The empirical chlorophyll algorithm for SeaWiFS: Testing the OC4.v4 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 OC4.v4 algorithm using the newly published NOMAD data set (Werdell *et al*. 2005). Results of the OC3M evaluation are presented in a separate paper.

Our purpose is to quantify the uncertainty associated with the OC4 algorithm (either OC4.v4 or a reparameterized version, OC4.v5), 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). In assessing the uncertainty, we address two issues: (1) what is the relationship between an RMS error expressed in log units and the relative or percentage error? and (2) should we adjust for differences between the data distribution and that of the global ocean chlorophyll? If so, how?

Methods

The NOMAD dataset contains data from 3,467 stations. Data provided include water-leaving radiance, $L_w(\lambda)$, and downwelling surface irradiance, $E_s(\lambda)$, in 20 bands from 405 to 683 nm. We calculated the remote-sensing reflectance, $R_{rs}(\lambda)$, as the ratio of $L_w(\lambda)$ to $E_s(\lambda)$ for the bands used in the OC4 algorithm: 443, 489, 510, and 555 nm. Initially, we eliminated stations having missing $R_{rs}(\lambda)$ data in any of these 4 band with the following exceptions. There were 275 stations where $R_{rs}(555)$ was missing, but $R_{rs}(550)$ and $R_{rs}(560)$ were measured. For these stations, $R_{rs}(555)$ was estimated by linear interpolation, and flagged to distinguish these stations from the others. In addition, there were 9 stations where $R_{rs}(510)$ was missing and $R_{rs}(520)$ was measured. For these stations, we estimated $R_{rs}(510)$ using equation (3) from the SeaWiFS Post-launch TM Vol. 11, chapter 2. The resulting subset contained data from 2514 stations.

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 OC4.v4 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

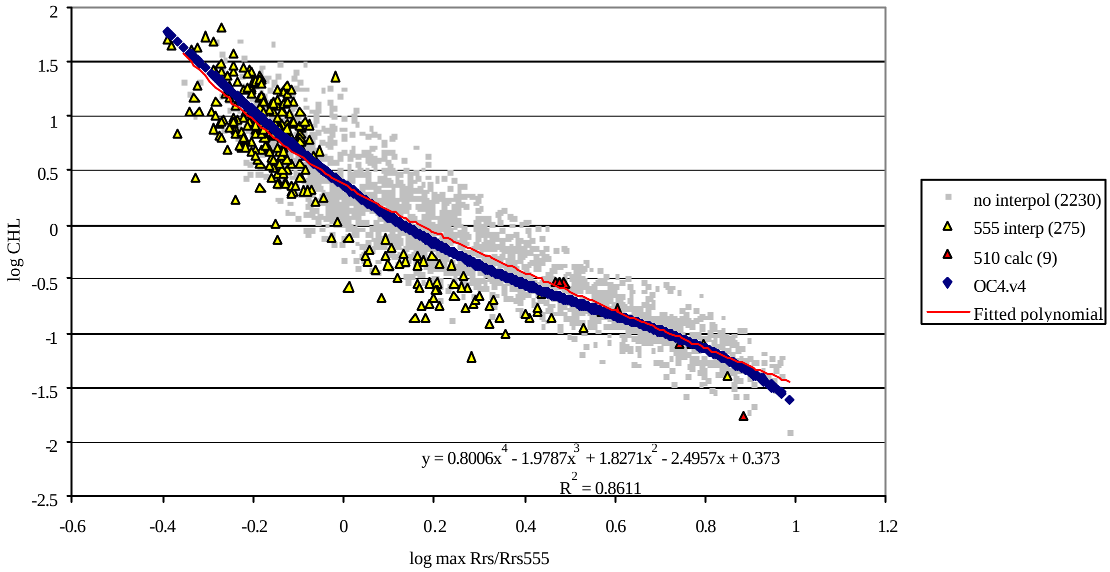
$$X = \log\left[\frac{\max(R_{rs}(443), R_{rs}(489), R_{rs}(510))}{R_{rs}(555)}\right] \qquad (2)$$

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. We fitted 4th-order polynomials to plots of log[Chl] vs. X and compared these to the OC4 algorithm. The algorithm with coefficients fitted to the NOMAD data is called OC4.v5.

