The empirical chlorophyll algorithm for MODIS: Testing the OC3M algorithm using NOMAD data

Janet W. Campbell & Hui Feng
University of New Hampshire
September 22, 2005

In preparation for a workshop to be held in September 2005 to consider semi-analytical algorithms for MODIS and SeaWiFS, we have evaluated the current empirical algorithms, OC4.v4 used for SeaWiFS and OC3M used for MODIS. Both are documented in Vol. 11 of the SeaWiFS Post-launch TM series. This paper details the evaluation of the OC3M algorithm using the newly published NOMAD data set (Werdell *et al*. 2005). Results of the OC4.v4 evaluation are presented in a separate paper.

Our purpose is to quantify the uncertainty associated with the OC3M algorithm (either the published version or a reparameterized version, OC3M.v2), thus establishing a baseline against which to compare any alternative algorithms that might be proposed at the workshop. This has been done with a subset of NOMAD (N = 2208) as explained below, and with a further partitioning of this dataset into stations with HPLC chlorophylls only (N = 870) and those with fluorometric chlorophylls only (N = 1338). The relationship between an RMS error expressed in log units and the relative or percentage error is addressed and explained.

Methods

The NOMAD dataset contains data for water-leaving radiance, $L_w(\lambda)$, and downwelling surface irradiance, $E_s(\lambda)$, in 20 bands from 405 to 683 nm. We eliminated stations having missing $L_w(\lambda)$ or $E_s(\lambda)$ data in any of the first 5 SeaWiFS bands. The resulting subset, containing 2208 stations, was designated the Evaluation Data Set to be used to test algorithms at the workshop. Remote-sensing reflectance, $R_{rs}(\lambda)$, was calculated as the ratio of $L_w(\lambda)$ to $E_s(\lambda)$ for all bands.

For the initial results, we use the HPLC chlorophyll ("chl_a") if it is present; otherwise, we use the fluorometric chlorophyll ("chl"). Later, we distinguish results for the two methods of measuring chlorophyll.

The form of the OC3M algorithm is:

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

where

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

and the coefficients a_0, a_1, a_2, a_3, a_4 are 0.283, -2.753, 1.457, 0.659, and -1.403, respectively. We fitted 4th-order polynomials to plots of log[Chl] vs. X and compared these to the OC3M algorithm. The algorithm with coefficients fitted to the NOMAD data is called OC3M.v2.

Results

Figure 1 is a plot of log[Chl] vs. X. Note that the OC4.v4 algorithm curve uses $R_{rs}(510)$ when that is maximum, whereas the OC3M uses only the 443 and 489 bands. The NOMAD data are plotted against the log max Rrs/Rrs555 values used for OC3M.

Fig. 1 - NOMAD chlorophyll vs. max Rrs ratio on log-log scale. Also shown is the OC3M algorithm (blue line) and OC4.v4 (red line). Only stations having measured Rrs in first 5 SeaWiFS bands are shown (N = 2208).

