African Journal of Environmental Science and Technology Vol. 5(9), pp. 732-747, September 2011
Available online at http://www.academicjournals.org/AJEST
ISSN 1996-0786 ©2011 Academic Journals

Full Length Research Paper

Improved ocean chlorophyll estimate from remote sensed data: The modified blending technique

Mathias A. ONABID

Department of Mathematics and Computer Sciences, Faculty of Sciences, P.O. Box 67 Dschang, University of Dschang, Cameroon. E-mail: mathakong@yahoo.fr.

Accepted 2 August, 2011

Gregg and Conkright (2001) who pioneered the use of the blending technique in an attempt to calibrate ocean chlorophyll, expressed the need for further work to be done in order to obtain improved results. One problem faced when using this technique with spatially sparse data, is distortion of the resulting blended field when approaching the coastal boundaries. In this paper, the causes of the distortion and alternative methods for solving it are discussed. One of these method herein termed the *corrector factor* method, appeared the most appropriate in correcting the problem. In it, the blending process is done twice. This method sees the reduction of the mean squared difference between the blended and satellite fields from 6.299 in the normal blending to 0.347 in the corrector factor blending. This figure is also below the tolerance margin (the mean squared difference between the satellite and *in situ* fields) for the real data which was 0.989. Furthermore, this method is backed by a standard statistical procedure which produces identical results to its own even though the two methods differ in structure. A mathematical proof as to why these results coincide is also outlined. Validation study carried out by the authors showed that at least 80% of the times these methods are used the corrector factor will provide a better estimate of chlorophyll concentration than the original blending method. It is expected that analysis on primary productivity and management in the ocean environment will be greatly enhanced by this new finding.

Key words: Satellite, *in-situ*, sea-WiFS, blending, corrector factor, pseudozeroes, noisy data, kernel smoothing.

INTRODUCTION

Aquatic life and production revolve about the distribution and biomass of phytoplankton in the upper layer of the ocean. These unicellular algae are of extreme importance to the ocean food web. There is therefore, a need to track their existence and population in this environment. Clarke et al. (2006) have shown that to measure the population of unicellular algae by cell count is very difficult, because of their resemblance to other non-alga carbon rich particles. Yet this can be done in terms of photosynthetic pigment content which is endemic across all taxonomic groups of algae. Therefore to better predict the abundance of this phytoplankton, it is important that the distribution of chlorophyll be determined as accurately as possible.

Modelling the distribution of ocean chlorophyll has been greatly hindered by the expensive nature of sea operations and the large areal coverage of the ocean which makes it difficult to carry out direct water sampling. The introduction of the orbiting satellite-borne sensors could provide samples at scale and resolution relevant to the problem since data can be obtained from the spectral properties of light reflected from the sea surface. This too, is not void of errors which could arise from the algorithm for converting reflected data to ocean chlorophyll, knowledge of atmospheric optical state and the chemical composition of the ocean. The coastal zone colour scanner (1978 to 1986) (CZCS) was the first satellite borne ocean reflectance sensor. More recently, the sea-viewing wide field-of-view sensor (sea-WiFS) was launched by NASA in 1997. These problems faced by the sensing equipment creates discrepancies between data obtained from the satellite and those obtained by ship and buoys (*in-situ*), making it more difficult to estimate chlorophyll concentration from any of them.

The idea of blending the two data fields of ocean chlorophyll was introduced by Gregg and Conkright (2001). It is believed that the blending of both data sources can maximize the strength of each data field and

produce a high quality, spatially large data field of ocean chlorophyll. The resulting blended field can then be used to predict chlorophyll concentration where bottle samples can not be obtained. Despite the constraints imposed by Gregg and Conkright (2001) during the blending process, they still expressed the need for more work to be done in order to obtain improved results.

Research objective

One of the problems faced by the blending method when applied to ocean chlorophyll calibration is distortion as one approaches the coastline. Therefore, the main objective of this research is to identify the cause(s) of this distortion and to provide possible corrections to such.

Research outline

To achieve this aim, a review of the literature on chlorophyll calibration using the blending method of Gregg and Conkright (2001) was carried out. Simulation studies based on data generated from a bi-variate Gaussian model was carried out to investigate the process. Possible corrective procedures to the problem identified in this example are presented. Real data from the North Atlantic Ocean obtained from April to June (second quarter) were used to justify the findings from the simulation study. These are the concerns of "The blending method of Gregg and Conkright, Data structure and pre-processing and Simulation studies." The second quarter is preferred because the highest sampling rates in both fields were obtained here. Thus the reality of the relationship between the two data fields could be well established. After application of these techniques to the real data was considered, the proof as to why results from the corrector factor method should coincide with results from a non parametric statistical technique is then outlined. This is followed by a validation study on the *corrector factor* and original blending methods, and finally conclusion.

Results obtained here are based on real data from the North Atlantic Ocean obtained from National Oceanic and Atmospheric Administration (NOAA) and the National Oceanographic Data Center (NODC) in the case of the *in situ* date field and from the output of the 2002 National Aeronautics and Space Administration (NASA) pre-processing of the Sea-WIFS data archive for the case of the satellite data field. The study gives a detailed description of the extraction and pre-processing of the data fields.

The program codes used in this research were written in C+, Press et al. (1992) and FORTRAN, Edgar (1992); and then interfaced in the R version (2.0.1) programming environment, Venables et al. (2003); where the data analysis was done. All the image plots were done using the graphic package of Paciorek (2006).

THE BLENDING METHOD OF GREGG AND CONKRIGHT

Gregg and Conkright (2001) are so far the only ones known to have used the blended analysis method in an attempt to model chlorophyll concentration in ocean water. However, this was restricted to the coastal zone colour scanner (CZCS) Era (1978 to 1986).

The *in-situ* data were obtained from the archive maintained by NOAA, NODC. Seasonal climatologies were constructed using the northern hemisphere conventions: Winter (January to March), Spring (April to June), Summer (July to September), and Autumn (October to December). To obtain high-quality data in the blended field, the method required that *in-situ* data are subjected to rigorous quality control procedures. The *in situ* field therefore contained data collected by ship and buoy, and was assumed to be of high quality and accurate but fell short of full spatial coverage.

For the satellite observations, monthly mean CZCS pigment data were obtained from the NASA goddard space flight centre distributed active archive centre during the lifetime of the CZCS. These pigment estimates were then converted to chlorophyll using the O'Reilly et al. (1998) formula:

$$log_{10}S = (log_{10}P - 0.127/0{:}983)$$

where S indicates the satellite-derived chlorophyll and P indicates satellite-derived pigment.

They constructed seasonal observations by first combining chlorophyll estimates for the individual months into seasons for each year in which the CZCS was operating and then averaging the seasons over the years. This enabled the removal of sampling alias occurring in the CZCS seasonal composites due to unequal sampling of months within seasons. The s*atellite* data field therefore contained observations obtained by satellite-borne sensor. This approach, can provide sampling at a much higher scale and resolution relevant to the problem but the results are subject to several sources of error that affect their accuracy and reliability.

Blending of the data fields

In order to obtain a blended data field, the following were the points under consideration:

(1) After undergoing a rigorous quality control process, the *in situ* data, are assumed to be accurate and thus inserted directly in the final blended field where they serve as boundary values during the blending process. Therefore the *in situ* values remain unadjusted in the final blended field.
(2) Satellite chlorophyll measurements can enter the final blended data field only through the use of the Poisson equation:

$$\nabla^2 U = \rho \quad (1)$$

where U is the final blended field and ρ the forcing term which is defined as the Laplacian of the gridded satellite chlorophyll data ($\nabla^2 S$) with S representing the satellite field.

