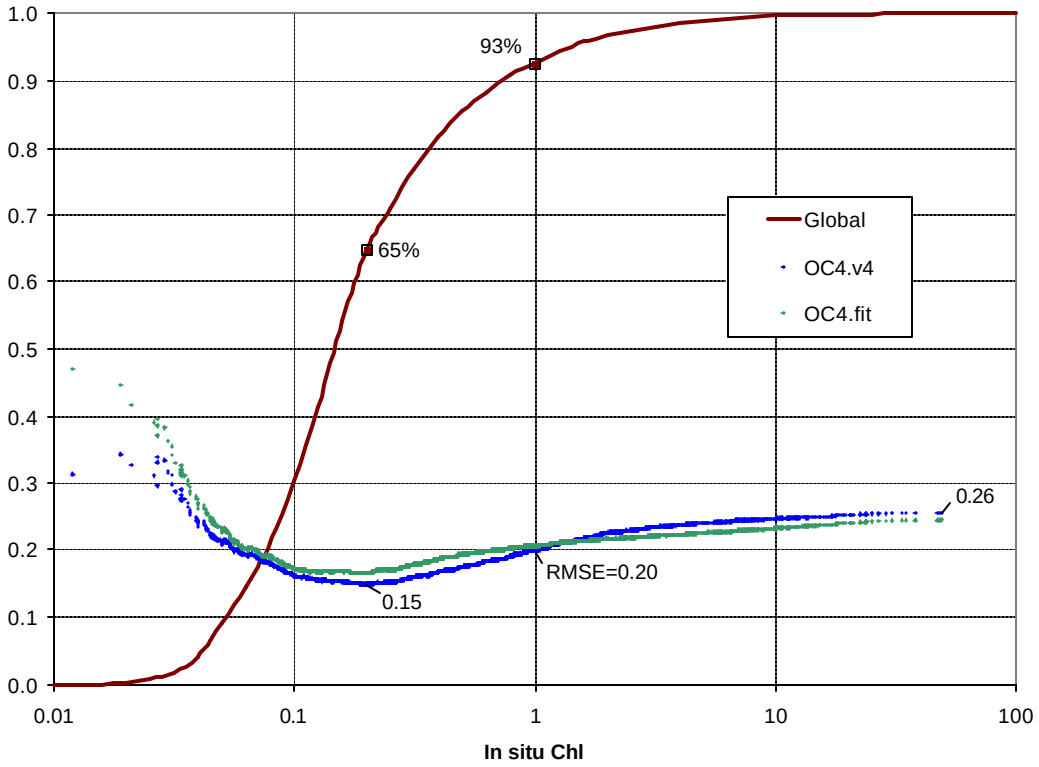

Fig. 8. The blue and green curves are the cumulative RMSEs for the two algorithms for stations with chlorophyll less than the value shown on the horizontal axis. Also shown is the cumulative distribution of Chl from the SeaWiFS climatology (1997-2005).

The lowest RMSE achieved with these algorithms is that of the OC4.v4 algorithm which has RMSE = 0.15 at chlorophyll levels below 0.2 mg m^{-3}. Based on the cumulative frequency curve, this is representative of 65% of the ocean. The RMSE = 0.20 is representative of 93% of the ocean where chlorophyll is below 1 mg m^{-3}. The RMSE values based on all the NOMAD data are 0.256 for OC4.v4 and 0.245 for OC4.fit. These values are clearly biased by the over-representation of high chlorophylls in that data set.

Weighted statistics

To arrive at a more globally representative assessment of an algorithm's uncertainty, its errors should be weighted to reflect the global distribution of chlorophyll. Figure 9 shows the bias and RMS errors within narrow bins along the log(Chl) axis. These were weighted by the relative frequency in each bin to arrive at more representative metrics for the performance of the algorithms (Table 3).

