

## Case Study 17

# Comparison of *In Situ* and Remotely-Sensed Chl-*a* Concentrations: A Statistical Examination of the Match-up Approach

E. Santamaría-del-Ángel<sup>\*1</sup>, R. Millán-Núñez<sup>1</sup>, A. González-Silvera<sup>1</sup> and R. Cajal-Medrano<sup>1</sup>

## 17.1 Background

With the launch of the Coastal Zone Color Scanner (CZCS) in November of 1978, a new era in oceanographic studies began. This was the first instrument dedicated to the measurement of ocean colour using satellite imaging, and its main purpose was to determine whether spectroradiometric observations could be used to identify and quantify suspended or dissolved matter in ocean waters (IOCCG, 2000; 2004; 2006). CZCS imagery encompassed large geographic areas and was collected over short periods of time, something that was not possible with previous measurement techniques (ships, buoys, airplanes). CZCS was a 'proof-of-concept' mission, to determine whether Chlorophyll-*a* concentration (Chl-*a*) could be estimated from space, based on spectrophotometric principles. Studies using CZCS data (Peláez and McGowan 1986; Yoder et al., 1987; Muller-Karger et al., 1991; Santamaria-del-Ángel et al. 1994a,b) showed that measurement of the colour of the ocean is a powerful tool for oceanographic studies, and that this method can yield information about the ocean surface at meso- to macro-scales. These studies provided justification to launch other sensors such as SeaWiFS (Sea-viewing Wide Field-of-view Sensor), MODIS-Aqua (MODerate resolution Imaging Spectroradiometer) and MERIS (MEdium Resolution Imaging Spectrometer).

Data extracted from ocean-colour images allows one to examine the temporal-spatial variability of the surface layer of the oceans. For example, Chl-*a* is an index of phytoplankton biomass, so a time series of Chl-*a* concentrations can be used in modelling studies that require phytoplankton biomass as an entry variable, such as primary productivity models (Platt et al. 1988; Barocio-León et al., 2007) or carbon

---

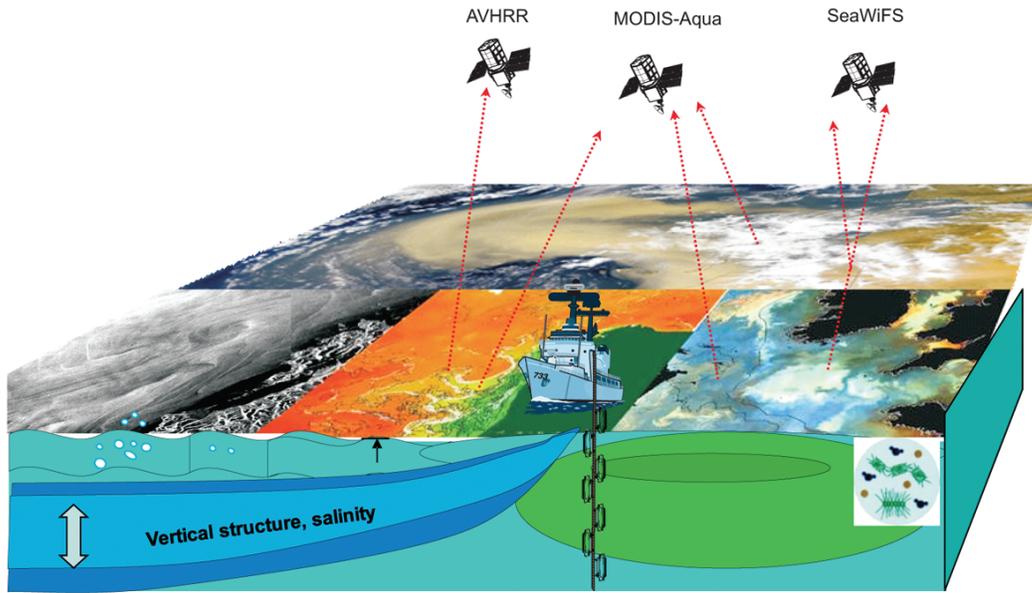
<sup>1</sup>Facultad de Ciencias Marinas, Universidad Autónoma de Baja California (FCM-UABC) Carretera Tijuana-Ensenada Km 106 C.P. 22860 Ensenada BC. Mexico. \* *Email address:* [santamaria@uabc.edu.mx](mailto:santamaria@uabc.edu.mx)

flux models (Camacho-Ibar et al., 2007). In addition, ocean-colour images can provide information about oceanographic surface structures at the meso-scale and allow for tracking of their space-time variations (Traganza et al. 1980; Santamaría-del-Ángel et al., 2002; González-Silvera et al., 2004, 2006; López-Calderón et al., 2008). Such technology may also be able to provide data for fishery studies (IOCCG, 2009; Dulvi et al., 2009).

One of the main challenges in using ocean-colour imagery is to determine the degree of correlation between the *in situ* measurements and the satellite-derived data. NASA uses the 'match-up' technique, which is based on a hypothetical linear relationship between satellite Chl-*a* concentrations ( $Chl_{a_s}$ ) and the *in situ* values obtained from water samples ( $Chl_{a_i}$ ). For most data, a 70% correlation (or 30% error) is considered a good fit (Gregg and Casey, 2004; Djavidnia et al., 2006). To understand the match-up approximation, and to consider the pros and the cons of this method, several statistical considerations must be taken into account: (a) the pattern of data variability, (b) the association indexes used to express the relationship between the *in situ* and satellite data, and (c) the number of data points considered when applying this approximation. It is also important to consider the data scales e.g., *in situ* measurements are generally based on ~1 liter of sea water, while remotely-sensed estimations are obtained from an area of ~1 km<sup>2</sup>. It is difficult to obtain an ideal match-up in space and time. Ideally, *in situ* measurements should be collected at the same time as the radiometric measurements required to validate ocean-colour algorithms.

The spatial distribution of phytoplankton on the ocean surface is not homogeneous; similarly, the vertical distribution in the water column is not homogeneous and generally exhibits a sub-surface maximum (Cullen and Eppley, 1981; Millán-Núñez et al., 1996). The distribution and size of the patches depends on a number of physical (light, turbulent mixing processes such as wind, surges), chemical (nutrients) and biological (algal type) factors. The  $Chl_{a_s}$  data provides information about the phytoplankton biomass in the first optical depth at a scale of ~1 km per pixel (Figure 17.1) while the  $Chl_{a_i}$  data are derived from discrete bottle samples near the ocean surface. Differences in sampling techniques are one of the factors contributing to the variability of the two datasets. Both approximations seek the concentration of Chl-*a*, but while the *in situ* samples are based on spectrophotometric, fluorometric, or HPLC determinations of ~1 liter of water, remote sensor measurements integrate data (through marine optics approximations) from a greater volume, yielding average values of Chl-*a* concentration (Fig. 17.1). Thus the *in situ* and remote sensor measurements evaluate processes on different space/time scales (Fig. 17.2). Satellite remote sensing allows the study of processes >10 km horizontal scale, encompassing several decades, while *in situ* measurements study processes over much smaller time and space scales (cm to meters, minutes to days).