During the blending process, the second order partial differential equation is used. The second order partial derivative of the satellite field is used to obtain the forcing term. The available *in situ* observations act as boundary values wherever they are found within the working area. The partial differential equation in the case of two spatial variables say, x and y is of the form:

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = \rho(x, y) \quad (2)$$

In the final analysis, the conditional relaxation analysis method (CRAM) described by Oort (1983) was used to solve the series of simultaneous equations that resulted from finite-differencing the Poisson equation. This solution is the blended field.

On the application of this technique to ocean chlorophyll, Gregg and Conkright (2001) expressed the need for more work to be done in order to take into account the wide range of chlorophyll values found in the ocean, and the extreme sparseness of *in situ* data, this will obviously improve the results. They stressed that these problems are not encountered with sea surface temperature (SST) because of the reduced range of variability of ocean temperature. In another attempt, Gregg and Conkright (2002) reanalysed the CZCS, the global ocean chlorophyll archive using compatible atmospheric correction and bio-optical algorithms with Sea-WIFS. This permitted them to be able to quantitatively compare the decadal trends in global ocean chlorophyll from the CZCS period to the sea-WIFS period (September 1997 to December 2000). They concluded that, blending both data archives with available *in situ* data improved the residual errors of each data record. Yet in another attempt, Gregg et al. (2002) reanalysed the CZCS data of the NOAA and the NASA data in an attempt to provide a high-quality blended satellite *in situ* data set with a consistent view of global ocean chlorophyll spanning two decades from 1978.

DATA STRUCTURE AND PRE-PROCESSING

The satellite data herein used is an extract of the output from the 2002 National Aeronautics and Space Administration (NASA) pre-processing of the Sea-WIFS data archive. It comprises observations from 1997 to 2002 averaged over a grid size of 0.75 longitude by 0.75 latitude and using the successive 8-day intervals over the year. After the extraction, there were still some grids with zero values as observations in the satellite field. To reduce the wide range between observation, transformation by natural logarithm was perform and during this process, grids with zero as observation were allocated a pseudozero.

The *in situ* data from the world ocean data base were extracted from the world ocean data base with additional Canadian data (DFO MEDS). A total of 378570 observations were recorded from 1933 to 2002 between latitude 30 and 80°N and between longitude 90°W and 90°E of the equator, representing the research area. Since there has been a lot of change in the environmental conditions over the past 40 years, the data field was restricted to observations collected from 1990 onwards. These were then averaged over the top 5 m of the ocean at grid nodes determined by the intersection of latitude and longitude. Observations with value more than 40 mg/m^3 were considered to be outliers and thus eliminated from the sample, as were observations identified as being on land. The resulting data field was unevenly distributed as most of the observations clustered around the coastal waters, with sparse data in the open waters.

The observations were further averaged over the season of the year determined by northern hemisphere convention. This seasonal division improved on the uneven distribution of the *in situ* field thus providing samples from most parts of the working area, though still sparse. Both data fields were measured in mg per cubic meter and stored in matrices of the order 241 by 67 with the same grid size. The final working fields have dimensions 230 × 65 which lie between latitude 31.5 and 79.5°*N* and between longitude 88.5°*W* and 83.25°*E* of the equator.

SIMULATION STUDIES

An exponential bi-variate Gaussian function was used to simulate data over the working arena. From it, data fields for both *satellite* and *in situ* were extracted. The fields were made to match the real data of the second quarter by position and then scaled to match the satellite data. On introducing these simulated data fields into the blending process, it took 271 iterations to attain convergence. The resulting blended field is shown in Figure 1.

The plot of the blended field did not reflect what was expected. Knowing that the data fields are from a function with a known shape, and considering the fact that the values used for both satellite and *in situ* are from the same database and considered correct one would expect the plot from the blended field to match that from the *satellite* very closely. Figure 1 shows that this is not the case.

Problem identification

The primary objective of the blending process is to combine *in situ* and *satellite* data using the Poisson equation with the intension of producing a blended field that can be used to predict ocean chlorophyll at positions where the *in situ* field has no observations.

It is expected that, in situations where both data fields are from the same database and assumed to be correct, the resulting output from the blending process should match the inputs. Therefore, to better appreciate the result of the process, a perspective plot of the entire satellite and blended data fields were plotted as shown in Figure 2.

Figure 2 reveals that the distortion at the coastline boundaries of the plots in Figure 1, are principally caused by the presence of zero values at the external boundary points. These points turn to pull down the values from the blended field towards themselves since they remain unchanged in the course of the process. These results might have conspired to urge Gregg and Conkright (2001), to suggest the need for more work to be done in order to obtain improved results.

MODIFICATIONS TO DEAL WITH THE DISTORTIONS

The simulation study, revealed that the distortion of the blended field along the coastline, and the wide difference between it and the satellite field, are the result of the pseudozeroes used on the external boundary and the setting of the forcing term to zeroes on land and at the coastline. Based on these findings, alternative methods for solving this distortion problem are explored. The approach is from two perspectives:

(i) Handling values at the external boundary and
(ii) Handling the values at the coastline boundary.

Working with the external boundary

Here the problem posed by the use of pseudozeroes at the external boundary is addressed.

Linearly interpolating at the external boundary

The external boundary is made up of all the locations along the longitudes and latitudes forming a rectangular enclosure of the working area. Locations without observed values such as those on land were considered to have missing values. Linear interpolation is one method which could be used to impute missing values. This would provide an almost smooth transition across missing external boundary values and thus remove the effects of the pseudozeroes assumed during the process.

During the interpolation process, a temporary vector *t* is created whenever a location containing missing value is encountered. This vector *t* will then contain the sequence of locations having missing