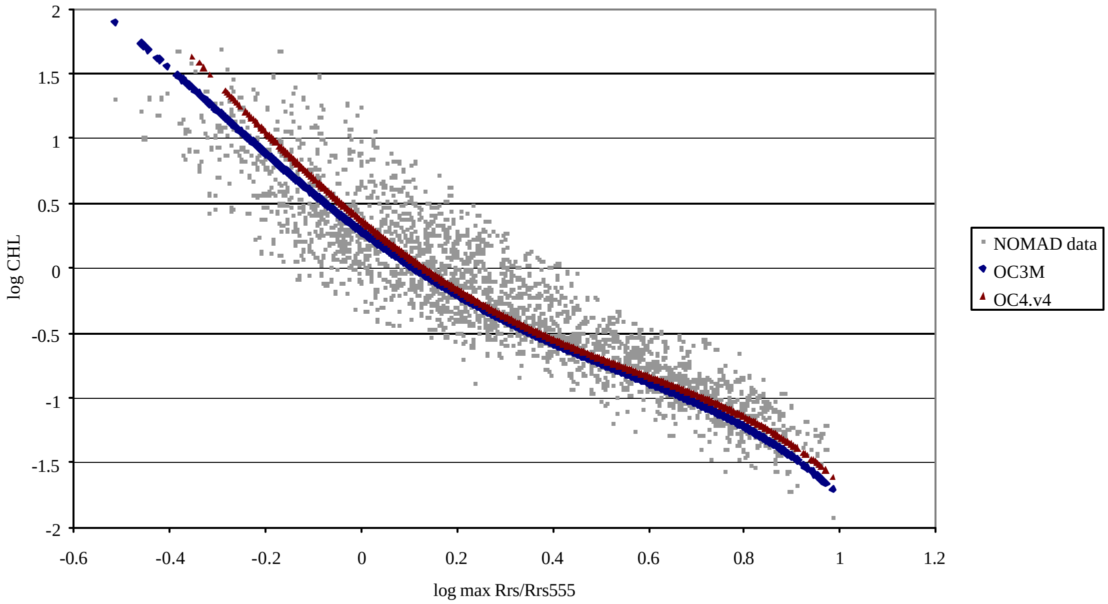


Fig. 2 - A fitted polynomial (OC3M.v2) is compared with OC3M algorithm.