Results

Figure 1 is a plot of log[Chl] vs. X for all the data in the subset. The points at which $R_{rs}(555)$ was interpolated are shown as yellow triangles, and the ones in which $R_{rs}(510)$ was calculated are red triangles. All other points are shown as grey squares. The 4th-order polynomial fitted to the remaining 2230 points is the red line, and OC4 is the blue line.

Fig. 1 - NOMAD chlorophyll vs. max Rrs ratio on log-log scale (N = 2514). Also shown is OC4.v4 algorithm (blue line) and 4th order polynomial fitted to the data. Yellow triangles are stations where Rrs(555) was interpolated from adjacent bands.



Many of the points having an interpolated $R_{rs}(555)$ are fall below the other points. We thus decided to eliminate those points and also the ones in which $R_{rs}(510)$ was calculated. Subsequently, because we intend to compare the OC4 results with those for semi-analytic algorithms, we eliminated all stations that were missing $R_{rs}(\lambda)$ data in any of the first 5 SeaWiFS bands. In other words, we also eliminated stations that were missing data for the 412 nm band. The resulting subset, containing N = 2208 stations, was designed the official Evaluation Data Set for the workshop. All subsequent plots and analyses are based on this subset.

On the next page, figure 2 shows only the OC4 algorithm compared with the NOMAD chlorophyll data, and figure 3 shows only the fitted polynomial.

Fig. 2 – NOMAD chlorophyll vs. max Rrs ratio on log-log scale. Also shown is the OC4.v4 algorithm (blue line). Only stations having Rrs data in the first 5 bands of SeaWiFS are used (N = 2208).

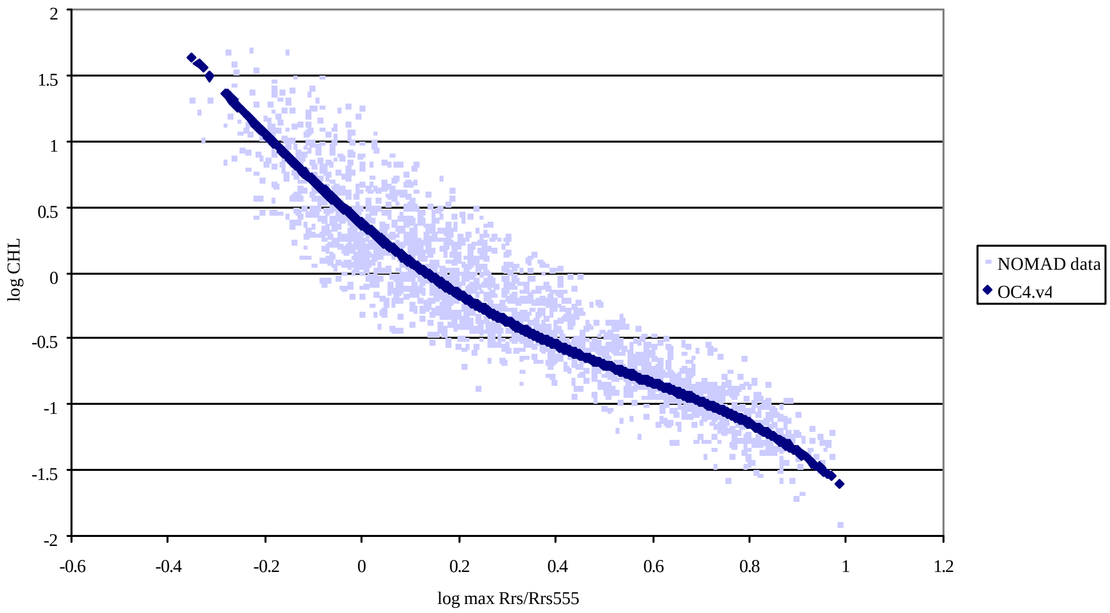


Fig. 3 - Same as figure 2 except the fitted polynomial is shown instead of the OC4 algorithm.