The 'match-up approximation', a graphical technique based on a theoretical straight line fitted to two variables with identical distribution patterns, can be used



**Figure 17.1** Schematic representation of the *in situ* and satellite-based sampling methods.

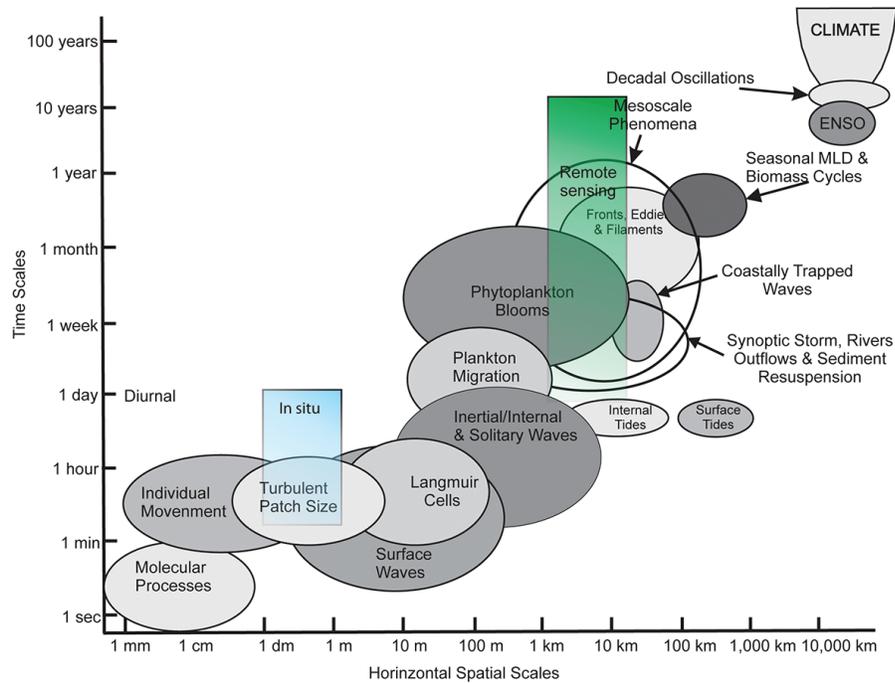
to compare data. If one variable is plotted against itself, or two variables with different magnitudes are plotted, the resulting graph yields a straight line with a 45-degree slope. As the data distribution differs, the dispersion increases. To determine the statistical validity of the observed patterns, a statistical analysis can be applied to examine the level of linear association between variables. The most common linear association index is Pearson's correlation coefficient ( $r_p$ ), often simply just called 'correlation coefficient', denoted  $r$ . The mathematical expression is:

$$r_p = \frac{\text{Cov}_{A,B}}{SD_A \times SD_B} \quad (17.1)$$

where  $r_p$  = Pearson's correlation coefficient;  $\text{Cov}_{A,B}$  = Covariance of A and B;  $SD_A$  = standard deviation of A;  $SD_B$  = standard deviation of B. It is a measure of the correlation (linear dependence) between two variables A and B, giving a value between +1 and -1 inclusive (1 indicates a direct linear relationship, -1 indicates an inverse linear correlation, and zero indicates no linear relationship). It is expressed by the covariance of the two variables divided by the product of their standard deviations. A hypothesis test known as a 'correlation analysis' is carried out to determine if the coefficient is significant:

$$\begin{aligned} H_0 : r_p &= 0 \\ H_a : r_p &\neq 0 \end{aligned}$$

To accept or reject  $H_0$ , two values must be compared; the calculated value ( $r_{p_{cal}}$ ),



**Figure 17.2** Schematic representation of the temporal space scales that cover each type of sampling method. Adapted from Dickey (2003).

derived from Equation 17.1, and the critical value ( $r_{p_{cr}}$ ) obtained from a table of critical values (found in any statistical textbook) based on the degree of freedom ( $df = n-1$ ) and the error  $\alpha$  (1-confidence level). Confidence levels are 90, 95 and 99% yielding errors of 10, 5 and 1, respectively.  $r_{p_{cr}}$  is the minimum significant value of  $r_p$ . If  $r_{p_{cal}} > r_{p_{cr}}$ ,  $H_0$  is rejected and is statistically significant. If  $r_{p_{cal}} < r_{p_{cr}}$ ,  $H_0$  cannot be rejected and  $r_p$  is not significant. Decision making becomes more robust with a greater number of data points, which is why the number of data points is critical. In general, only a small number of data points are obtained if one uses only the samples collected close to the time of the satellite overpass, or on sunny days.

A significant value does not imply a cause-and-effect relationship. For example, a correlation coefficient of 0.975 between Chl-*a* and sea surface temperature (SST) does not imply an increase of Chl-*a* with an increase in temperature, but rather that SST can be used as an indicator of temperature surge e.g. upwelled cool nutrient-rich water can cause an increase in the phytoplankton biomass in the euphotic zone. Furthermore, it should be noted that Chl-*a* is expressed on a logarithmic scale while  $r_p$  is not, so a logarithmic transformation of the Chl-*a* data is required. To perform the match-up in a more direct manner, the use of Spearman's Non-Parametric Correlation Coefficient can be used (Equation 17.2):

$$r_s = \frac{COV_{RA,RB}}{SD_{RA} \times SD_{RB}} \quad (17.2)$$

where:  $r_s$  = Spearman's correlation coefficient;  $COV_{RA,RB}$  = covariance of the ranges of A and B;  $SD_{RA}$  and  $SD_{RB}$  = standard deviation of the ranges of A and B, respectively. The outcome for  $r_s$  is very similar to that of  $r_p$ , with a range of -1 to 1. The statistical significance of  $r_p$  can be determined through hypothesis testing similar to the tests described for  $r_s$  using a table of critical values of Spearman's coefficient. Case studies of two cruises are presented below as a practical demonstration:

1. **Case Study 1:** Only oceanographic stations sampled close to the time of the sensor overpass are considered, using data from the R/V IOFFE 2002 Ushuaia-Montevideo cruise (8-12 March, 2002).
2. **Case Study 2:** A combination of ten cruises in the CalCOFI (California Cooperative Oceanic Fisheries Investigations (<http://www.calcofi.org>) region is used from 2004 to 2006, sampled during daylight hours using both MODIS-Aqua and SeaWiFS data to help increase the number of observations.