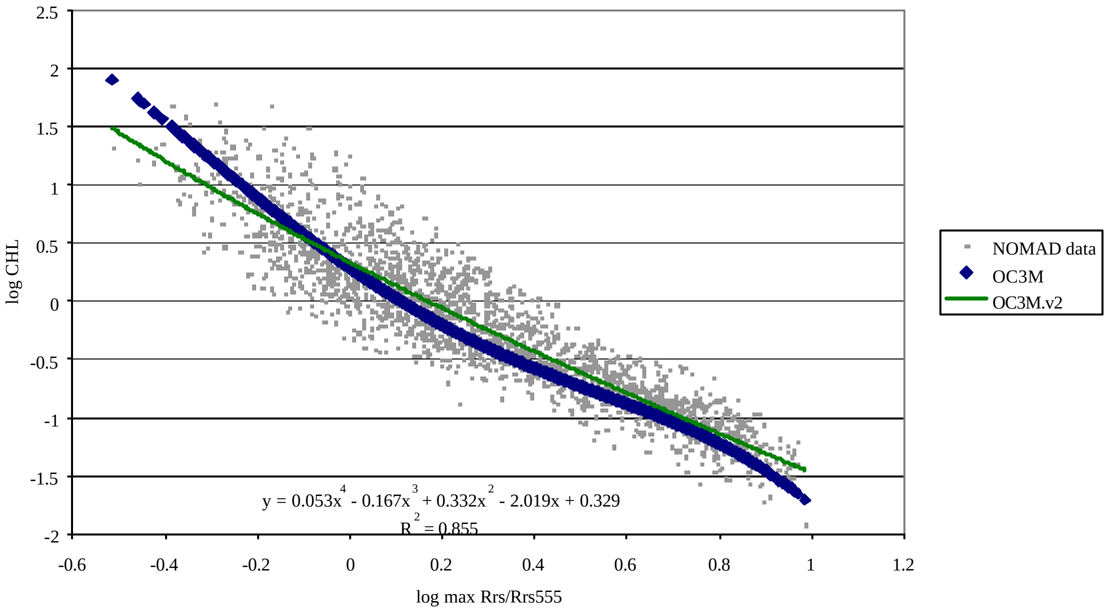
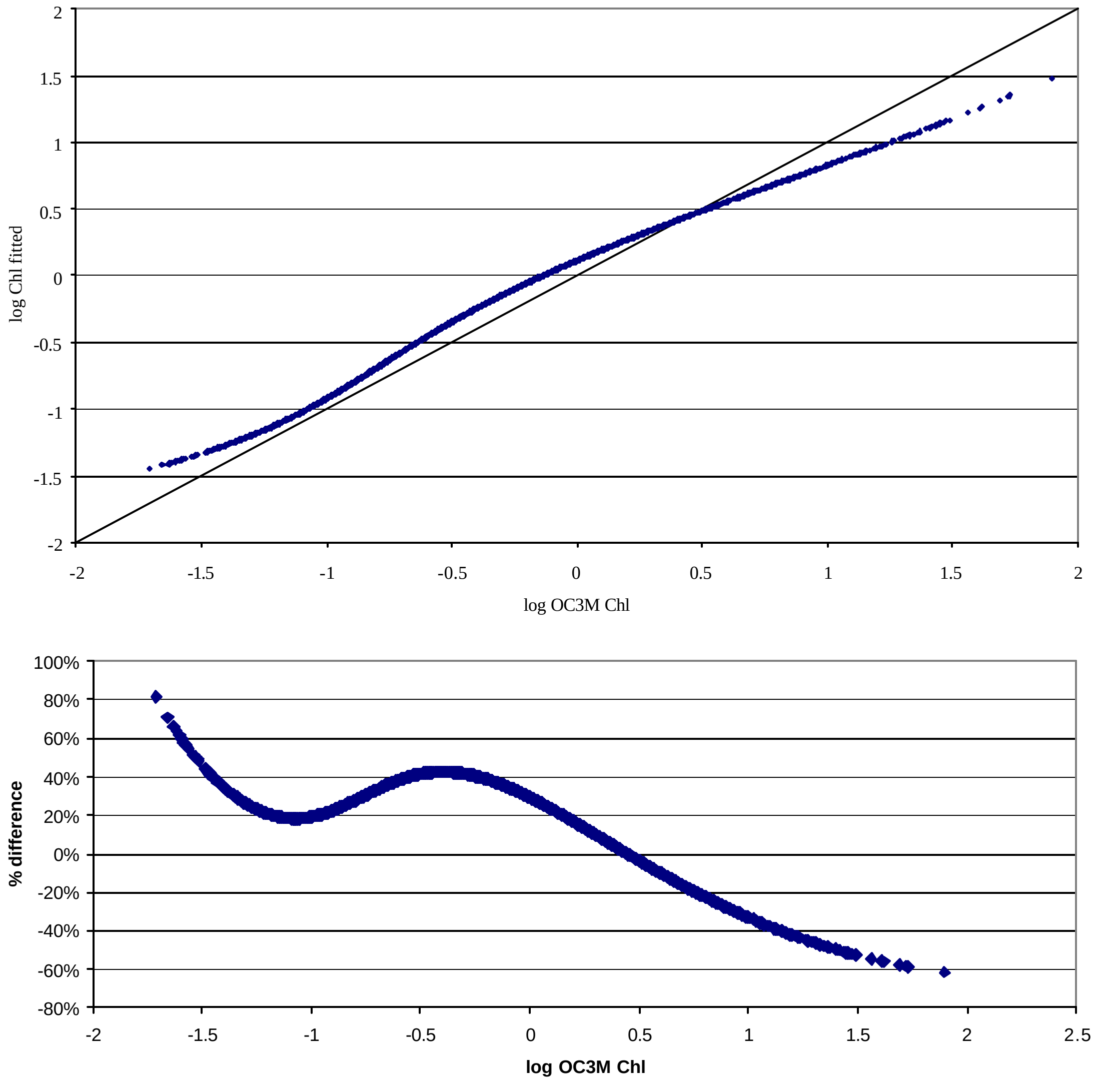


Figure 3 compares chlorophyll derived by the fitted polynomial to that derived with the OC3M algorithm. The fitted polynomial yields higher chlorophyll values below 2.5 mg m^{-3}, with values up to 43% higher at OC3M Chl = 0.4 mg m^{-3} and even higher values for OC3M Chl < 0.03 mg m^{-3}. Above 2.5 mg m^{-3} the fitted polynomial yields chlorophylls that are lower than the OC3M algorithm. In this range, values are up to 60% lower.

Fig. 3 - Comparison of the chlorophyll algorithm fitted to the NOMAD data vs. the OC3M algorithm.



Does the chlorophyll method (HPLC vs. fluorometric) make a difference?

Figure 4 shows both the HPLC (red symbols) and fluorometric (grey symbols) on a plot similar to figure 1. Also shown are polynomial curves fitted to the two subsets. Figure 5a compares the HPLC data and fitted curve to the OC3M curve, and figure 5b compares the fluorometric data and fitted curve to the OC3M curve. Chlorophyll computed with the fitted curves for both subsets are plotted against the OC3M chlorophyll in figure 6.