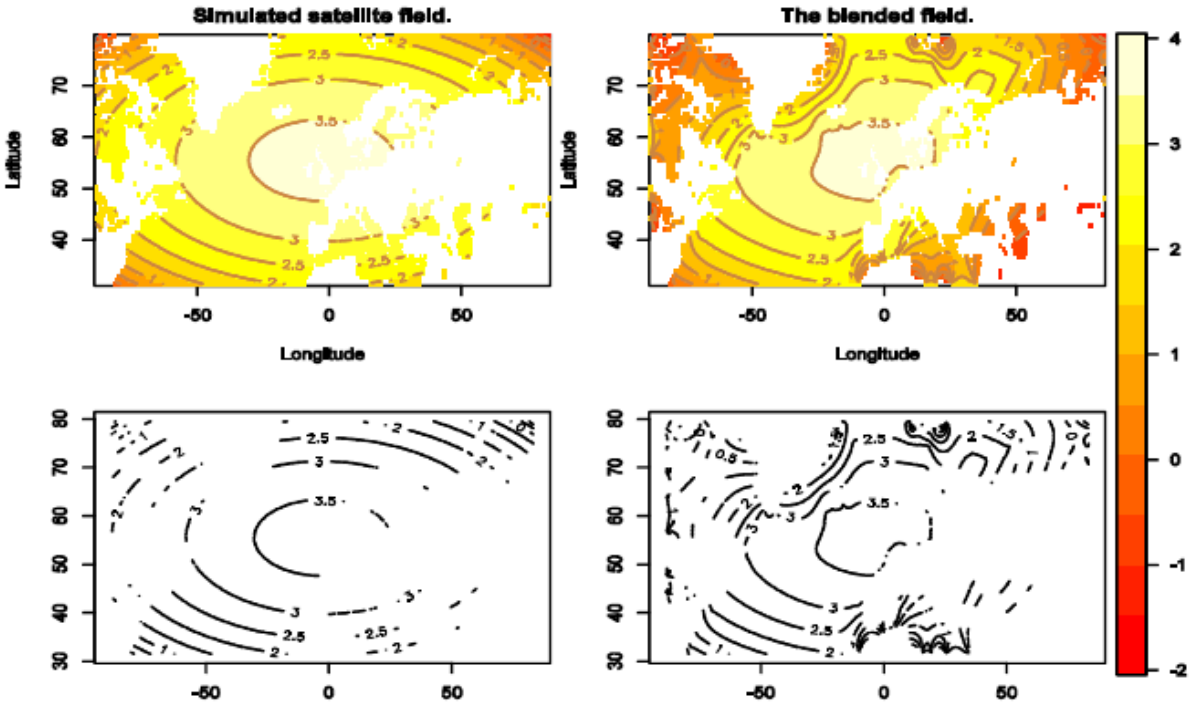

Figure 1. Image and contour plots from satellite and blended fields obtained from the blending process using simulated data.

values that follow until a new location with an observed value is encountered.

Then observations for each of the locations t_i with a missing value was then estimated using the linear interpolation formula:

$$f(t_i) = f(t_0) + [f(t_{n+1}) - f(t_0)] / [t_{n+1} - t_0] (t_i - t_0) \quad \text{for} \quad i = 1,..., n$$

where:

(i) i denote the index of the locations of the sequence of missing values in vector *t*, and
(ii) t_0 represent the index of the location with an observed value immediately before this sequence, with $f(t_0)$ its observed value,
(iii) t_{n+1} represent the index of the location with an observed value immediately after this sequence, with $f(t_{n+1})$ its observed value.
(iv) $f(t_i)$ is the estimate for the observation at location t_i.

There are therefore *n* locations with missing values in *t*.

The interpolated external boundary was then attached to the blended field as an initial solution to the blending process. The image plot of the resulting blended field is shown in Figure 3. This shows that the distortion has not been solved completely. The value of the mean squared difference at the coastline was reduced significantly, but the overall difference between the blended and the satellite data fields was only slightly affected as can be seen on Table 1.

From these plots, there is evidence that the external boundary is not the sole cause of the distortion in the blended field since a smooth transition of values has been provided by interpolating along this external boundary.

Working with the coastline

Working with the coastline gave an insight as to why the blending method worked well with the sea surface temperature. As stated by Reynolds (1988), it was possible to obtain *in situ* values at almost all the coastline because of the heavy ship traffic. Thus most of the coastline values were from the *in situ* data field and could serve as boundary points. This is not the case with ocean chlorophyll data. The *in situ* values at the coastline in the quarters where you could find some, were very few and may be highly contaminated, considering the activities at the coastal waters and the effects of run-offs from the land. In approaching the solution to the distortion problem using the coastline, two options were investigated.

Calculating the forcing term in the presence of missing values

The blending process currently sets forcing terms at the coastline and on land to zero because at least one of the satellite observations needed to compute them is not available. This results to a stiff jump in forcing terms from the points on the land edge to those in the sea coast. However, the blending method assumes a smooth transition in ocean chlorophyll concentration over the working arena though in reality there is no sea surface chlorophyll as soon as one gets to land. The aim here is to study the effect of resolving this mismatch in forcing terms by calculating the forcing terms at the coastline and on land in order to create a smooth field of forcing terms over the working area. For this to be possible, all the missing values in the satellite field were replaced by pseudozeroes. This permitted the calculation of forcing terms at the coastline and on land. The calculated forcing terms were then used in the blending process. The image plots of the resulting blended field showed a very close match to the satellite field as can be seen in Figure 4 and shown by the mean squared differences in Table 1. Despite this match, the method was discarded because it had no mathematical backing. This is because in reality, it is not possible to calculate the forcing term using finite difference whenone of the terms in the equation is missing.

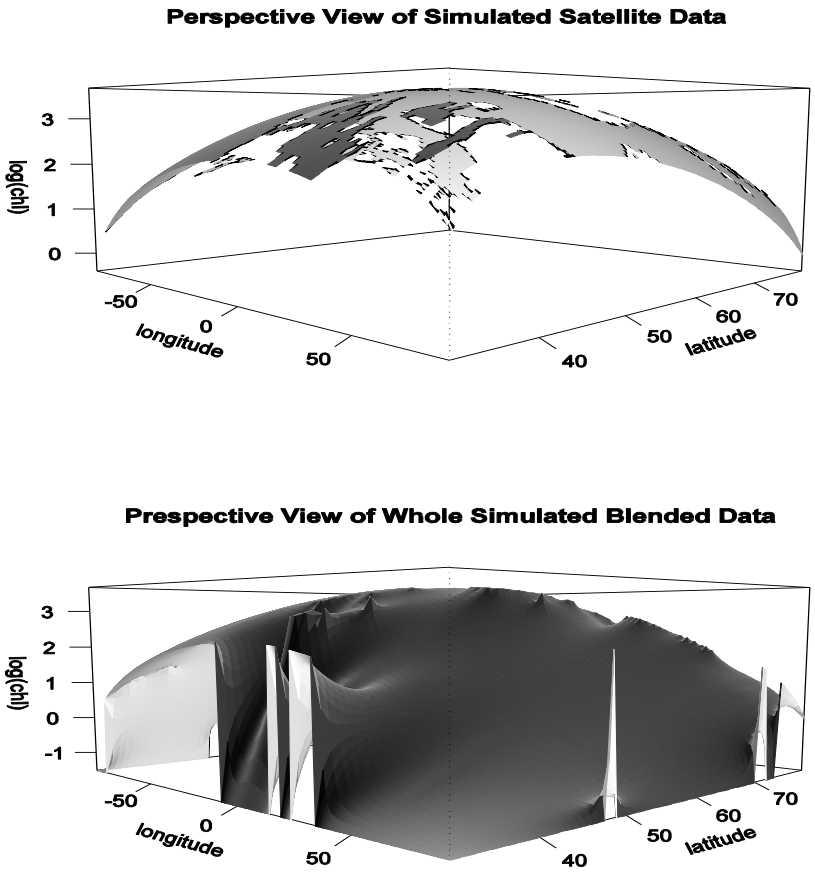

Figure 2. Perspective plots of the satellite field and the entire blended fields. The effect of the pseudozero values at the boundary can be seen clearly as they tend to pull down the final blended field (lower plot) with log (chl) = natural logarithm of chlorophyll.

Using the coastline as boundary points

The idea behind this technique is that the values at the coastline are assumed to be correct.

These values therefore serve as boundary values to the blended field during the blending process. The resulting blended field is shown in Figure 5.

Figure 5, shows a complete match between the blended and satellite fields at the coastline, this also seen numerically on Table 1 containing the mean squared differences. This is as expected since the satellite and *in situ* data fields were from the same database. This method could have been very successful if most of the coastal boundaries of the *in situ* field of the real data had observations. But this is not the case since in the real data fields, most of the values at the coastal boundary are from the satellite data field. This data field is known to be incorrect and the aim of the blending process is to correct it. Therefore, adopting this technique will mean assuming that the satellite values at these coastlines are correct which contradicts the initial claim.