SeaWiFS images with 1-km pixel resolution were used to make daily composites for both cruises. The concentration of Chl-*a* was calculated using the OC4-V4 algorithm (O'Reilly et al. 2000, Equations 17.3):

$$Chla = 10^{0.366 - 3.067R + 1.930R^2 + 0.649R^3 - 1.532R^4}, \quad (17.3)$$

where  $R = \log_{10} \left[ \frac{R_{rs443} > R_{rs490} > R_{rs510}}{R_{rs555}} \right]$ . The OC3M-V4 algorithm was used for the MODIS-Aqua images (O'Reilly et al. 2000, Equation 17.4):

$$Chla = 10^{0.283 - 2.753R + 1.457R^2 + 0.659R^3 - 1.403R^4}, \quad (17.4)$$

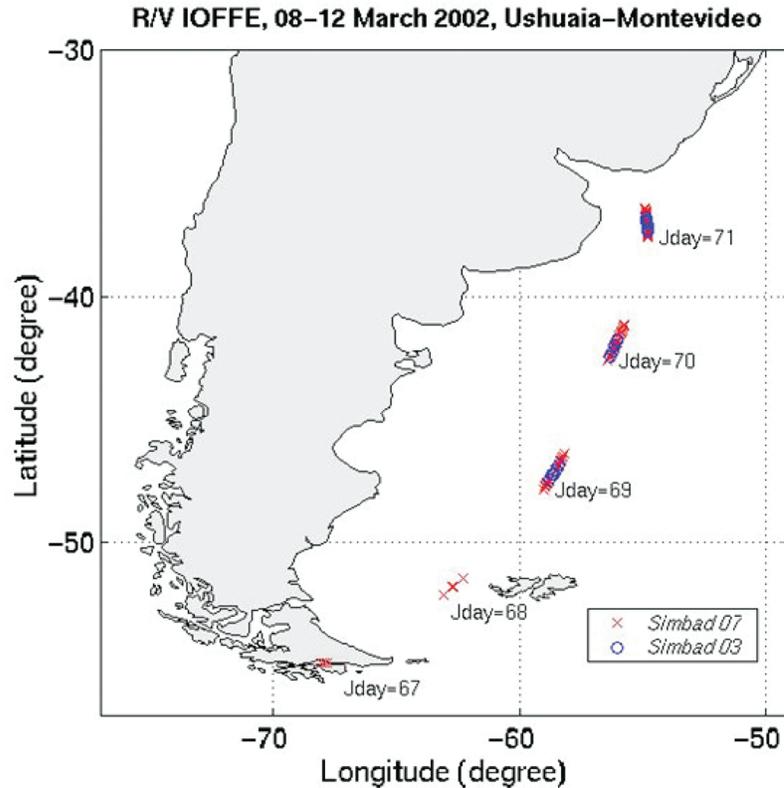
where  $R = \log_{10} \left[ \frac{R_{rs443} > R_{rs488}}{R_{rs551}} \right]$ .

## 17.2 Demonstration

### 17.2.1 Case Study 1

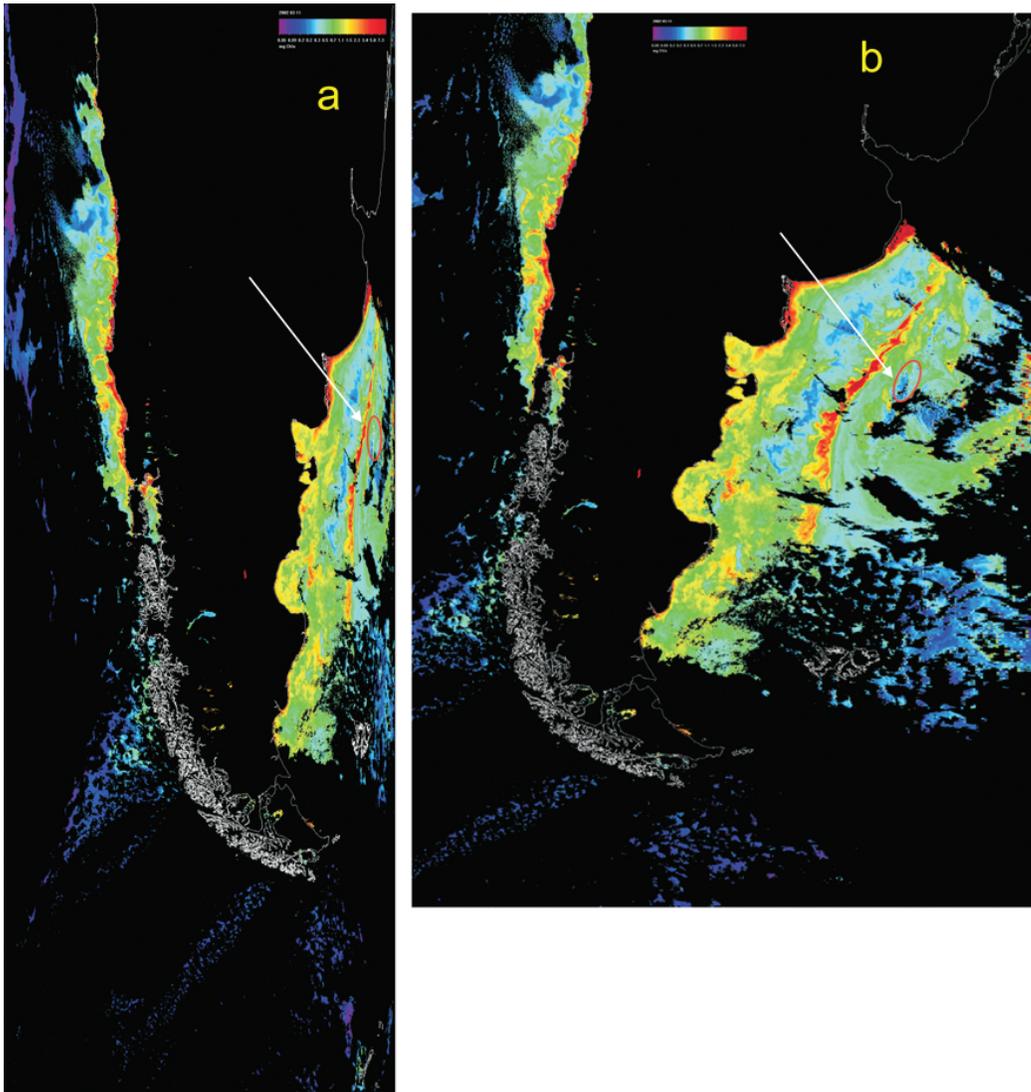
*In situ* data for this case study was collected during the R/V IOFFE Ushuaia-Montevideo cruise (8-12 March 2002) (Fig. 17.3), and was compared to SeaWiFS satellite data. Of the 337 oceanographic stations sampled during the cruise, only 14 fulfilled the requirements for match-up analysis i.e. samples collected between the hours of 10:00 and 14:00 (i.e. 2 hours before or after SeaWiFS overpass). In cases where the study area has high cloud coverage, all available satellite images are needed for analysis. Furthermore, some satellite images may not be centered directly over the sampling area, so some *in situ* sampling stations may not have adequate satellite data because of pixel degradation at the extreme edge of the sensor sweep (see Figure 17.4). In addition, clouds can prevent satellite data collection over a sampling

station. It is thus recommended that a  $3 \times 3$  pixel box centered over the station coordinates be used when extracting satellite data over a sampling station. There are several data extraction software packages available, including MatLab, WIM, ENVI and SEADAS.