Fig. 4 - Comparison between HPLC and fluorometric chlorophyll vs. max Rrs ratio. Separate polynomials are fitted to each data set.

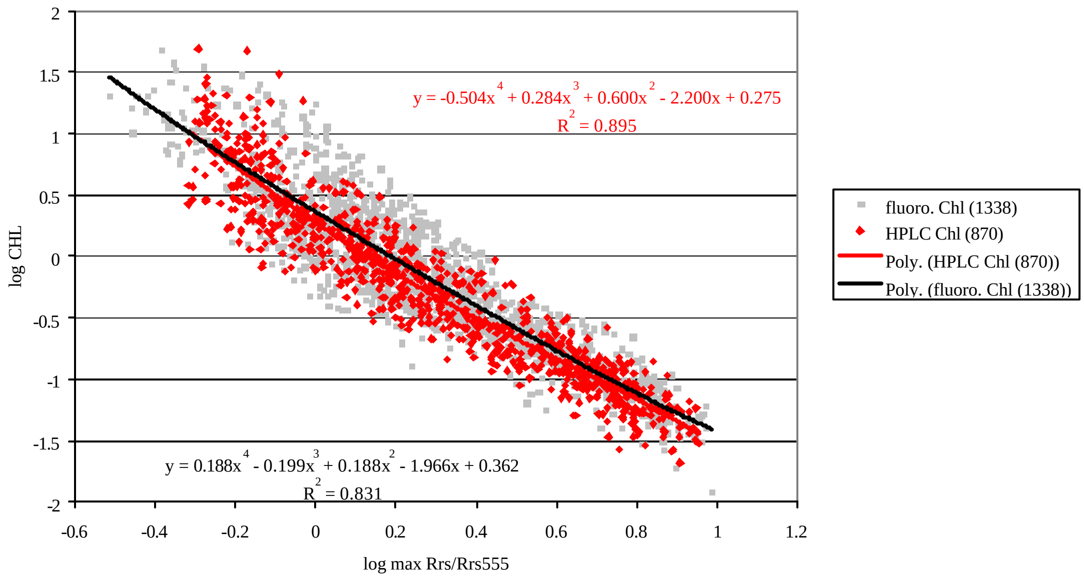


Fig. 5a - Comparison of the algorithm fitted to the HPLC data vs. the OC3M algorithm.