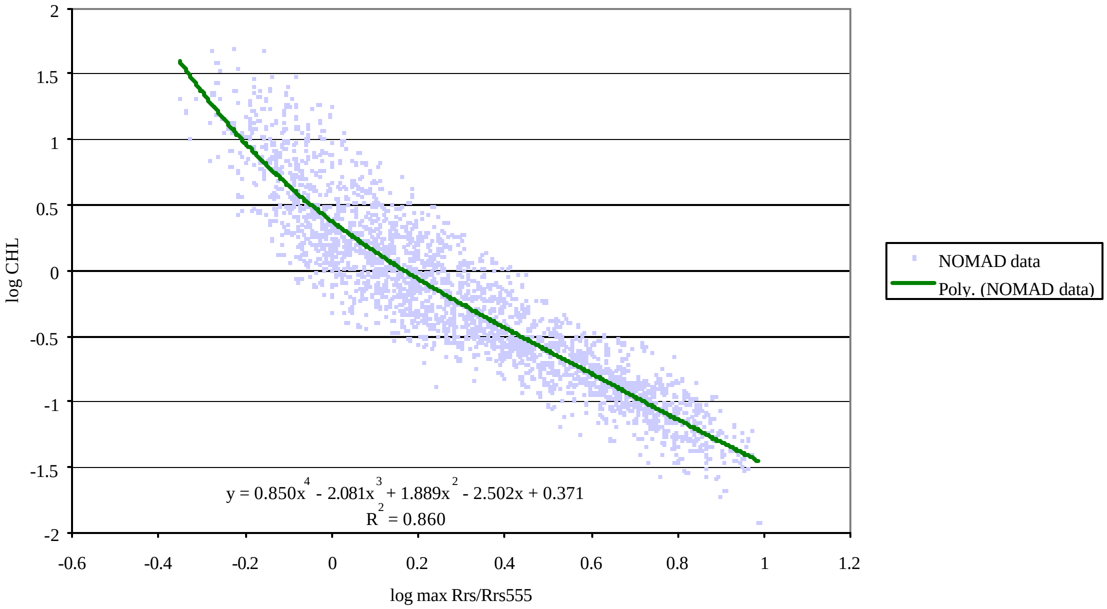
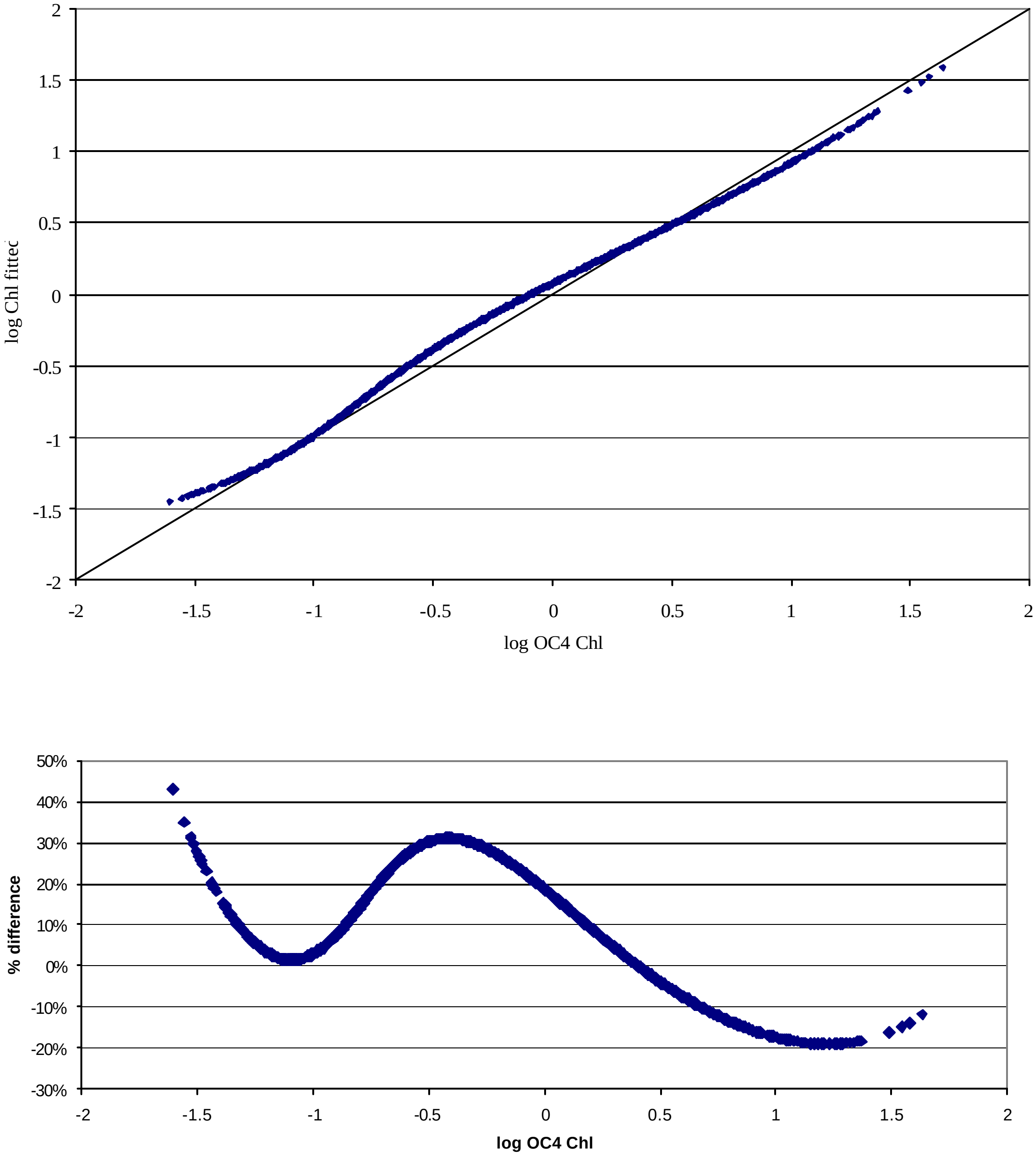