**Figure 17.3** Study area of the R/V IOFFE Ushuaia-Montevideo cruise (8–12 March 2002).

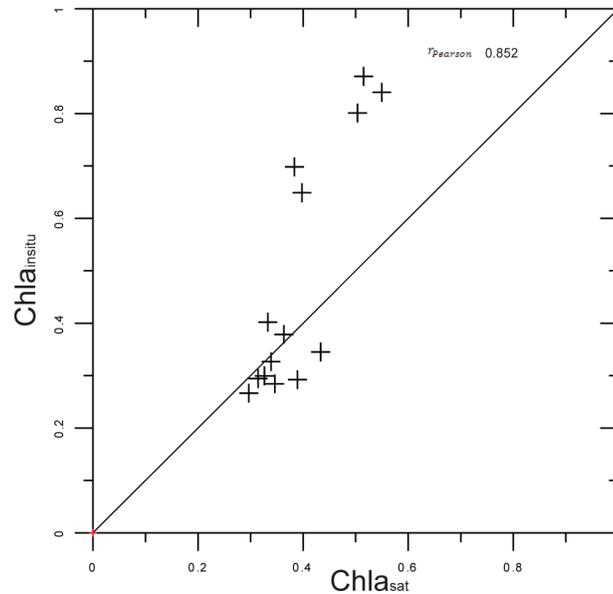
All files used in this case study can be downloaded from the IOCCG website at <http://www.ioccg.org/handbook/matchup/>. The Excel file 'case1data.xls' shows 14 stations with  $Chl a_i$  (determined by HPLC) and the averages of the  $3 \times 3$  box centered over the sampling station coordinates ( $Chl a_s$ ). Since  $Chl a_s$  represents integration over the first optical depth, samples within the first optical depth must be integrated for  $Chl a_i$ . The correlation between  $Chl a_s$  and  $Chl a_i$  is determined using  $r_p$ . Although 14 data points is a relatively small number,  $r_p = 0.852$  indicating that 85.2% of the total variability can be explained by one, or several, linear models. This value is statistically significant ( $\alpha = 5\%$ ,  $r_{p_{cr}} = 0.532$ ). Figure 17.5a shows that both *in situ* and satellite chlorophyll concentrations around  $0.3 \text{ mg m}^{-3}$  are remarkably similar. However, when  $Chl a_i > 0.6 \text{ mg m}^{-3}$ ,  $Chl a_s$  is underestimated.



**Figure 17.4** Example of SeaWiFS chlorophyll image S2002070154515 processed to (a) Level 2, and (b) Level 3.

### 17.2.2 Case Study 2

This example will demonstrate how to increase the number of matchup data points in areas with high cloud coverage, using data from more than one satellite sensor. Data from 10 cruises in the CalCOFI region were used (2004 to 2006) in conjunction with MODIS-Aqua and SeaWiFS satellite imagery (Figure 17.6). Using data from two satellite sensors increases the possibility of matchup data over a given sampling station because of the different overpass times of the sensors, and changes in cloud cover patterns throughout the day. Figure 17.7 shows SeaWiFS and MODIS-Aqua images for 7 and 8 February 2006. A common area is defined by a yellow circle in



**Figure 17.5** Relationship between *in situ* chlorophyll measurements from the R/V IOFFE cruise, and satellite SeaWiFS-derived chlorophyll.

each image to highlight changes in cloud cover patterns. The database for this case study can be found in the Excel file ‘case2data.xls’ on the IOCCG website. There are five columns: station-cruise,  $Chla_s$  for MODIS-Aqua and SeaWiFS, the geometric mean of both, and  $Chla_i$ . The arithmetic mean ( $\bar{X}_a$ ) for each station is the sum of the data values divided by the total number of data points.

$$\bar{X}_a = \frac{\sum x}{n} \quad (17.5)$$

This mean uses all pixels, even those with no geophysical values, so the geometric mean ( $\bar{X}_g$ ) should be used to generate a value that is representative of the data:

$$\bar{X}_g = \frac{\sum x}{N_{in}} \quad (17.6)$$

i.e., the ratio of the sum of the valid data and the number of pixels that yielded these valid data points ( $N_{in}$ ). Using equations 17.1 and 17.2,  $r_P$  and  $r_S$  coefficients were calculated for the data as well as the base 10 log-transformed data (Table 17.1). Using only the MODIS-Aqua satellite data for the 10 cruises over almost 3 years would yield 128 match-up data points. If only SeaWiFS data were used, this number would increase to 142. Combining data from the two sensors and using the geometric mean, the number of data points increases to 172. Note that all the coefficients are statistically significant at  $\alpha = 5\%$ ; but only at concentrations  $< 1 \text{ mg Chla m}^{-3}$ .

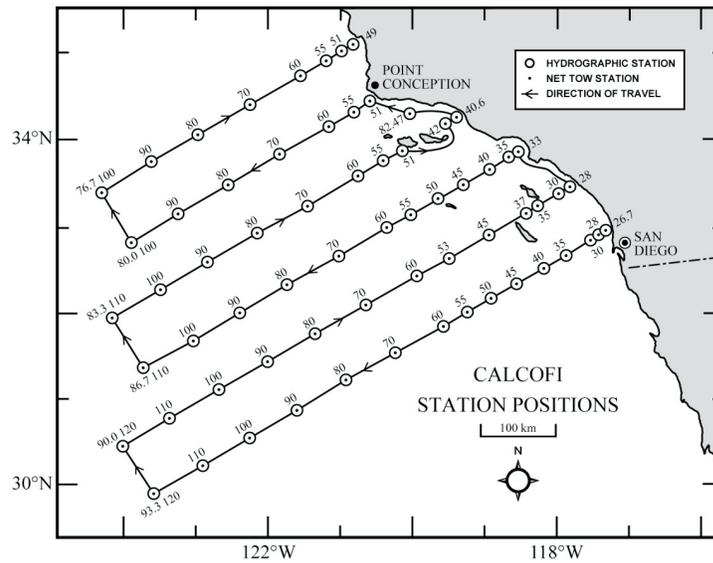
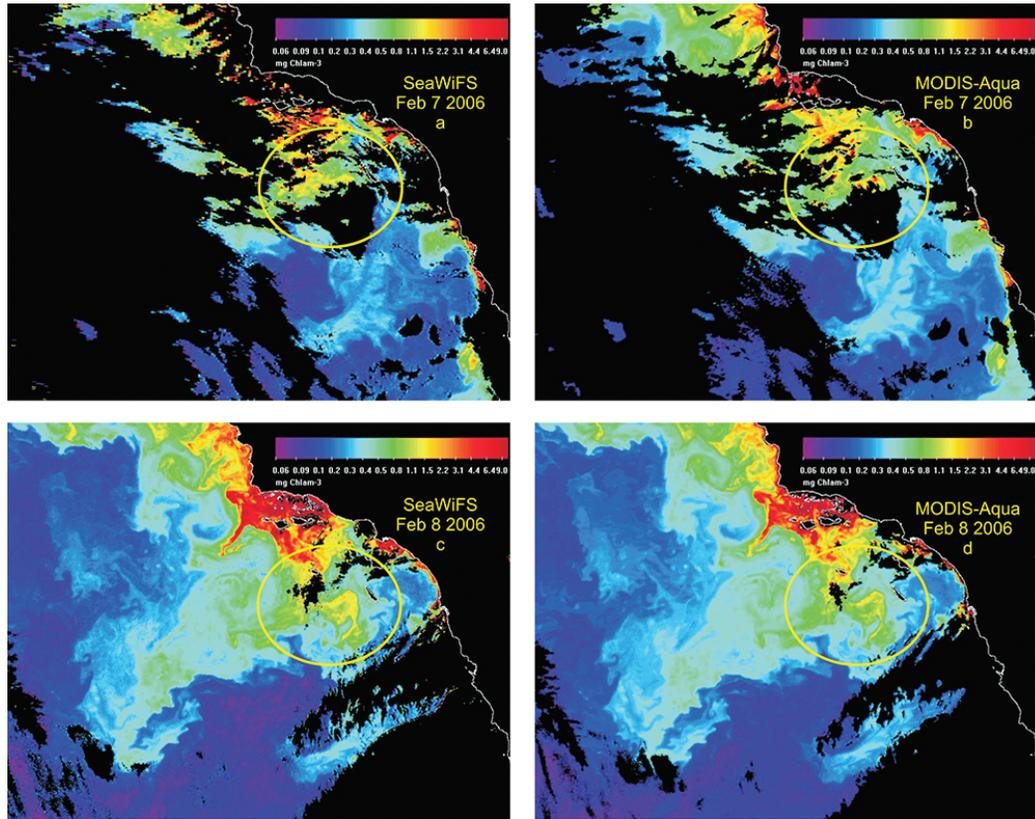


Figure 17.6 Location of the CALCOFI study area.