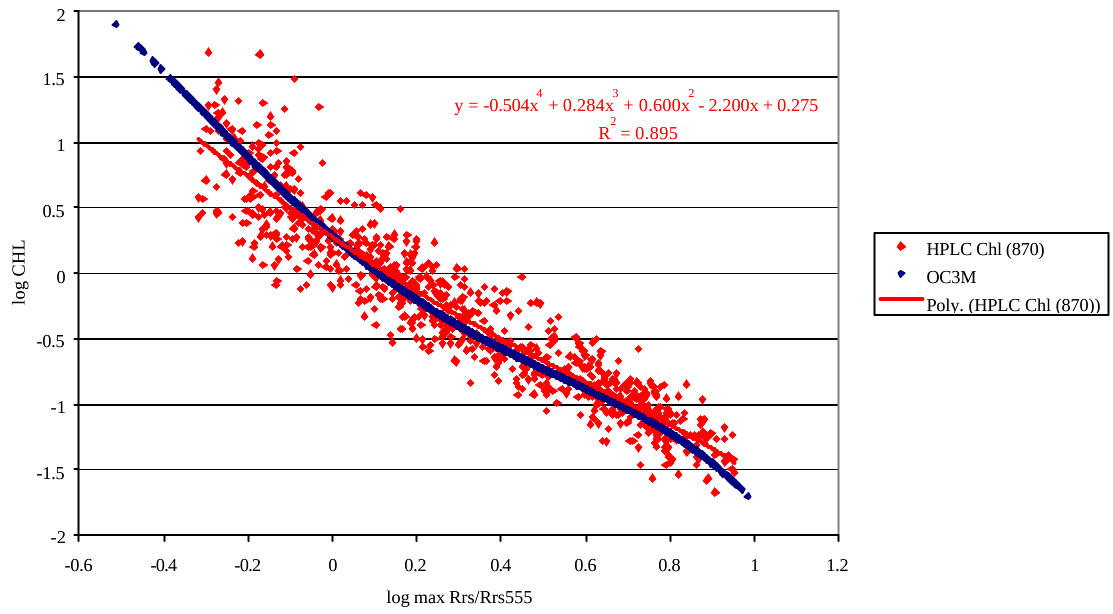
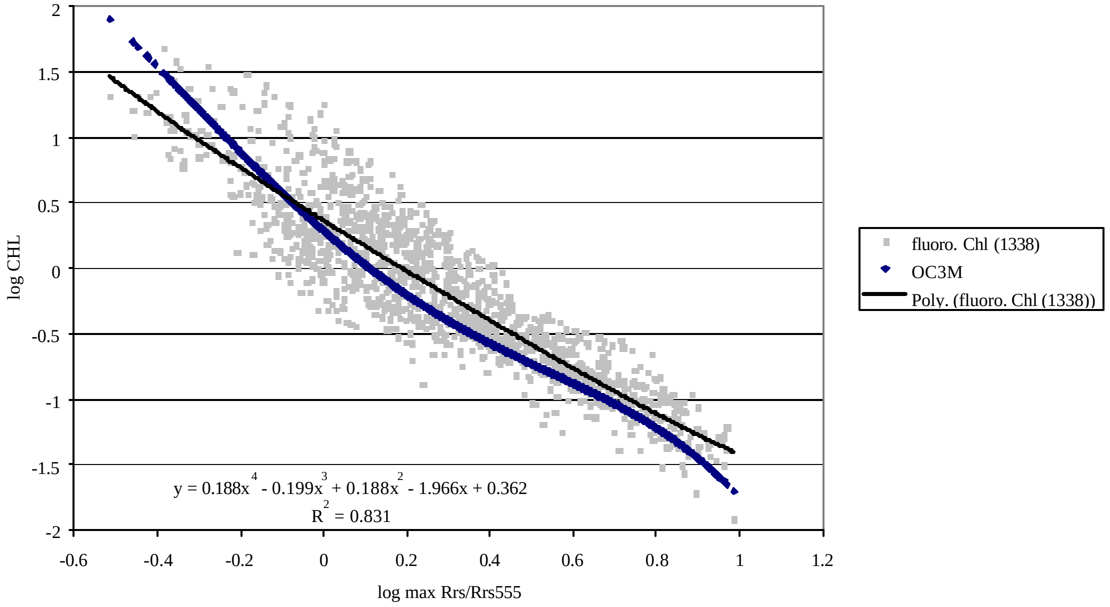
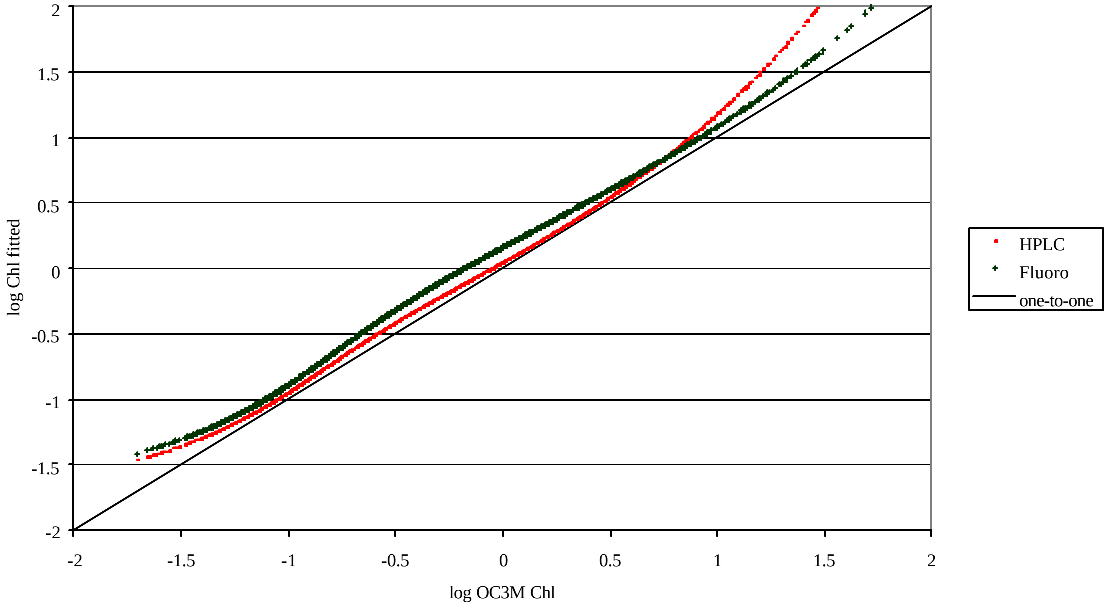