The corrector factor method

In this technique, the blending process is performed twice. In the first run, an *in situ* field is extracted from the generated database asthe *satellite* field so that the *in situ* and satellite values are identical at positions where the original *in situ* field had observations. These fields are then introduced into the blending process. The difference between the satellite and the resulting blended fields in the first run is used as correction to the blended field resulting from the second run. In the second run, the blending is done as normal, making use of the original data from both *in situ* and *satellite* fields to obtain a blended field. The corrector factor is then subtracted from this blended field. The resulting field is the corrected blended field. The image plot of the corrected field matches completely the original satellite field. Moreover, the mean squared differences between the corrected blended and the satellite fields at both the coastline and over the entire working arena were very insignificant and were essentially zero as they were at the limit of tolerance for defining a number on the computer. This technique is seen to be successful after its implementation with the simulated data set.

The image plots in Figure 6 summarizes the results obtained at the difference stages of the 'corrector factor' technique using the simulated data fields.

The smoothed in-fill method

With this method, the idea is to smooth the transition of the second derivatives from sea to land. This was achieved by doing local linear kernel regression to fill in values on land.

The *sm.regression* function found in the *sm package* by Bowman

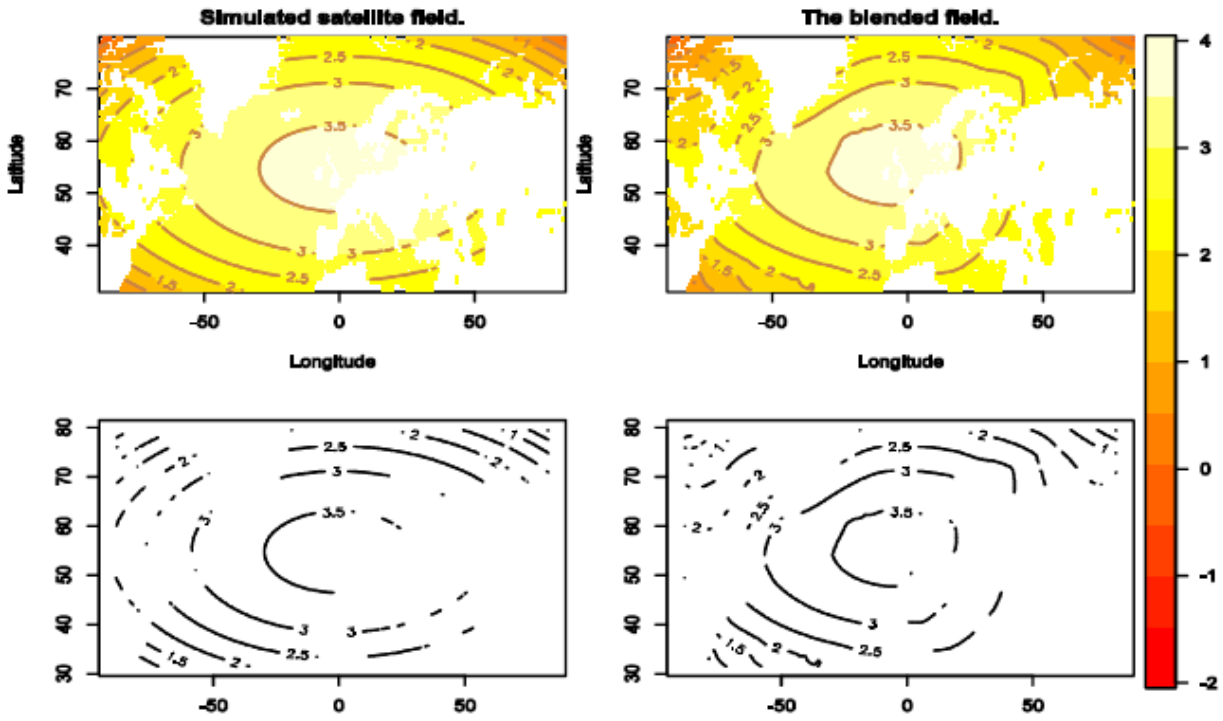

Figure 3. Image and contour plots from the simulated satellite and blended fields obtained when making use of the interpolated boundary.

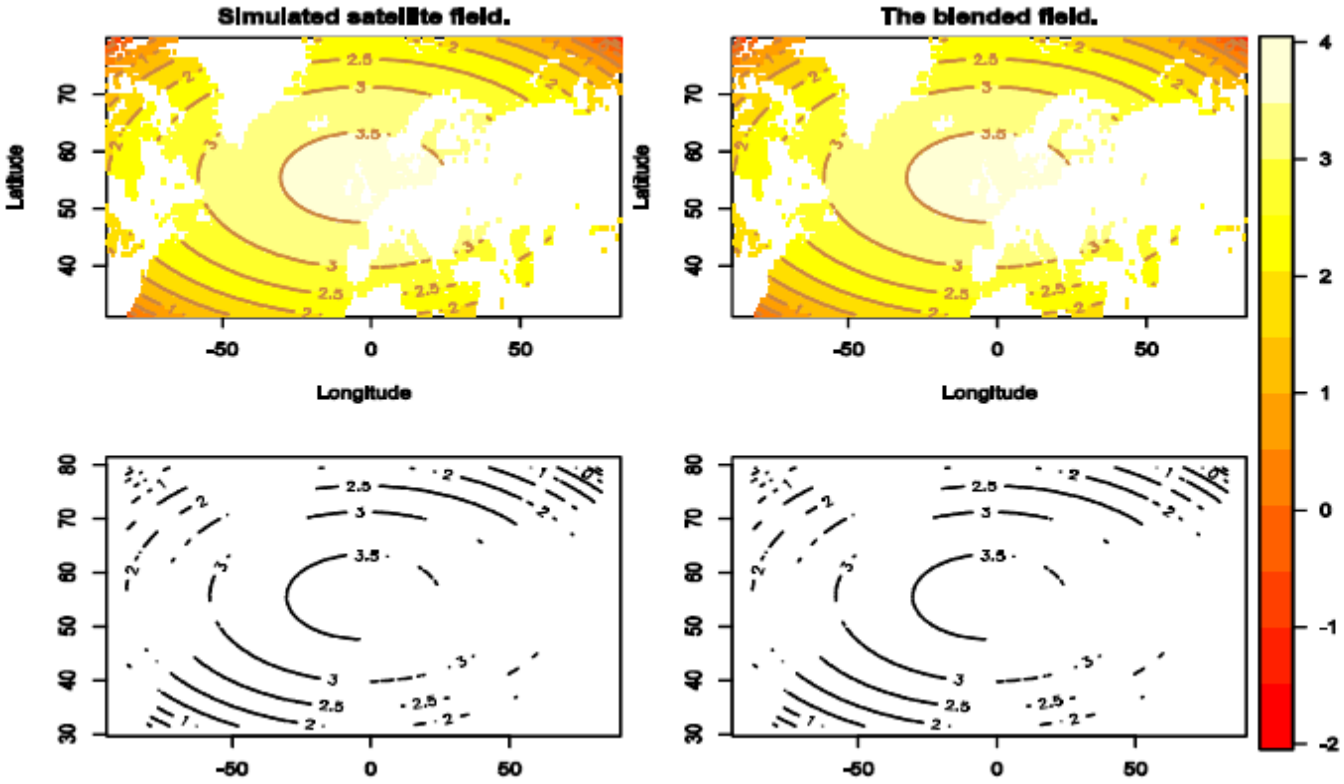

Figure 4. Image and contour plots from the simulated satellite and blended fields obtained from the blending process using forcing terms calculated in the presence of missing values.