Figure 4 compares chlorophyll derived by the fitted polynomial to that derived with the OC4 algorithm. The fitted polynomial yields higher chlorophyll values below 2.5 mg m^{-3}, with values up to 31% higher at OC4 Chl = 0.4 mg m^{-3}. Above 2.5 mg m^{-3} the fitted polynomial yields chlorophylls that are lower than the OC4 algorithm. In this range, values are up to 20% lower.

Fig. 4 - Comparison of the chlorophyll algorithm fitted to the NOMAD data vs. the OC4.v4 algorithm.



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

Figure 5 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 6a compares the HPLC data and fitted curve to the OC4 curve, and figure 6b compares the fluorometric data and fitted curve to the OC4 curve. Chlorophyll computed with the fitted curves for both subsets are plotted against the OC4 chlorophyll in figure 7.

Fig. 5 - Comparison between HPLC and fluorometric chlorophyll vs. max Rrs ratio. Separate polynomials are fitted to each data set.

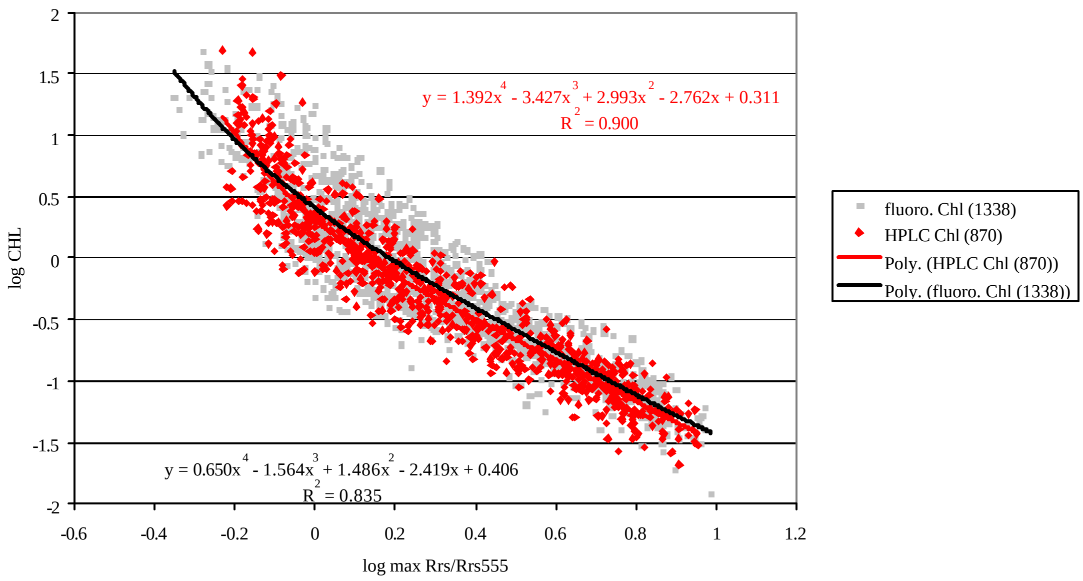


Fig. 6a - Comparison of the algorithm fitted to the HPLC data vs. the OC4.v4 algorithm.