Table 17.1 Correlation between the *in situ* concentrations of chlorophyll-*a* and the concentrations derived from MODIS-Aqua, SeaWiFS and the combination of both sensors for CalCOFI cruises from 2004 to 2006.

	Data		Log <sub>10</sub> Data	
	$r_p$	$r_s$	$r_p$	$r_s$
MODIS-Aqua (n=128)	0.690	0.839	0.807	0.839
SeaWiFS (n=142)	0.588	0.859	0.802	0.859
Both (n=172)	0.664	0.882	0.834	0.882

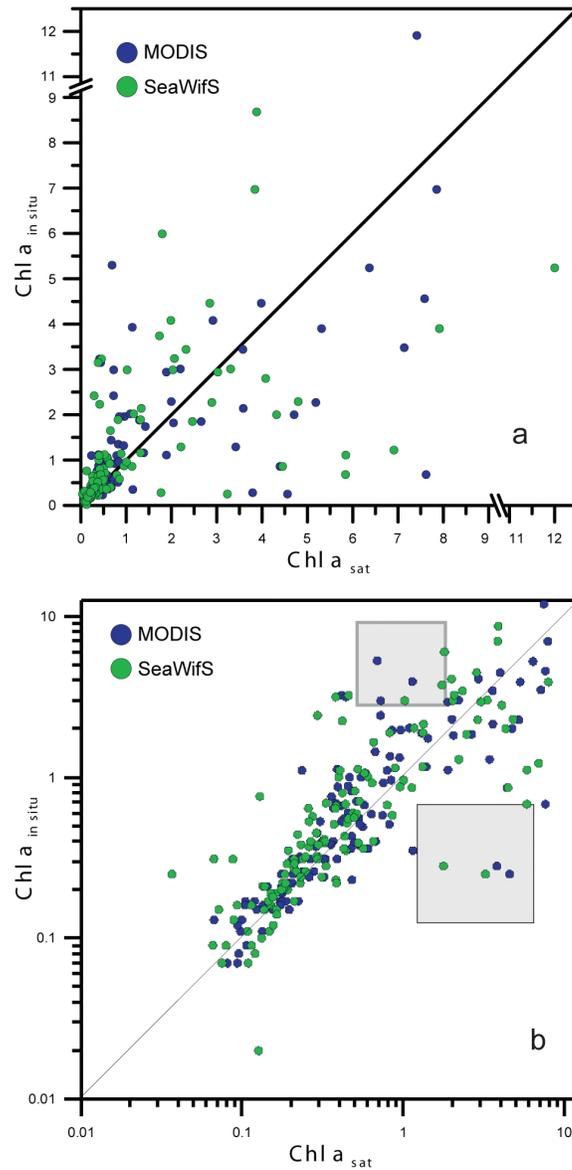
Since  $r_p$  determines the degree of correlation expressed by the variability explained by linear relationship, while  $r_s$  determines the degree of correlation (including that explained by linear models), Spearman values will be greater than Pearson values, which is apparent in the non-transformed data.  $Chl_a_s$  is generally expressed on a logarithmic scale and it can be seen that  $r_p$  increases if the data is log-transformed, while  $r_s$  remains the same, suggesting that Spearman's correlation coefficient is better for establishing the degree of correlation between  $Chl_{a_i}$  and  $Chl_{a_s}$ . Figure 17.8a shows that with non log-transformed data at low chlorophyll concentrations, there appears to be a high correlation between the satellite and  $Chl_{a_i}$  data, but at concentrations  $> 1 \text{ mg Chl a m}^{-3}$ , data dispersion increases considerably. This is less evident when expressing the same relationships on a logarithmic scale (Figure 17.8b). Note that expressing these relationships on a log scale only changes the visual representation, not the distribution pattern of the data points. Sampling can only provide a window of data into the global variability (Figure 17.8b gray



**Figure 17.7** Example images for the CalCOFI case study from 7–8 February 2006. a) SeaWiFS image from 7 February 2006, b) MODIS-Aqua image from 7 February 2006, c) SeaWiFS image from 8 February 2006, and d) MODIS-Aqua image from 8 February 2006. The yellow circle delineates an area to examine variability in cloud coverage.

squares), which is why linear relationships cannot explain all cases. For this reason, Spearman's correlation coefficient is a better indicator than Pearson's because it explores all types of relationships between two variables.

The CalCOFI region represents a system with a high space-time variability that is affected by climate fluctuations (Bograd et al., 2003). The increase in the water temperature and thermocline depth and the stratification of the water column are accompanied by changes in the populations of algae, invertebrates, zooplankton, fish and birds (Bograd et al., 2003). The CalCOFI database and the information derived from satellite imagery offers the potential to construct robust models that can explain the high variability of this area. This region is characterized by a strong oceanographic structure at the mesoscale, with the generation and evolution of meanders, eddies and filaments along the coast. The combination of data from several sources (including satellites and *in situ* measurements) in numerical models can be used to complement the descriptions of this variability at the mesoscale. Di



**Figure 17.8** Relationship between *in situ* chlorophyll data and SeaWiFS (green circles) and MODIS-Aqua (blue circles) chlorophyll data for the CALCOFI cruises from 2004 to 2006, plotted on a linear scale (a), and a logarithmic scale (b). The grey squares represent a hypothetical window when using data from one cruise only.

Lorenzo et al. (2004) used the CalCOFI database in combination with data derived from SeaWiFS to model the dynamic nature of the California Current system. They noted that the comparison of *in situ* Chl-a data with that derived from SeaWiFS was difficult due to the different sampling scales employed by each approximation. This case study proposes a better approximation to compare the *in situ* and satellite data,

allowing for the space-time resolution of both to be maximized.

### 17.3 Training

The files in the folder entitled "trainingfiles" (<http://www.ioccg.org/handbook/matchup/>) will be used in this section. First we will focus on data extraction from stations in the  $3 \times 3$  pixel boxes. The stations and images of the CalCOFI 0507 cruise (July 2005) will be used. Before starting, three points must be considered:

1. It is important for all images to have the same geographic projection to facilitate preparation of the script for data extraction (based on latitude, longitude and geophysical value data matrices). If the images do not have the same projection, the matrices will have different dimensions.
2. The text file '0507stations.txt' lists the details of the sampling stations in three columns: longitude, latitude (both in degrees and tenths of a degree) and station identification. The first row is used for column headers. Note that latitudes are positive in the northern hemisphere and negative in the southern hemisphere, and longitudes are positive in the eastern hemisphere and negative in the western hemisphere.
3. A text file must be generated with a list of addresses where the images are stored (see '0507imagery.txt').