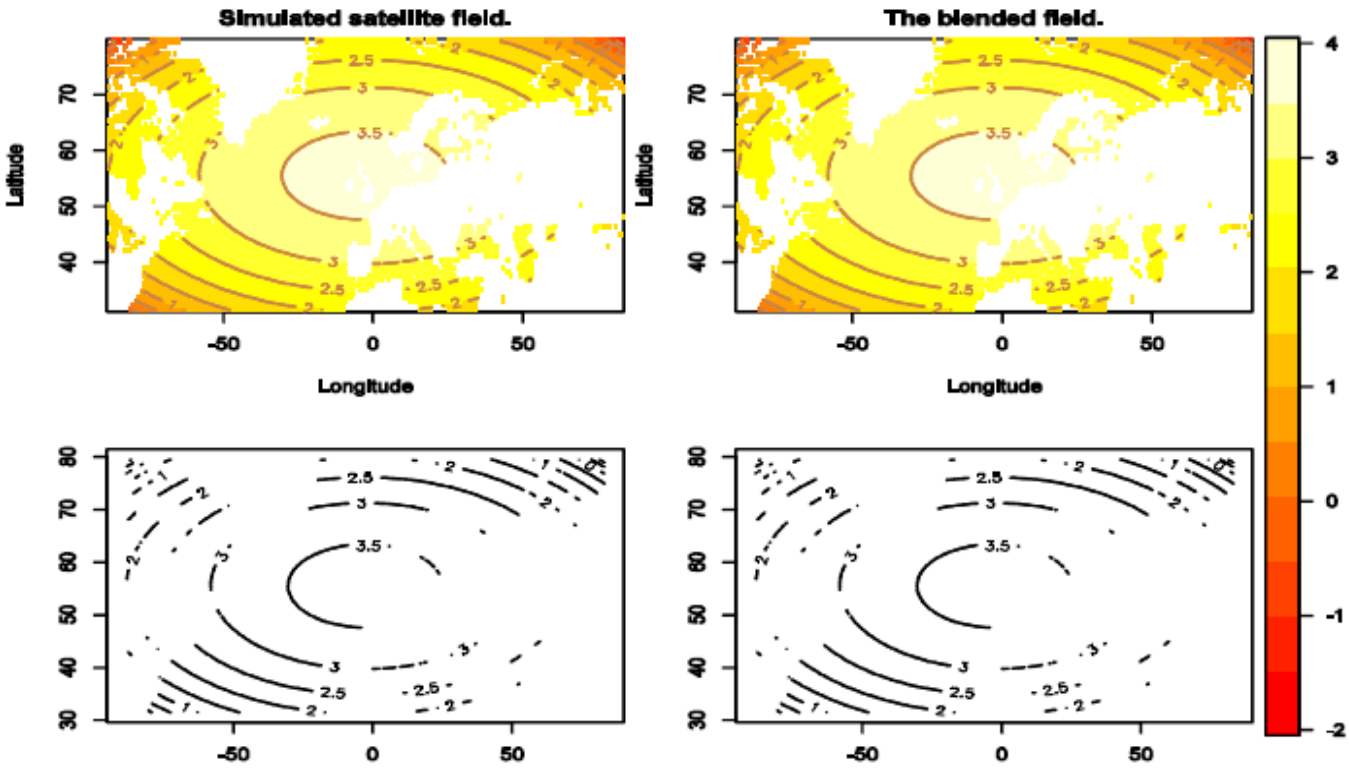

Figure 5. Image and contour plots from the simulated satellite and blended fields obtained from the blending process using coastline points as boundary.

Table 1. Mean squared differences between blended and satellite data fields when the simulated data sets were used in the normal blending and all the corrective methods.

Method	Difference at coastline	Difference between the whole blended fields
Normal blending procedure	0.6134563	0.2719859
Linearly interpolated boundary	7.22e-003	3.01e-002
Derivatives calculated at coastal point	2:02e-014	1:07e-014
Using coastline as boundary	0:00e + 00	6.20e-015
Corrector factor method	2.36e-034	9.36e-034
Using smoothed satellite data	2.97e-15	7.55e−015

and Azzalini (1997) running in the R programming environment was used with a bandwidth (h=(2,2)). This function creates a nonparametric regression estimate from data consisting of a single response variable and spatial covariates. The kernel regression estimator known as *the local polynomial kernel estimator* described by Wand and Jones (1995) was used. This was useful because it could permit the estimation of the regression functions at individual nodes in the working arena by locally fitting polynomials of degree zero, that is a constant, using weighted least squares. With this, all the grid points on land were filled with values and thus permitted the calculation of the forcing term over land.

For the *sm.regression* function to be used properly, grid points with observed values were identified and used as covariates to estimate values over the entire grid space spanning the working area. This was used as the response variable. This permitted the smoothing of the entire satellite field. The smoothed satellite field was then introduced into the normal blending process with the *in situ* data field. Figure 7, shows the resulting blended field.

RESULTS FROM THE VARIOUS TECHNIQUES USING SIMULATED DATA

Discussion of the simulated data results

The study on the simulated data identified the problem of the blending method as being distortion of the solution field as one approaches the coastline in the working area. This was seen to be caused by the fact that the second derivatives on land are set to zero and the external boundary points without observations from either the *in situ* or satellite fields are set to pseudozeroes. Alternative