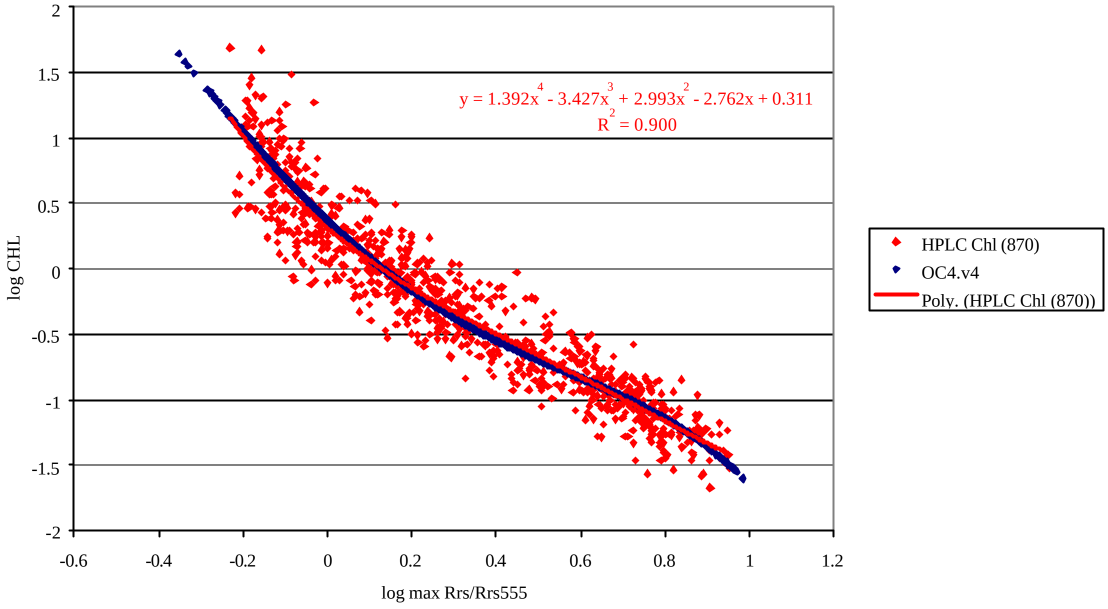
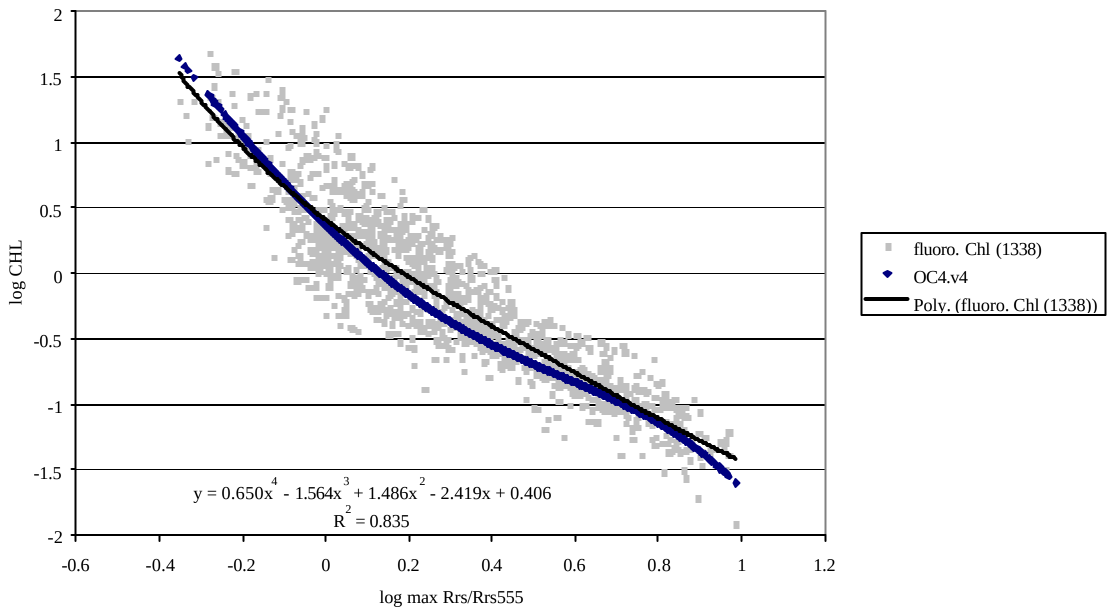
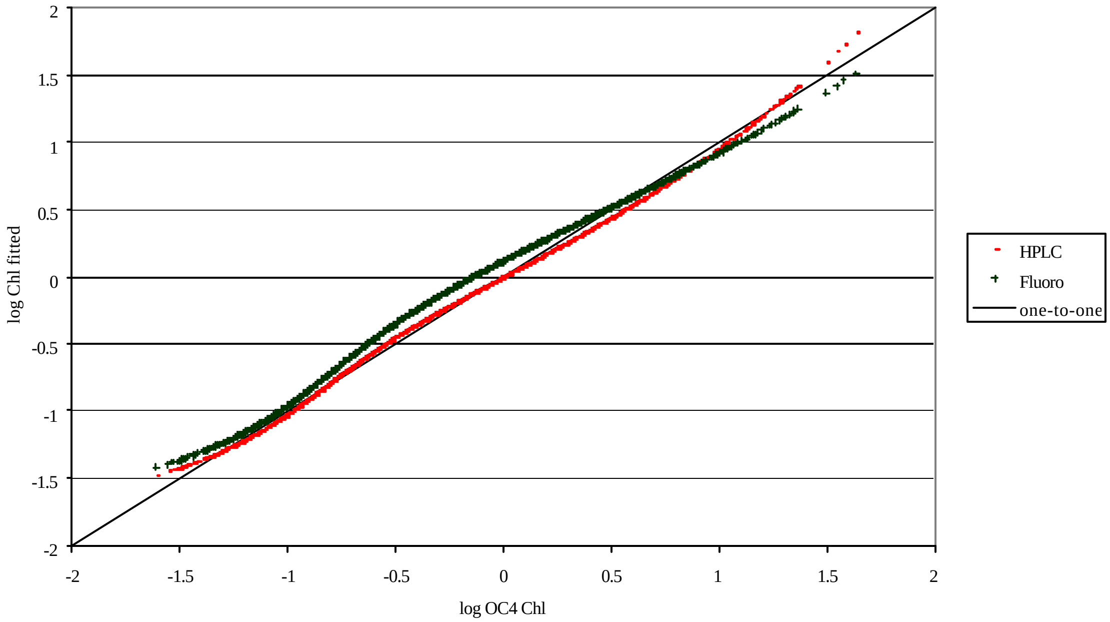


Fig. 6b - Comparison of the algorithms fitted to the fluorometric chlorophylls.



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 = 0.835, N = 1338). Moreover, the HPLC polynomial was closer to the OC4 curve than the fluorometric one (Fig. 7). Our conclusion is that there is more “noise” in the fluorometric chlorophyll measurements.

Fig. 7 - Comparison of the polynomials fitted to the HPLC and fluorometric chlorophylls vs. the OC4.v4 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.