We used the WIM (Windows Image Manager) software (<http://www.wimsoft.com>), specifically its WAM (WIM Automation Module) module called 'wam\_statist'. In the upper left hand corner, there is a window labeled "List of Images", where the file name '0507imagery.txt' is placed. In the upper right hand corner, there are two windows: the top one is labeled "Mask or Station File Name" where the name and address of the station file ('0507stations.txt') is placed. The name and address of the file where the data are stored ('0507wam\_statist\_result.csv') is placed in the bottom window. This type of file can be opened in Excel and yields 23 columns. Column A consists of the image names, columns B and C are the start and end years (if the images are composites). Columns D and E are the start and end days (if the images are weekly or monthly composites). This case study uses daily LAC images with B and C values of 2005 and D and E having the same value (until another image is analyzed). Column F identifies the station that is named in the third column of the station file ('0507stations.txt'). Column G indicates the number of pixels in the  $3 \times 3$  box that have data ( $G, N_{in}$ ) and column H indicates the number of pixels that do not have data ( $H, N_{out}$ ). The maximum value in each column is 9 and the minimum is 0, so that if column G has a value of 9, all the pixels in the  $3 \times 3$  box have data.

The basic geometric statistical parameters can be extracted from the data in column G, (based only on the valid pixels): minimum (I), maximum (J), mean (K), standard deviation (L), and median (M). When there are no data in the  $3 \times 3$  box due to cloud coverage, signal saturation, or other factors (i.e. column H has a value

of 9), the value in these columns is -99. Column N denotes the pixel centered in the geographical coordinates where the station was located. Columns P through W contain the values of the remaining pixels in the 3×3 box. These data allow comparisons to be made with other statistical parameters, e.g. the mode.

The next step is to generate a file where the extracted data can be combined with the *in situ* data. In this case, this file was generated by combining the data from CalCOFI cruise 0507 and the results of the extraction file '0507wam\_statist\_result.csv'. The resulting file ('0507match-up.xls') will be used in the second part of this section, where the focus will be on the calculation of Pearson's and Spearman's correlation coefficients to establish the degree of correlation between  $Chl_s$  and  $Chl_i$ . We will use data from MODIS-Aqua and SeaWiFS to increase the number of data points for the calculation of the two coefficients (using Equations 17.1 and 17.2), for both the log-transformed and raw data (Table 17.2).

**Table 17.2** Correlation between *in situ* chlorophyll-*a* and chlorophyll derived from MODIS-Aqua, SeaWiFS, and the combination of the two sensors, for the CalCOFI 0507 cruise.

	Data		Log <sub>10</sub> Data	
	$r_p$	$r_s$	$r_p$	$r_s$
MODIS-Aqua (n=5)	0.822	1.000	0.947	1.000
SeaWiFS (n=4)	0.933	1.000	0.881	1.000
Both (n=4)	0.873	0.964	0.946	0.964

Next, the statistical range must be calculated. This is done by labeling the smallest number in the data series 1, the next smallest 2 and so on, until the whole data series has been labelled. Table 17.3 shows three sets of data. Set A has no repeating values, so the range is calculated starting at 1 and ending in 10, since  $n = 10$ . Set B has repeating values (number 90 is repeated twice). The corresponding ranges would be 1 and 2, so a mean of the ranges is calculated and each would be assigned a value of 1.5 ( $\frac{1+2}{2}$ ). The next range to assign would be 3. Set C has a triple repeat of 124 and a double repeat of 128. In this case, 124 would have the corresponding ranges 4, 5, and 6, so a mean range of 5 is assigned to each, leaving the next range value as 7. For the 128 repeat, the corresponding ranges are 7 and 8 so a range value of 7.5 is assigned to each, leaving the next range value as 9.

Even with only a few data points (Table 17.2)  $Chl_s$  displays a high correlation with  $Chl_i$ . As noted previously,  $r_p$  coefficients are generally lower than  $r_s$  (= 100 in this study). Even when the  $r_p$  values are large, this does not imply a 1:1 relationship (Figure 17.9a). Rather, it implies that a high percentage of the variability can be explained by linear models. If the same graph is expressed on a log scale (Figure 17.9b), an apparent 1:1 linear relationship is observed. Note, the relationship of  $Chl_s$  to  $Chl_i$  is not the same as  $\log_{10}Chl_s$  to  $\log_{10}Chl_i$ .

**Table 17.3** Data demonstrating the calculation of ranges for one variable.

Set A	Rank A	Set B	Rank B	Set C	Rank C
133	6	129	6.0	128	7.5
137	8	132	8.0	124	5.0
99	3	90	1.5	110	3.0
138	9	136	9.0	131	9.0
92	2	90	1.5	98	2.0
89	1	93	3.0	84	1.0
130	4	114	4.0	147	10.0
132	5	129	6.0	124	5.0
141	10	150	10.0	128	7.5
135	7	129	6.0	124	5.0

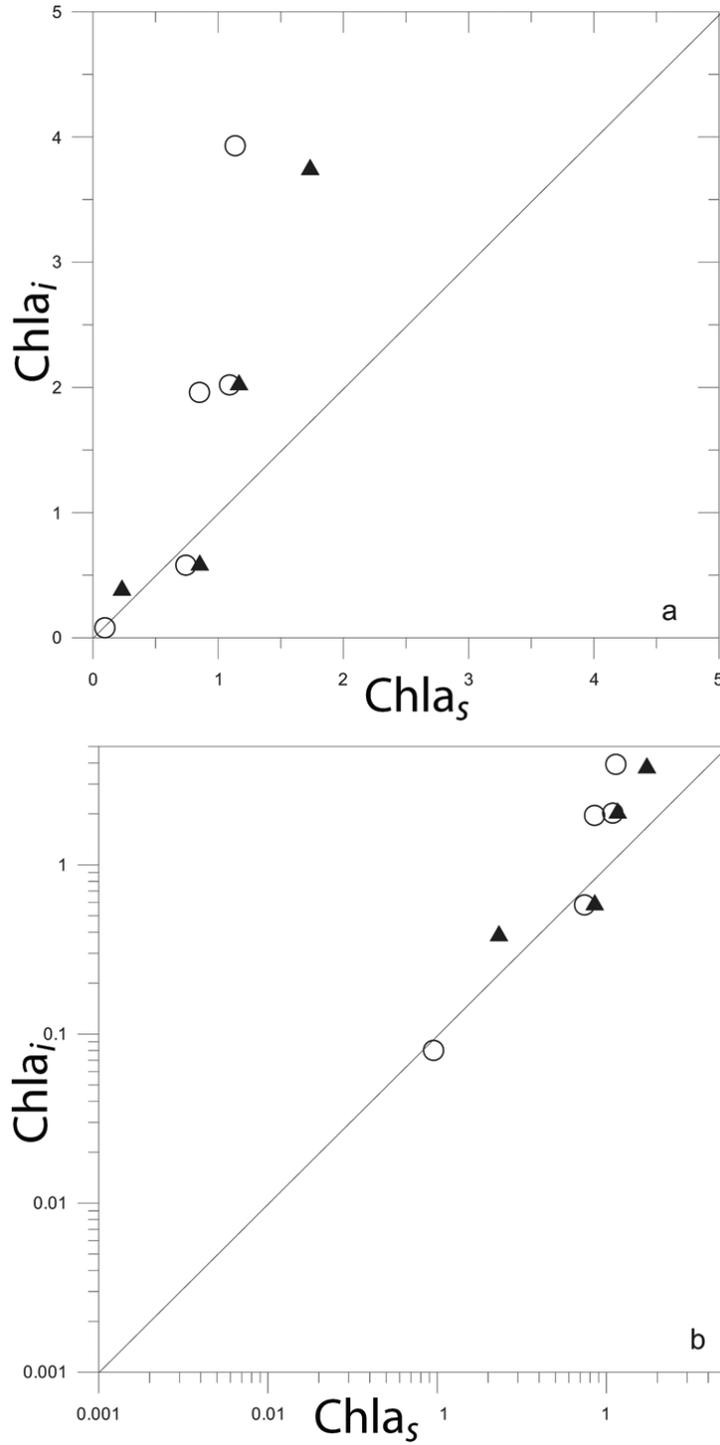
If extrapolation of data (modelling one concentration based on the other) is desired in addition to generation of the linear model, tests on the significance of the intercept, the slope and the global significance of the model must be carried out. However, none of this is needed if only the degree of match-up is desired.  $r_p$  yields the degree of variability that can be explained by linear models. If one concentration is to be modeled based upon the other, the analysis can be based on empirical (linear) or mechanistic models. All the models have a determination coefficient ( $R^2$ ); i.e. the percentage of variability explained by a specific model. It is calculated as follows:

$$R^2 = \frac{SSM}{SST_o} \quad (17.7)$$

$$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (17.8)$$

$$SST_o = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (17.9)$$

where  $SSM$  is the sum of the squared differences between the modelled data ( $\hat{y}_i$ ) and the mean of the observed data ( $\bar{y}$ ).  $SST_o$  is the sum of the squares between the observed data ( $y_i$ ) and the mean of the observed data ( $\bar{y}$ ) and defines of the total variability of the dependent variable.  $R^2$  is the ratio of the two sums of squares. When a linear model is used, it is assumed that  $R^2 = r_p^2$ , but this may not hold true for all data.



**Figure 17.9** Correlation between *in situ* chlorophyll and satellite-derived chlorophyll (triangles = SeaWiFS; circles = MODIS-Aqua) for the CALCOFI 0507 cruise, plotted on a linear (a) and (b) logarithmic scale.

## 17.4 Questions

1. Why is it important to have an *in situ* database for a defined grid, sampled over a long period of time?
2. Why is the CalCOFI area so important in this regard?
3. What is the weakness of the CalCOFI database, and how can that weakness be minimized?
4. What are the advantages and disadvantages of using satellite-derived data in this area?
5. Why is the relationship between *in situ* measurements and those derived from remote sensors important?
6. Are "normalized" data required to carry out match-up approximations? Do they have to be normalized with logarithms?
7. How should match-up study results be expressed?
8. What is the difference between  $R^2$  and  $r_P$ ?
9. Is Spearman's coefficient better than Pearson's?
10. How do you calculate the range for  $r_S$ ?
11. What is the geometric mean?
12. Is the geometric mean representative of the  $3 \times 3$  data extraction box?
13. Why is it important to have all the images at the same projection?

## 17.5 Answers

1. This sampling scheme allows variations over seasons, years, decades and longer time scales to be assessed in a more reliable manner and also allows the system to be modelled.
2. The CalCOFI area has a sampling record, for a defined grid, going back more than 60 years.
3. A possible weakness of the CalCOFI database is that it only provides data four times a year, leaving nearly nine months with no monitoring in the area. The *in situ* observations of CalCOFI can be complemented with the use of ocean colour and SST images. Although these images only provide information about the surface of the ocean, they can provide a synopsis of changes in space and time.

4. Advantages include access to data on a daily time scale over a broad area, which allow the synoptic description of space-time variability and highlight oceanographic structures at the mesoscale. Currently, a long time series with a 1-km pixel size can be generated. Disadvantages of these data are that they only yield surface information and require cloudless days. Weekly or monthly data composites and long time series can be derived from remotely-sensed data.
5. The relationship between *in situ* measurements and those derived from remote sensors has three components: a) synoptic complementary data for space-time studies in windows where the *in situ* sampling does not yield any data; b) representation of indirect approximations, as well as those from remote sensors; and c) entry variables used to model the system.
6. No, if "normalized" means that the data fit a Gaussian distribution. Pearson's and Spearman's correlation coefficients do not require that the internal distribution of the variables fits a Gaussian curve.
7. The calculated value of the chosen coefficient must be presented as well as the significance given by the hypothesis tests, indicating the number of data points and the error,  $\alpha$ . A graph can be constructed with axes that have the same scale. A 45° straight line denoting the 1: 1 line should be included.
8.  $r_p$  is the degree of variability that can be explained by linear models (one or several), while  $R^2$  represents the variability explained by a given model. When a linear model is used, it is assumed that  $R^2 = r_p^2$ . However, this assumption may not be true in all cases. Furthermore,  $r_p$  calculated for variables AB is the same as that calculated for BA, but  $R^2$  is exclusive of a particular model.
9. Pearson's coefficient measures the degree of linear association, while Spearman's simply measures the degree of association. Spearman's coefficient is more robust if all that is sought is the degree of association.
10. Statistical ranges are defined as hierarchical indicators of a data set. A value of 1 is assigned to the smallest number in the series, the next smallest number is labeled with 2 and so on. The maximum range is equal to the number of data points. In the case of data points with the same values, the mean of the ranges assigned to the repeated number is calculated and assigned.
11. It is the sum of the valid data points divided by the number of pixels contributing to the valid data.
12. In general, the mean is considered representative of the data set but for satellite data sets, the geometric mean is more representative than the arithmetic mean.

13. It is important that all the images have the same geographic projection because it facilitates writing a data extraction program based on data matrices of latitude, longitude and geophysical values. If the images did not have the same projection, the matrices would have different dimensions, which would require another entry variable in the data extraction process.