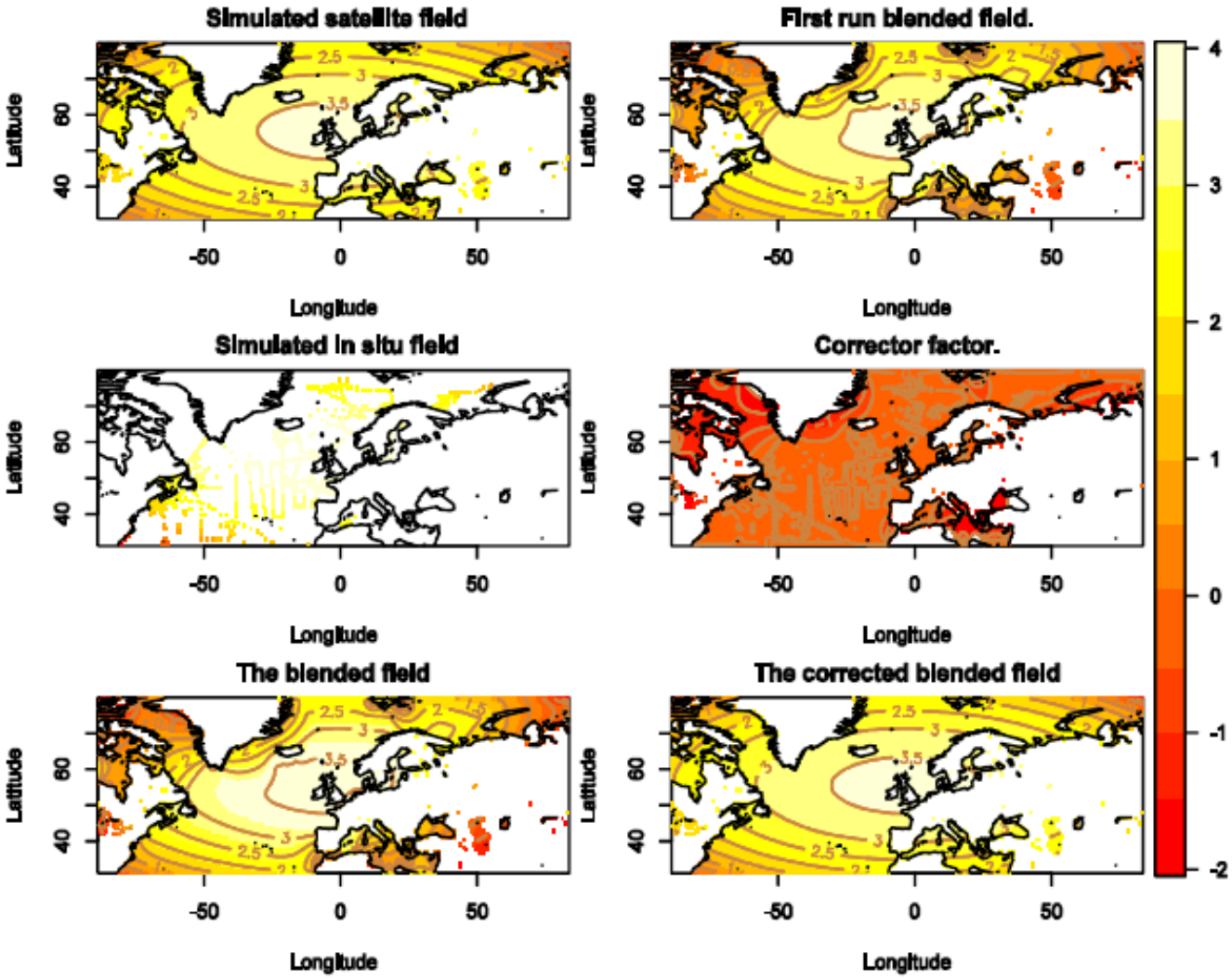

Figure 6. The bottom plots show the resulting blended fields from both normal blending and corrector factor techniques. These can be compared with the satellite field at the top left corner to see the differences and similarities respectively.

methods to solve the problem were proposed, discussed and implemented, and the results tabulated as previously shown. These results are interpreted as follows:

(1) The interpolated boundary procedure ameliorated thesituation at the coastline boundary but had little effect on the overall difference between the satellite and the blended data fields. Though this improvement could be seen numerically, the plot of the resulting blended field still showed some distortions as was seen on the contour plot. Therefore this technique could not correct the problem.

(2) Calculating the forcing term from the satellite field in the presence of missing values yielded lower values as differences between the satellite and the blended fields both at the coastline and on the entire working are during the illustration. The assumptions made during this process were mathematically unrealistic hence it can not be used.

(3) Using the coastline as boundary alongside the external boundary did quite well in the simulation study. This is expected because all the data fields are from the same database, thus the corresponding values are expected to be the same. This technique gave an insight as to why the blending technique worked well with the sea surface temperature. In it, most of the coastline values were from the *in situ* field and were correct. This is not the case with ocean chlorophyll. Thus it was dropped.

(4) The smallest mean squared differences between the satellite and blended fields during the study with using the simulated data were obtained when the 'corrector factor' and the smooth in-fill' methods were used. The mean squared differences from both techniques were far below the tolerance margin. The error margin (the difference between *satellite* and *in-situ* data fields) for the simulated data was set to zero.

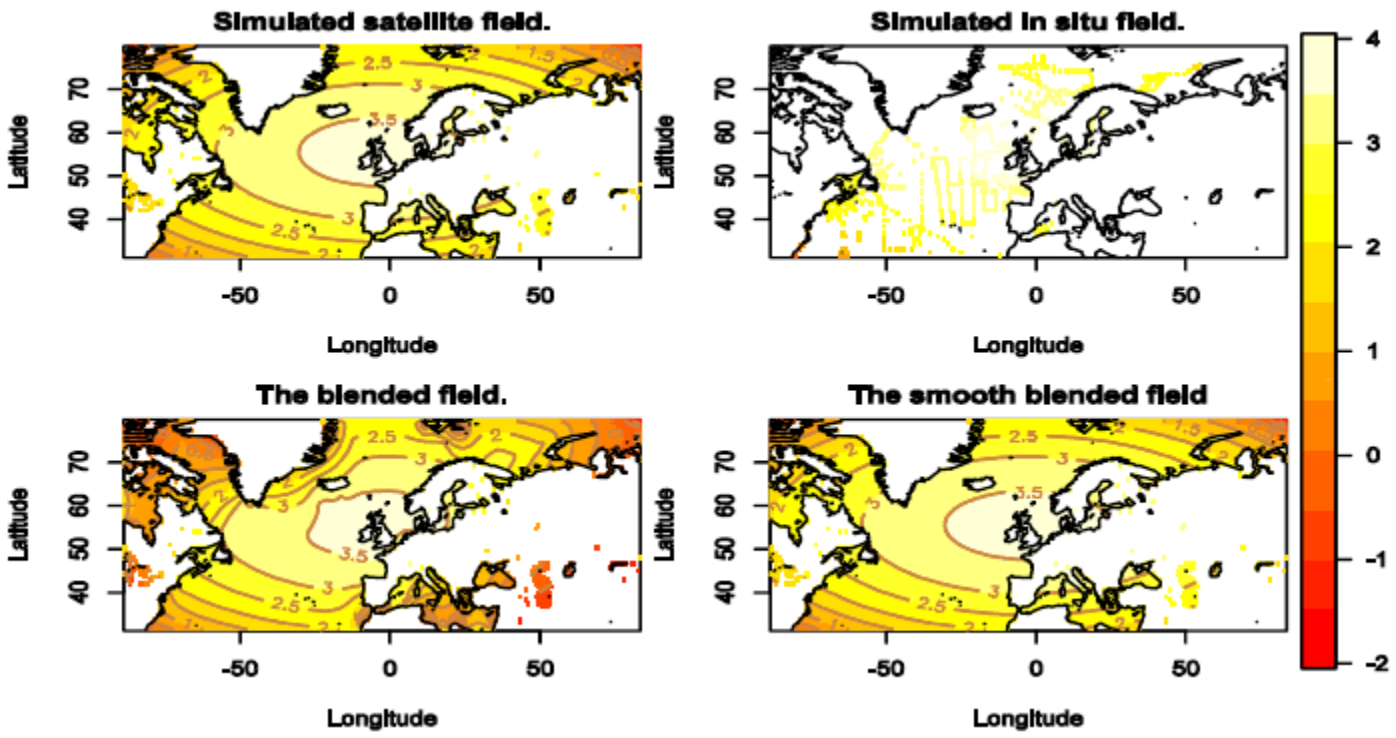

Figure 7. Blended field obtained from using smooth in-fill satellite data in the blending process.

APPLICATION TO REAL DATA

Application of the blended technique to the calibration of ocean chlorophyll requires further modification if improved results are to be obtained as suggested by Gregg and Conkright (2001). This suggestion has been confirmed in this paper as shown in the simulation studies where the two data fields are from the same database. There is therefore the need to modify this procedure, especially if it is to be used in situations where data are sparse. The corrector factor method has proven to be the most appropriate in correcting the distortion problem experienced by the blending technique in chlorophyll calibration.

The alternative methods of improvement discussed during the simulation studies were applied to the real data for the second quarter from the North Atlantic Ocean. Although the results obtained are from the analysis of data of the second quarter (April to June), the process could be successfully applied to all the four seasons.