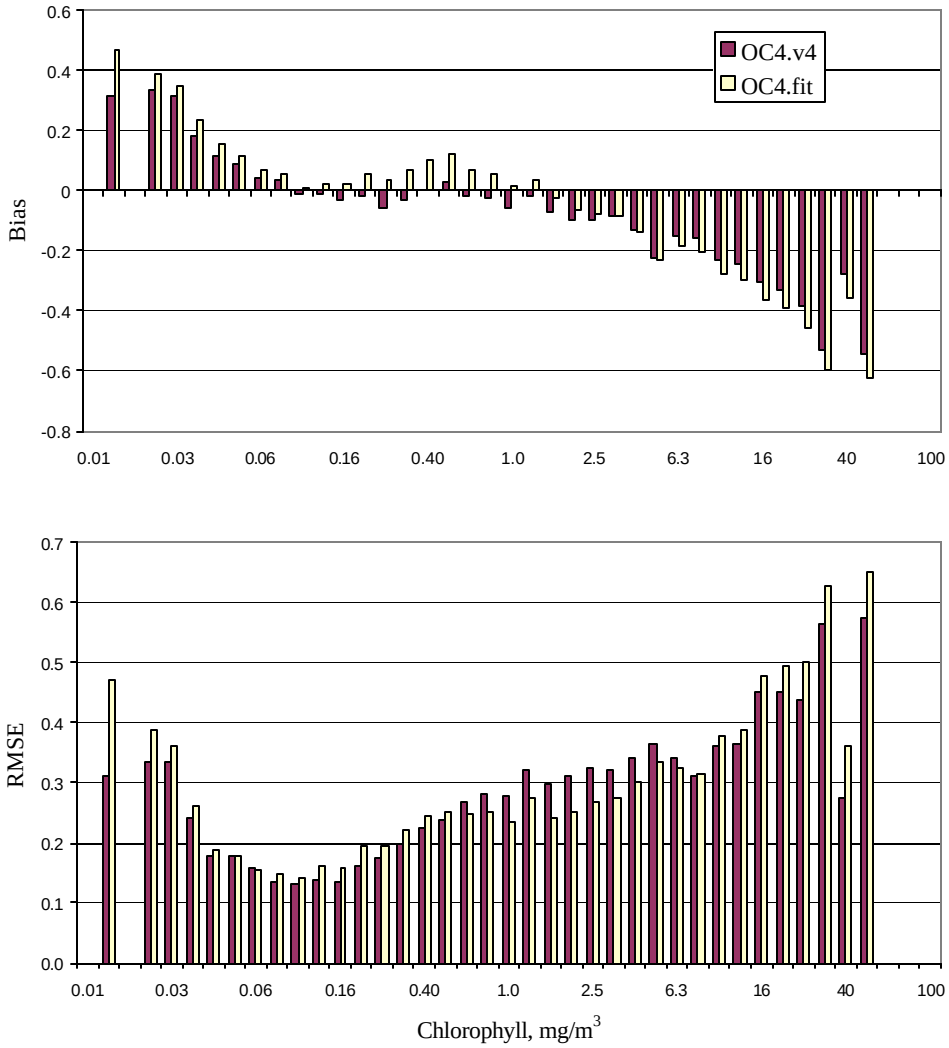

Fig. 9. Error statistics within narrow bins in log(Chl).

Table 3. Globally representative error statistics arrived at by weighting binned statistics (Fig. 9) by the frequency in each log(Chl) bin as determined by the 1997-2005 SeaWiFS climatology (Fig. 6). Results are shown in the two right columns. The unweighted results previously presented (Tables 1 and 2) are shown on the left for comparison.

	Unweighted Statistics		Weighted Statistics	
	OC4.v4	OC4.fit	OC4.v4	OC4.fit
δ (log units)				
bias	-0.047	0.000	-0.006	0.047
RMSE	0.256	0.245	0.193	0.199
relerr (%)				
mean	6%	17%	9%	23%
median	-10%	0%	-1%	11%
std dev	67%	72%	51%	58%
$\hat{C}/C$				
lower	0.50	0.57	0.63	0.71
median	0.90	1.00	0.99	1.11
upper	1.60	1.76	1.54	1.74

Discussion and conclusions

Metrics for evaluating the uncertainty of a chlorophyll algorithm have been presented. These procedures can be applied to evaluate other algorithms or predictive models where the dynamic range of interest spans several orders of magnitude and calls for relative rather than absolute errors. We recommend that the error be defined as the difference between log-transformed values of predicted and measured chlorophyll. A large data set such as NOMAD can be used to create a statistical sample of errors that can be characterized in terms of the mean error (bias) and root-mean-square error (RMSE). In the example of the OC4 algorithm described here, the log errors were normally distributed, and thus the ratio of predicted to measured was lognormally distributed. Statistics of the relative errors can be derived from the same data set, or estimated based on an assumed normal distribution of the log errors.