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	OC4	all data	HPLC	fluoro
N	2208	2208	870	1338
a_0	0.366	0.371	0.311	0.406
a_1	-3.067	-2.502	-2.762	-2.419
a_2	1.930	1.889	2.993	1.486
a_3	0.649	-2.081	-3.427	-1.564
a_4	-1.532	0.850	1.392	0.650
Error Statistics (log-log)				
bias	-0.047	0.000	0.000	0.000
RMSE	0.256	0.245	0.217	0.257
R^2	0.85	0.86	0.90	0.84

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:

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

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*:

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

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:

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

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

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

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 (%)	OC4	all data	HPLC	fluoro
mean	6%	17%	13%	19%
median	-7%	4%	7%	5%
std dev	66%	68%	60%	74%
Statistics based on lognormal assumption (eq. 7-9)				
mean	6%	17%	13%	19%
median	-10%	0%	0%	0%
std dev	67%	72%	60%	77%

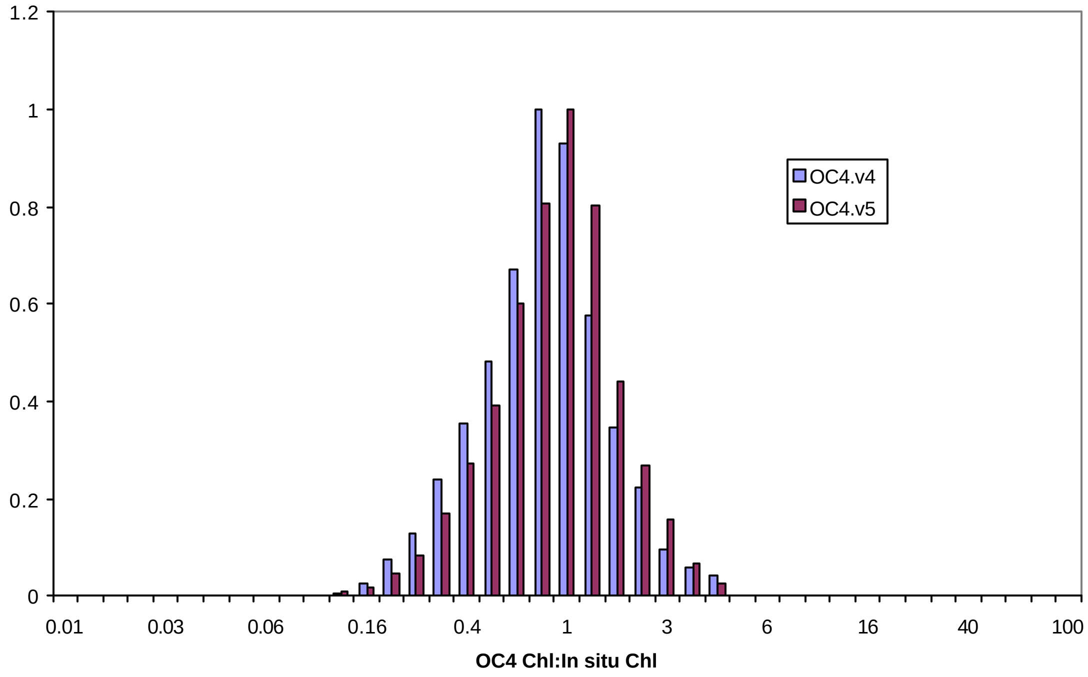
The first thing to notice is that the errors are no longer unbiased; the mean relative error is positive. The median relative error is close to zero, or would be zero under the assumption of a lognormally distributed error. The second point is that the standard deviations are quite large; it doesn't 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.