Figure 8 show the distribution of observed chlorophyll from both the *in situ* and satellite fields from the second quarter followed by the plots of blended fields obtained using the normal, corrector factor and the smooth in-fill blending techniques.

The differences between the corrected blend and the normal blend can be seen clearly around the coastline. This is very prominent around the Black sea. In this area, there are virtually no *in situ* observations, hence it is expected that at the end of the process, the blended field should be more like the satellite field. This expectation is not met by the normal blending process, but it is seen in the corrector factor method. This confirms that this technique is a better way of handling sparse data during the blending process.

Table 2 shows the results obtained when all these methods were applied to the real data. The convergence criterion was set to 1×10^{-6}. The 'difference at coastline' refers to the mean squared difference between the resulting blended and the satellite fields at all the points along the coastline while the 'difference between the entire fields' refers to the mean squared difference between the blended and satellite fields over the whole working arena.

RESULTS FROM THE VARIOUS TECHNIQUES USING REAL DATA

Discussion on real data application

The original blending method exhibited distortion along the coastal area. This was best seen in areas that were almost completely surrounded by land. From the image plots, the area around the Black Sea and the coast of Greenland had extremely low and very high chlorophyll density respectively and so constituted the areas with

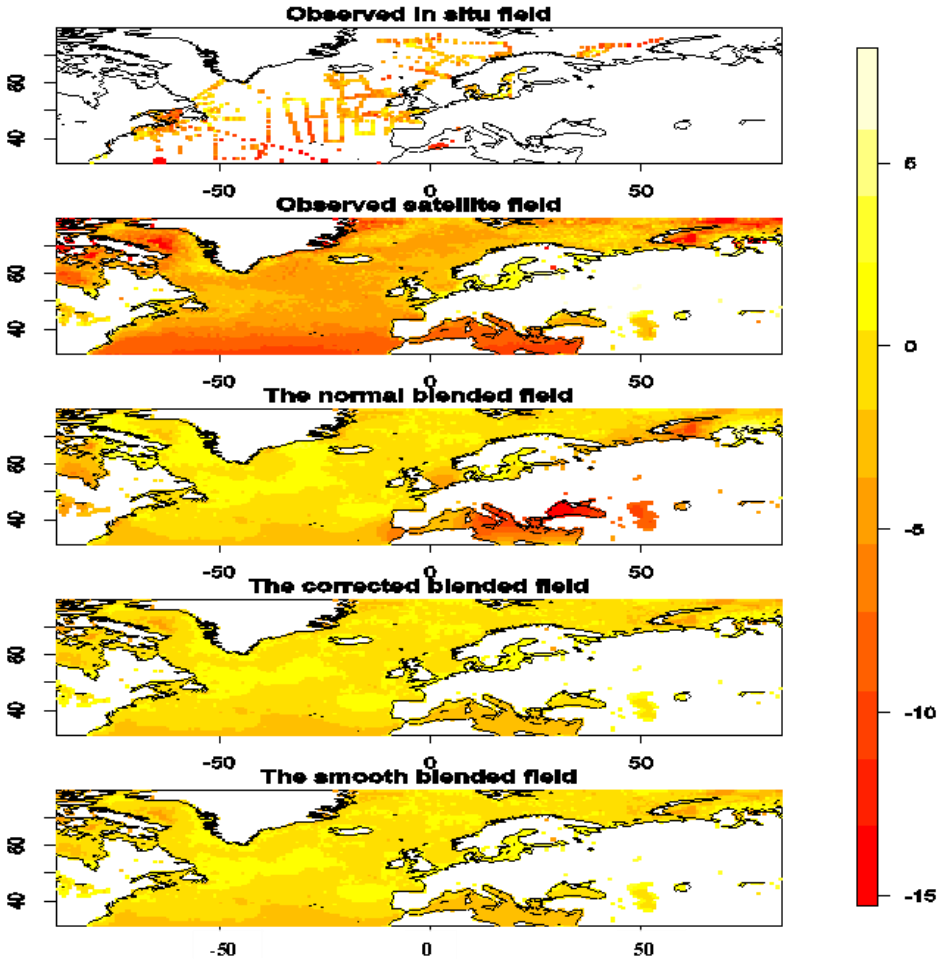

Figure 8. Image plots showing the distribution of chlorophyll form the *in situ*, satellite fields and the resulting blended fields from the normal, corrector factor and smooth in-fill techniques.

Table 2. Mean squared differences between blended and satellite data fields obtained when the normal blending and all the corrective methods were applied to the real data.

Method	Difference at coastline	Difference between the whole blended fields
Normal blending procedure	13.933980	6.298749
Linearly interpolated boundary	0.642991	5.915782
Derivatives calculated at coastal point	1.717635	1.021056
Using coastline as boundary	0.111843	0.260496
Corrector factor method	0.264751	0.3465249
Using smoothed satellite data	0.264751	0.3465249

extreme distortions. These areas were also outstanding in the simulation studies.

The alternative methods to solve the problem were proposed, discussed and verified. The results are shown on Table 2. These methods were applied to real ocean chlorophyll data from the North Atlantic with the following observations:

(1) The interpolated boundary procedure ameliorated the situation at the coastline as could be seen from the mean squared difference; but this procedure modified the overall difference in the real working area only slightly. Thus it is not suitable to correct the problem.

(2) When the forcing term from satellite field was calculated even in the presence of missing values, the differences between the resulting blended field and satellite were lower than those from the original blended field. Although this procedure violated mathematical

principles, the differences still exceeded the stated tolerance level. This method too does appear be acceptable as a possible correction to the problems resulting from the original blending method.
(3) The use of the coastline as boundary alongside the external boundary was actually an awkward decision considering the distribution of our *in situ* field observations. However, the differences were quite small when compared to the results produced by the original method. This procedure could have been a better option if the coastline was filled by values from the *in situ* field where they could serve as internal boundary points. But this is not the case with the real data. Most of the coastline values are from the satellite data field which is incorrect based on the assumptions of the process. Therefore, if this technique were to be used, it would have meant accepting that satellite values are correct. This contradicts the assumption that the satellite field is wrong.
(4) The smallest mean squared differences between the satellite and blended fields were obtained when the 'corrector factor' and 'the smooth in-fill' methods were used. In this case there is a direct match between the results obtained from both methods. This necessitated further investigation as these two methods do not have the same structure nor do they use the same principles in their mode of operation. From results obtained, an attempt to mathematically prove why these two methods should have identical results is outlined and an illustration using 1-dimensional partial differential equation is given.
The smallest mean squared differences between the satellite and blended fields were obtained when the '*corrector factor*' and the smooth in-fill methods were used. Because of the simplicity of the *corrector facto r*method and the fact that it removes the boundary artefacts without introducing other artefacts nor violating some underlying assumptions of the method, it is hereby highly recommended as the correction to the blending method.

PROOF AS TO WHY THE 'CORRECTOR FACTOR' AND THE 'SMOOTH IN-FILL' METHODS SHOULD COINCIDE

The blending technique as a whole is intended to produce a smooth field of ocean chlorophyll over the whole working area. It makes use of the available *in situ* data being used as boundary points and the second derivatives from the satellite data field to reproduce a blended field over the whole area contained within the external boundaries. The smooth in-fill method also uses the available satellite data to reproduce values in positions with missing values. It employs local linear kernel regression to estimate values at neighbouring locations. These two methods seek to achieve the same goal, but they do not have the same structure, neither do they use the same strategies towards achieving this objective. How then can they arrive at the same estimates?