Fig. 5b - Comparison of the algorithms fitted to the fluorometric chlorophylls.



Although it is hard to see any systematic difference between the HPLC and fluorometric chlorophyll data in figure 5, the HPLC polynomial had a better fit (R^2 = 0.90, N = 870) compared with the fluorometric fit (R^2 = 0.83, N = 1338). Moreover, the HPLC polynomial was closer to the OC3M curve than the fluorometric one (Fig. 6) except at high chlorophyll levels. Our conclusion is that there is more “noise” in the fluorometric chlorophyll measurements.

Fig.6 - Polynomials fitted to the HPLC and fluorometric chlorophylls vs. the OC3M algorithm.



Quantifying Uncertainty in the Empirical Algorithm

There are two issues related to quantifying uncertainty. One is the interpretation of errors in a log-log regression, and the other concerns the distribution of the data used to fit the polynomial compared with the distribution in the world's ocean. The second issue is discussed in the *Evaluation of OC4* paper. Here we repeat the equations for converting the log-based statistics to percentage errors and present results for the OC3M algorithm.

Table 1 lists the polynomial coefficients for the fits to all the data, and to the HPLC and fluorometric chlorophyll subsets. Also shown are error statistics associated with these fits. Note that in each case, the fits eliminate the average error (bias) in the log-log regressions.

Table 1. Polynomial fits to the NOMAD data. Coefficients are are defined based on equation (1). Error statistics based on the samples of size N where errors are defined by equation (4).

Variable	OC3M	all data	HPLC	fluoro
N	2208	2208	870	1338
a_0	0.283	0.339	0.275	0.362
a_1	-2.753	-2.019	-2.200	-1.966
a_2	1.457	0.332	0.600	0.189
a_3	0.659	-0.167	0.284	-0.199
a_4	-1.403	0.053	-0.504	0.188
Error Statistics (log-log)				
bias	-0.077	0.000	0.000	0.000
RMSE	0.277	0.249	0.222	0.260
R^2	0.84	0.86	0.90	0.83

How should log-log error statistics be interpreted?

In fitting polynomials to log-log data, the resulting curves minimize the mean squared error (MSE) between the logarithm of the predicted chlorophyll and the logarithm of the measured chlorophyll. That is, they minimize:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\log \hat{C}_i - \log C_i)^2 \qquad (3)$$

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

$$\delta_i = \log \hat{C}_i - \log C_i = \log\left(\frac{\hat{C}_i}{C_i}\right) = \log\left(1 + \mathrm{relerr}_i\right) \qquad (4)$$

where relerr_i is the relative error given by:

$$\mathrm{relerr}_i = \left(\frac{\hat{C}_i - C_i}{C_i} \right) \qquad (5)$$

Since there's a direct relationship between δ_i and relerr_i, this method is generally considered to minimize relative errors. This is appropriate for chlorophyll algorithms because of the large dynamic range of chlorophyll values found globally which vary by 4 orders of magnitude. The reasoning is that a 10% error in the open ocean is as important as a 10% error in coastal waters.

Table 1 gives estimates of the bias and RMSE for δ_i, but what do these tell us about statistics of the relative error? What about the mean and "one sigma" range for the relative error? These statistics can be calculated empirically from the data (see Table 2) or we can estimate them from the data in Table 1 given some reasonable assumptions.