There are two problems associated with the use of the NOMAD data to evaluate algorithms. One is that its chlorophyll distribution is quite different from that of the world's ocean. Thus, adjustments should be made to account for differences between the distributions. Calculation of a cumulative RMS error (fig. 8) is an informative way to assess the effects of the errors as a function of the chlorophyll level. This was appropriate in the case of the NOMAD data and the OC4 algorithms which performed better at low chlorophyll levels than in high chlorophyll areas. On the other hand, if there are large errors at the low end of the chlorophyll distribution, the cumulative RMS will be inflated by those errors thoughout its range. A more rigorous approach

is to sort the data into bins defined for different ranges of log Chl, calculate the bias and RMS errors in each bin, and weight the results by the global frequency of chlorophyll associated with each bin. When this was done, the OC4.v4 algorithm performed better than the OC4.fit algorithm that was fitted to the NOMAD data.

A second problem concerns the fact that the NOMAD data have been used in parameterizing the algorithms. In principle, the data used to evaluate an algorithm's performance should be independent from the data used to parameterize the algorithm. At the Ocean Color Bio-optical Algorithm Mini Workshop (OCBAM) held at the University of New Hampshire in September 2005, participants debated the importance of this requirement. Some argued that the data should be randomly divided into two data sets and only one set used to parameterize the algorithm, while the other set used to evaluate it. Others argued that this would prove very little because half the data randomly selected would have essentially the same distribution as the whole data set. It was decided instead to create a data set in which chlorophyll measured *in situ* is matched with satellite radiances measured by SeaWiFS or MODIS. The *in situ* data will not include any of the NOMAD data. This data set will be used to evaluate algorithms.

The Ocean Biology Data Processing Group at NASA Goddard is prepared to run algorithm codes and produce metrics for anyone wishing to have an algorithm considered as a candidate for the next generation of ocean color algorithms. The OCBAM participants concluded that little improvement in the chlorophyll algorithm can be achieved without accounting for the effects of other optically active constituents. It is generally believed that this will require a model-based algorithm instead of the empirical algorithms such as OC4.

Since the early days of the Coastal Zone Color Scanner, the goal has been to achieve a chlorophyll algorithm that was accurate to within ±35%. Many claims have been made that this goal has been achieved, and yet the accuracy of the OC4.v4 algorithm has much less accuracy according to the methods described here. After adjusting for the distribution of chlorophyll globally, the algorithm is accurate to within ±50% of the median (based on the standard deviation of the relative error). The algorithm-derived chlorophyll would be on average 9% higher than an *in situ* chlorophyll, but log-transformed chlorophyll values are nearly unbiased.

Where did the ±35% specification come from? According to Jim Mueller, who was a member of the Nimbus Experiment Team for the CZCS (pers. comm.), it was understood that the *in situ* methods for measuring chlorophyll concentration were no more accurate than ±35%, and therefore, the satellite algorithm should not be held to a higher accuracy. This became the goal for the chlorophyll algorithm accuracy. Because it has been generally believed that this goal was achieved, many erroneous methods have been employed to substantiate this claim.

We have demonstrated that the uncertainty in the SeaWiFS algorithm is approximately 50% globally and that the lowest uncertainty is associated with oligotrophic areas where chlorophyll is less than 0.2 mg m^{-3}. Empirical algorithms such as OC4.v4 are parameterized by fitting polynomials to log-transformed data, and this results in lognormally distributed errors whose statistics can be estimated from the statistics of the errors in log Chl.

Finally, we have pointed out several erroneous methods for estimating the relative uncertainty of the algorithm, and have shown that measures of uncertainty based on a large *in situ* database such as NOMAD are sensitive to the distribution of the data in the database.

References

O'Reilly, J.E., S. Maritorena, B.G. Mitchell, D.A. Siegel, K.L. Carder, S.A. Garver, M. Kahru, and C. McClain (1998), "Ocean color chlorophyll algorithms for SeaWiFS" *Journal of Geophysical Researc*h, 103, 24937-24953.

O'Reilly, J.E., S. Maritorena, M.C. O'Brien, D.A. Siegel, D. Toole, D. Menzies, R.C. Smith, J. L. Mueller, B.G. Mitchell, M. Kahru, F.P. Chavez, P. Strutton, G.F. Cota, S.B. Hooker, C. McClain, K.L. Carder, F. Muller-Karger, L. Harding, A. Magnuson, D. Phinney, G.F. Moore, J. Aiken, K.R. Arrigo, R. Letelier, and M. Culver (2000), "Ocean color chlorophyll *a* algorithms for SeaWiFS, OC2, and OC4: Version 4." In S.B. Hooker and E.R. Firestone (Eds.), *SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3* (pp. 9-23). NASA Tech. Memo. 2000-206892, Vol. 11, Greenbelt: NASA Goddard Space Flight Center.

Werdell, J. and S. Baily, "An improved in situ bio-optical data set for ocean color algorithm development and satellite data product validation," *Remote Sensing of the Environment*, (submitted, March 2005).