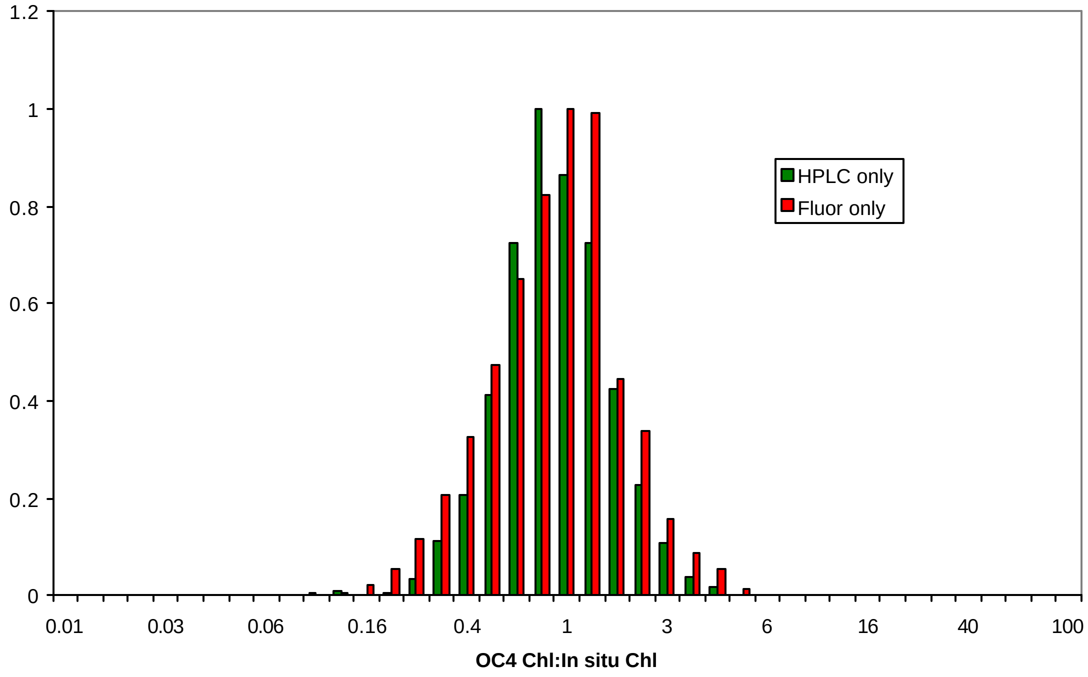
Expressing uncertainty in terms of ±1 standard deviation has meaning if the distribution is symmetric about the mean, but in the case of large relative errors, the distribution is skewed. Negative errors can't be larger than -100% whereas positive errors can be arbitrarily large. Use of the log error 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 8). 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 OC4-derived chlorophyll to the NOMAD chlorophyll. (a) Comparison of δ distributions for the OC4.v4 algorithm currently in use and a re-parameterized version (OC4.v5) 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
OC4 Chl:In situ Chl
OC4.v4
OC4.v5


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
OC4 Chl:In situ Chl
HPLC only
Fluor only

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. 9).

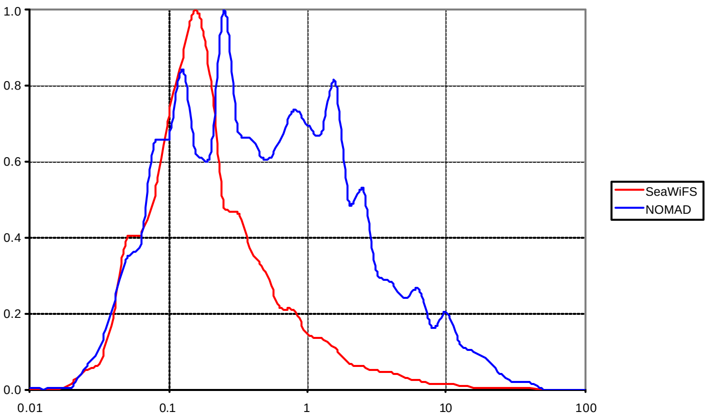


Fig. 9 – Histograms of chlorophyll data. NOMAD data distribution (blue) is compared with global SeaWiFS distribution (red) over its lifetime (1997-2005).

Ideally, the algorithm should achieve the same relative accuracy everywhere, that is, in high as well as low chlorophyll waters. To achieve this, the approach would be to sort the data into bins according to X (= log of the maximum reflectance ratio, equation 2), and fit a polynomial to the bin averages of log(Chl) vs. X. This scheme weights each bin equally regardless of how much data is in that bin. This should improve the fits in the high and low tails of the distribution where there are fewer data points. The average or RMS errors can be calculated for each bin using all the data in the bin, but later weighted by the distribution of X in the world's ocean.

In preparation for the Workshop in September, we will come up with sets of weights based on the global distributions of X and $R_{rs}(\lambda)$ in the SeaWiFS record (1997-present). These can be used to prescribe a single uncertainty value (as a percentage) to algorithms based on X or the remote sensing reflectance spectra.