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*:

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

where *m* = bias. Under this assumption, the ratio $\hat{C}/C$ is lognormally distributed. To compute the mean, median, and standard deviation of $\hat{C}/C$, we need the mean and standard deviation of $\ln(\hat{C}/C)$ which are given by M = *m* ln(10) and S = *s* ln(10). The statistics of $\hat{C}/C$ are:

$$\text{mean} \quad = \quad \exp\left(M + \frac{S^2}{2} \right) \qquad (7)$$

$$\text{median} \quad = \quad \exp(M) \qquad (8)$$

$$\text{std dev} \quad = \quad \text{mean} \sqrt{\exp(S^2) - 1} \qquad (9)$$

Statistics for relative errors associated with the polynomial fits in Table 1 are given in Table 2. We show statistics calculated empirically and based on the lognormal assumption (eqs. 7-8). The mean and median percentage errors are derived by subtracting 1 from the mean (eq. 7) or median (eq. 8) and then multiplying by 100%. The standard deviation of the percentage error is the same as the standard deviation of $\hat{C}/C$ (eq. 9) multiplied by 100%.

Table 2. Statistics of the percentage errors associated with the polynomials fitted to the NOMAD data.

relerr (%)	OC3M	all data	HPLC	fluoro
mean	2%	17%	14%	19%
median	-7%	2%	0%	2%
std dev	73%	68%	61%	74%
Statistics based on lognormal assumption (eq. 7-9)				
mean	1%	18%	14%	20%
median	-16%	0%	0%	0%
std dev	68%	74%	62%	79%

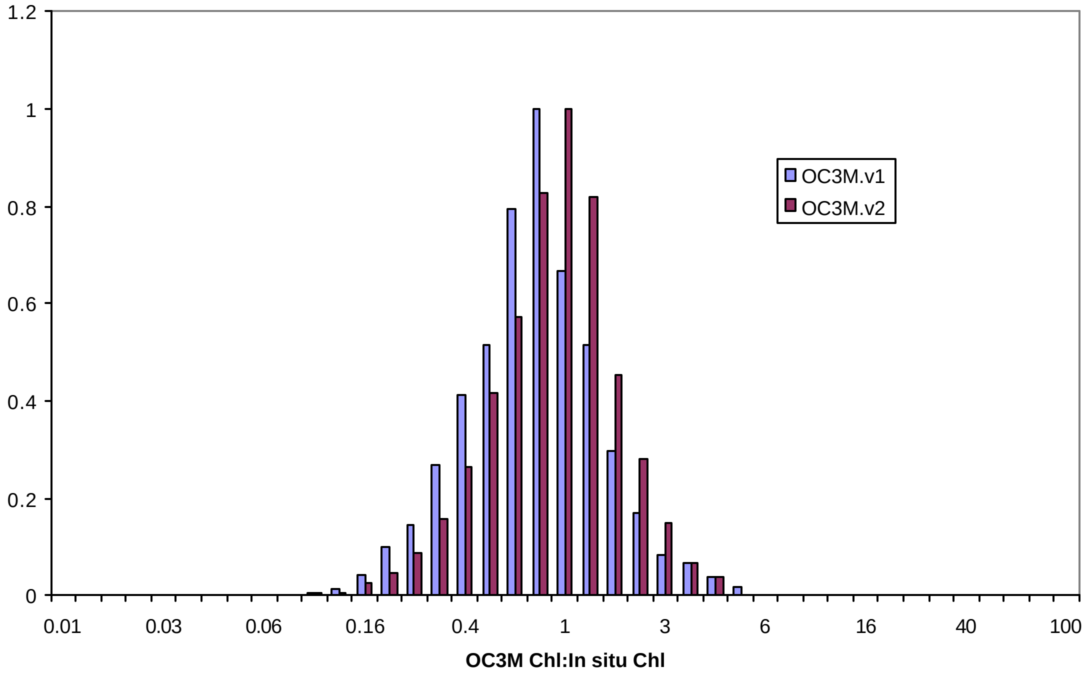
The first thing to notice is that the errors are no longer unbiased; the mean relative error is positive. The median relative error is close to zero, or would be zero under the assumption of a lognormally distributed error. The second point is that the standard deviations are quite large; it doesn't make sense to think of the errors as have a range of ±3 standard deviations as would be the case if errors were normally distributed.