## 17.6 References

- Barocio-León O, Millán-Núñez R, Santamaría-del-Ángel E, González-Silvera A (2007) Phytoplankton primary production in the euphotic zone of the California Current System from CZCS imagery and modeling. *Ciencias Marinas* 33(1): 59-72
- Bograd SJ, Checkley DA, Wooster WS (2003) CalCOFI: a half century of physical, chemical, and biological research in the California Current System. *Deep Sea Res II* 50 (14-16): 2349-2353
- Camacho-Ibar VF, Hernández-Ayón M, Santamaría-del-Ángel E, Dásele-Heuser LW, Zertuche-González JA (2007) Correlation of surges with carbon stocks in San Quintín Bay, a coastal lagoon in northwest Mexico. In: Hernández de la Torre B, Gaxiola Castro G (eds). *Carbon in Aquatic Ecosystems of Mexico*. Co-edited by the Department of Environmental and Natural Resources, National Ecology Institute NEI and the Center for Scientific Research and Higher Education of Ensenada CICESE, Mexico City pp 355-370
- Cox PM, Betts RA, Jones CD, Spall SA, Totterdell IJ (2000) Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature* 408, 184-187
- Cullen, JJ, Eppley RW (1981) Chlorophyll maximum layers of the Southern California Bight and possible mechanisms of their formation and maintenance. *Oceanol Acta* 4: 23-32
- Dickey, T (2003) Emerging ocean observations for interdisciplinary data assimilation systems. *J Marine Sys* 40-41: 5-48
- Djavidnia S, Mélin F, Hoepffner N (2006) Analysis of Multi-Sensor Global and Regional Ocean Colour Products. MERSEA - IP Marine Environment and Security for the European Area - Integrated Project Report on Deliverable D.2.3.5 European Commission - Joint Research Centre Ref: MERSEA-WP02-JRC-STR-0001-01A.pdf. 228 pp
- Dulvy N, Chassot E, Heymans J, Hyde K, Pauly D, Platt T, Sherman K (2009) Climate change, ecosystem variability and fisheries productivity. In: Forget M-H, Stuart V, Platt, T (eds), *Remote Sensing in Fisheries and Aquaculture*. IOCCG Report No. 8, IOCCG, Dartmouth, Canada pp 11-28
- Di Lorenzo E, Miller AJ, Neilson DJ, Cornuelle BD, Moisan JR (2004) Modelling observed California Current mesoscale eddies and the ecosystem response. *Int J Rem Sens* 25(7-8): 1307 - 1312
- Eber LE, Hewitt RP (1979) Conversion Algorithms for CALCOFI Station Grid. *CalCOFI Rep* XX: 135-137
- Friedlingstein, P, Cox PM, Betts RA, Bopp L, Von Bloh W, Brovkin V et al (2006) Climate-carbon cycle feedback analysis: results from the C4MIP model intercomparison. *J Climate* 19(14): 3337-3353
- Gregg WW, Casey NW (2004) Global and regional evaluation of the SeaWiFS chlorophyll data set. *Rem Sens Environ* 93: 463-479
- González-Silvera AG, Santamaría-del-Ángel E, Millán-Núñez R, Manzo-Monrroy H (2004) Satellite observations of mesoscale eddies in the Gulfs of Tehuantepec and Papagayo (Eastern Tropical Pacific). *Deep Sea Res II* 51: 587-600
- González-Silvera AG, Santamaría-del-Ángel E, Millán-Núñez R (2006) Spatial and temporal variability of the Brazil-Malvinas Confluence and the La Plata Plume as seen by SeaWiFS and AVHRR imagery. *J Geophys Res* 111: C06010, doi: 10.1029/2004JC002745
- Haidvogel DB, Wilkin JL, Young RE (1991) A semi-spectral primitive equation ocean circulation model using vertical sigma and orthogonal curvilinear horizontal coordinates. *J Comput Phys* 94: 151-185
- IOCCG (2000) *Remote Sensing of Ocean Colour in Coastal, and Other Optically-Complex, Waters*. Sathyendranath, S (ed), Reports of the International Ocean-Colour Coordinating Group, No. 3, IOCCG, Dartmouth, Canada

- IOCCG (2004) Guide to the Creation and Use of Ocean-Colour, Level-3, Binned Data Products. Antoine, D (ed), Reports of the International Ocean-Colour Coordinating Group, No. 4, IOCCG, Dartmouth, Canada
- IOCCG (2006) Remote Sensing of Inherent Optical Properties: Fundamentals, Tests of Algorithms, and Applications. Lee, Z-P (ed), Reports of the International Ocean-Colour Coordinating Group, No. 5, IOCCG, Dartmouth, Canada
- IOCCG (2007) Ocean-Colour Data Merging. Gregg W (ed), Reports of the International Ocean-Colour Coordinating Group, No. 6, IOCCG, Dartmouth, Canada
- IOCCG (2009) Remote Sensing in Fisheries and Aquaculture. Forget M-H, Stuart V, Platt T (eds), Reports of the International Ocean-Colour Coordinating Group, No. 8, IOCCG, Dartmouth, Canada
- Lewis MR (1992) Satellite ocean colour observations of global biogeochemical cycles. In: Falkowski PG, Woodhead AD (eds). Primary Productivity and Biogeochemical Cycles in the Sea. Plenum Press NY pp 139-154
- Lopez-Calderon J, Martinez A, Gonzalez-Silvera A, Santamaria-del-Angel E, Millan-Nuñez R (2008) Mesoscale eddies and wind variability in the northern Gulf of California. *J Geophys Res* 113: C10001, doi:10.1029/2007JC004630
- Millán-Núñez R, Alvarez-Borrego S, Trees CC (1996) Relationship between deep chlorophyll and surface chlorophyll concentration in the California Current System. *CalCOFI Rep* 37: 241-250
- Müller-Karger FE, Walsh JJ, Evans RH, Meyers MB (1991) On the seasonal phytoplankton concentration and sea surface temperature cycles of the Gulf of Mexico as determined by satellites, *J Geophys Res* 96(C7): 12,645-12,665
- O'Reilly JE, and 24 other authors (2000) Ocean colour chlorophyll-*a* algorithms for SeaWiFS, OC2, and OC4: version 4 In: Hooker SB, Firestone ER (eds) SeaWiFS Postlaunch Tech Rep Ser, Vol 11. SeaWiFS postlaunch calibration and validation analyses, Part 3. NASA, Goddard Space Flight Center, Greenbelt, MD, pp 9-23
- Pelaez J, McGowan JA (1986) Phytoplankton pigment patterns in the California Current as determined by satellite. *Limnol Oceanogr* 31(5): 927-950
- Platt T, Sathyendranath S, Caverhill CM, Lewis MR (1988) Oceanic primary production and available light: further algorithms for remote sensing. *Deep-Sea Res* 35: 855-879
- Santamaria-del-Angel E, Alvarez-Borrego S, Müller-Karger FE (1994a) Gulf of California biogeographic regions based on coastal zone color scanner imagery. *J Geophys Res* 99(C4): 7411-7421
- Santamaria-del-Angel E, Alvarez-Borrego S, Müller-Karger FE (1994b) The 1982-1984 El Niño in the Gulf of California as seen in coastal zone color scanner imagery. *J Geophys Res* 99(C4): 7423-7431
- Santamaria-del-Angel E, Millán-Núñez R, González-Silvera AG, Müller-Karger FE (2002) The color signature of the Ensenada Front and its seasonal and interannual variability. *CalCoFi Rep* 43: 156-161
- Strub PT, Kosro PM, Huyer A, CTZ Collaborators (1991) The nature of the cold filaments in the California Current System. *J Geophys Res* 96(C8): 14,743-14,768
- Traganza ED, Nestor DA, McDonald AK (1980) Satellite observations of nutrients upwelling off the coastal of California. *J Geophys Res* 85: 4101-4106
- Yoder JA, McClain CR, Blanton JO, Oey LY (1987) Spatial Scales in CZCS-Chlorophyll Imagery of the Southeastern U. S. Continental Shelf *Limnol Oceanogr* 32(4): 929-941

### 17.6.1 Further reading

- Rebstock GA (2003) Long-term change and stability in the California Current System: lessons from CalCOFI and other long-term data sets *Deep Sea Res II* 50(14-16): 2583-2594