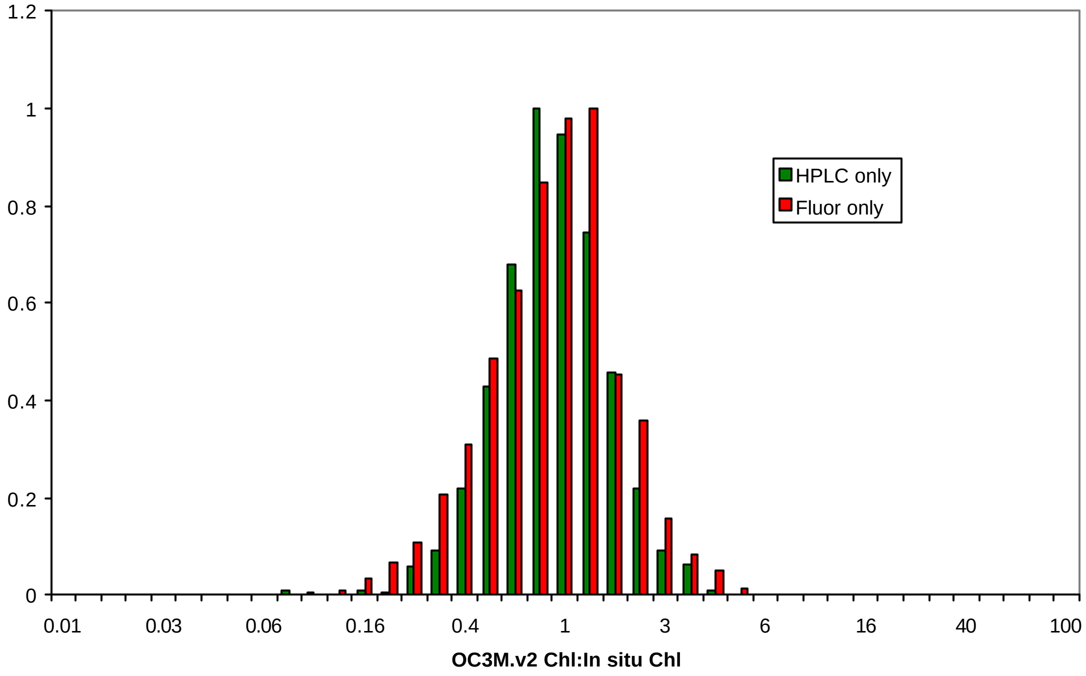
Expressing uncertainty in terms of ±1 standard deviation has meaning if the distribution is symmetric about the mean, but in the case of large relative errors, the distribution is skewed. Negative errors can't be larger than -100% whereas positive errors can be arbitrarily large. Use of the log error δ helps to alleviate this problem, but then the units are decades of log which are not easily interpreted.

Error histograms are a good way to express errors, where the horizontal axis is on a log scale (see figure 7). 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%.

On the following page:

Fig. 7. Histograms of the log error (δ). The horizontal axis labels are the ratio of the OC3M-derived chlorophyll to the NOMAD chlorophyll. (a) Comparison of δ distributions for the OC3M algorithm currently in use and a re-parameterized version based on a 4th-order polynomial fitted to the NOMAD data. (b) Comparison of δ distributions for the polynomials fitted to the HPLC and fluorometeric chlorophylls separately. Error statistics are shown in Tables 1 and 2.


1.2
1
0.8
0.6
0.4
0.2
0
0.01
0.03
0.06
0.16
0.4
1
3
6
16
40
100
OC3M Chl:In situ Chl
OC3M.v1
OC3M.v2


1.2
1
0.8
0.6
0.4
0.2
0
0.01
0.03
0.06
0.16
0.4
1
3
6
16
40
100
OC3M.v2 Chl:In situ Chl
HPLC only
Fluor only