Proof of the general case

In an attempt to prove why these two methods should have the same solution, we consider solving the second order partial differential equation in two-dimensions;

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \rho(x, y) \qquad (3)$$

given the ocean boundary set $B = \{f_i; x_i; y_i : i = 1; :::; n\}$ and known ρ. Let the solution be f'. If everything is keptthe same, except the k^{th} boundary point which is changed to $f_k + \Delta_k$, x_k, y_k, then given this modified boundary point, the new solution is easily seen to be $f'' = f' + g_k$, where g_k is the solution to

$$\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} = 0 \qquad (4)$$

subject to the boundary conditions $\{0, x_i, y_i : i = 1,\ldots, k-1, k+1,\ldots,n, \Delta_k, x_k, y_k\}$. Any other change would worsen the consistency of f'' with ρ, relative to f', or be inconsistent with the boundary conditions. So, if we start with satellite derived boundary values and change these to *in situ* derived boundary values, the corresponding change in the blended field depends only on the difference between the satellite and the *in situ* values and not on f'. Hence, any two methods which exactlyreproduce the satellite field over the ocean, when given satellite derived boundary points will yield identical results over the ocean when these boundary point values are replaced with *in situ* values. Therefore, both the infill-smoothing and the corrector-factor methods will produce identical results over the ocean. In fact, both methods produce blended fields of the form,

$$f_{blend}(x, y) = f_{sat}(x, y) + \sum_{k=1}^{n} g_k(x,y) \qquad (5)$$

where $g_k(x, y)$ is the solution to equation (3) with Δ_k set to the difference between the *in situ* and satellite values at boundary point k.

Given the ability of the method to reconstruct the satellite field from derived ρ and boundary points, then Equation (5) provides a proof by construction of the existence of (3) for a general blending problem.

Considering the 1-dimensional case

A better understanding of the proof can be achieved by

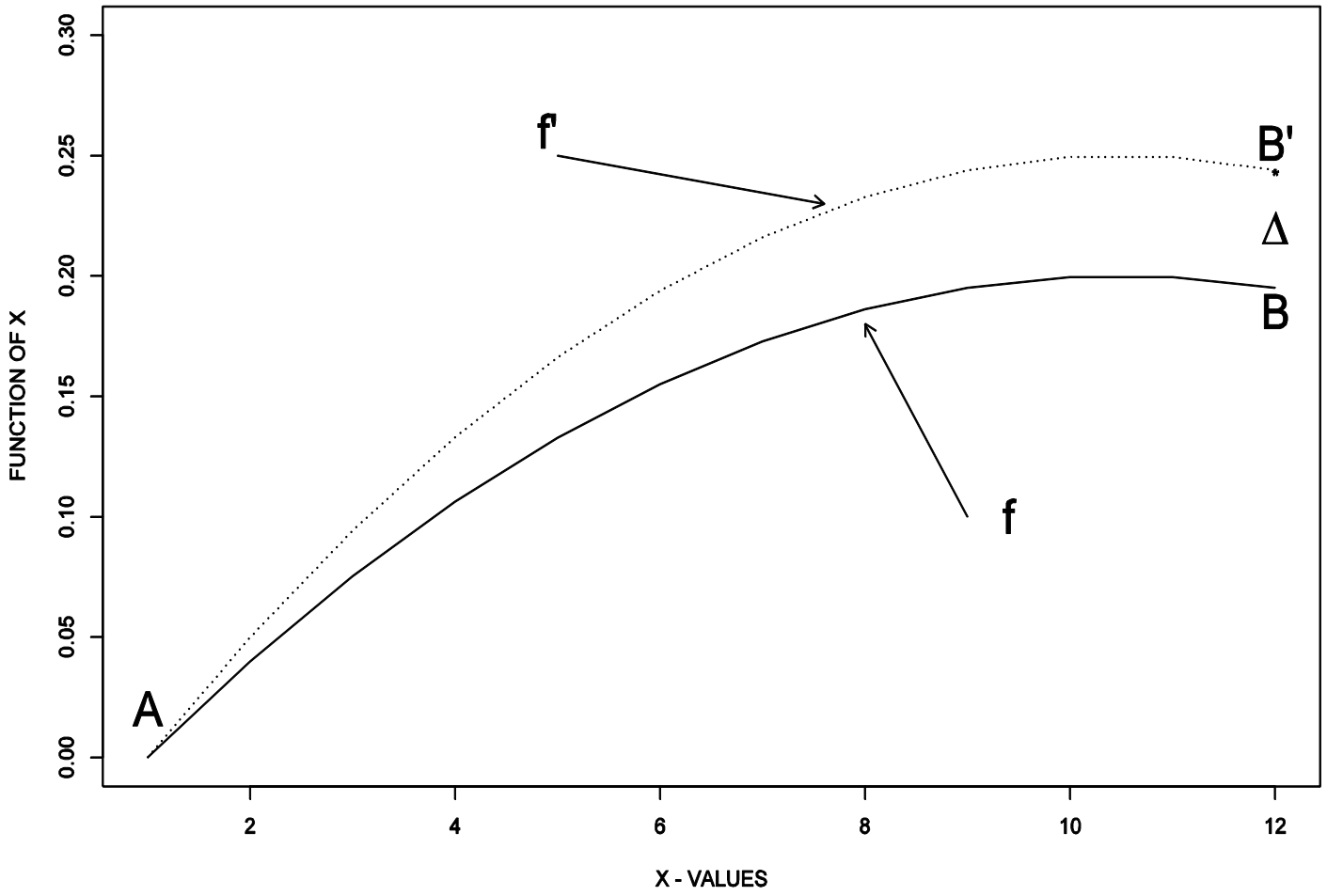

Figure 9. Dotted line for estimates and bold line for real function.

looking into the 1-dimensional situation. This can then be extended to higher order dimensions. Consider a section between two boundary points [A B] in 1-dimension as shown in Figure 9. If point B is moved up by an amount Δ to *B′*, making it inconsistent with the original field *f* then the method will simply increase the gradient of the reconstructed field between *A* and *B′* without changing the 2^{nd} derivatives. Any other change would worsen the 2^{nd} derivative match over the section or fail to hit *A* and *B′*. So, if *x* measures the distance along [A,B] and x_b is the distance between A and B, each point on the estimated field *A* and *B′* becomes,

$$f' = f^* + \Delta\frac{x}{x_b}$$

Now if there is no data over [A, C] say and the smoothing method is used to get an in-filled field *f** as shown in Figure 10. This filled in field f* is self-consistent and will be reproduced exactly from self-consistent data. If we move B up by Δ and apply the method as before, the reconstructed field becomes;

$$f'' = f^* + \Delta\frac{x}{x_b}$$

but *f** and *f* agree over [C, B] in which case so do *f ″* and *f′*.

Now consider the inconsistent situation Figure 11 in which missing segment AC is not filled in but just has 2^{nd} derivative set to zero.

The inconsistency means that, the reconstruction *g* does not match *f* over [C, B], but we can produce a correction, ∂ , that makes it match exactly. That is,

$$g' = g + \partial = f,$$

over [C, B].

Now if B is again moved up by amount Δ to *B′*, the method will again respond by increasing the gradient of *g* to get

$$g^* = g + \Delta\frac{x}{x_b}$$

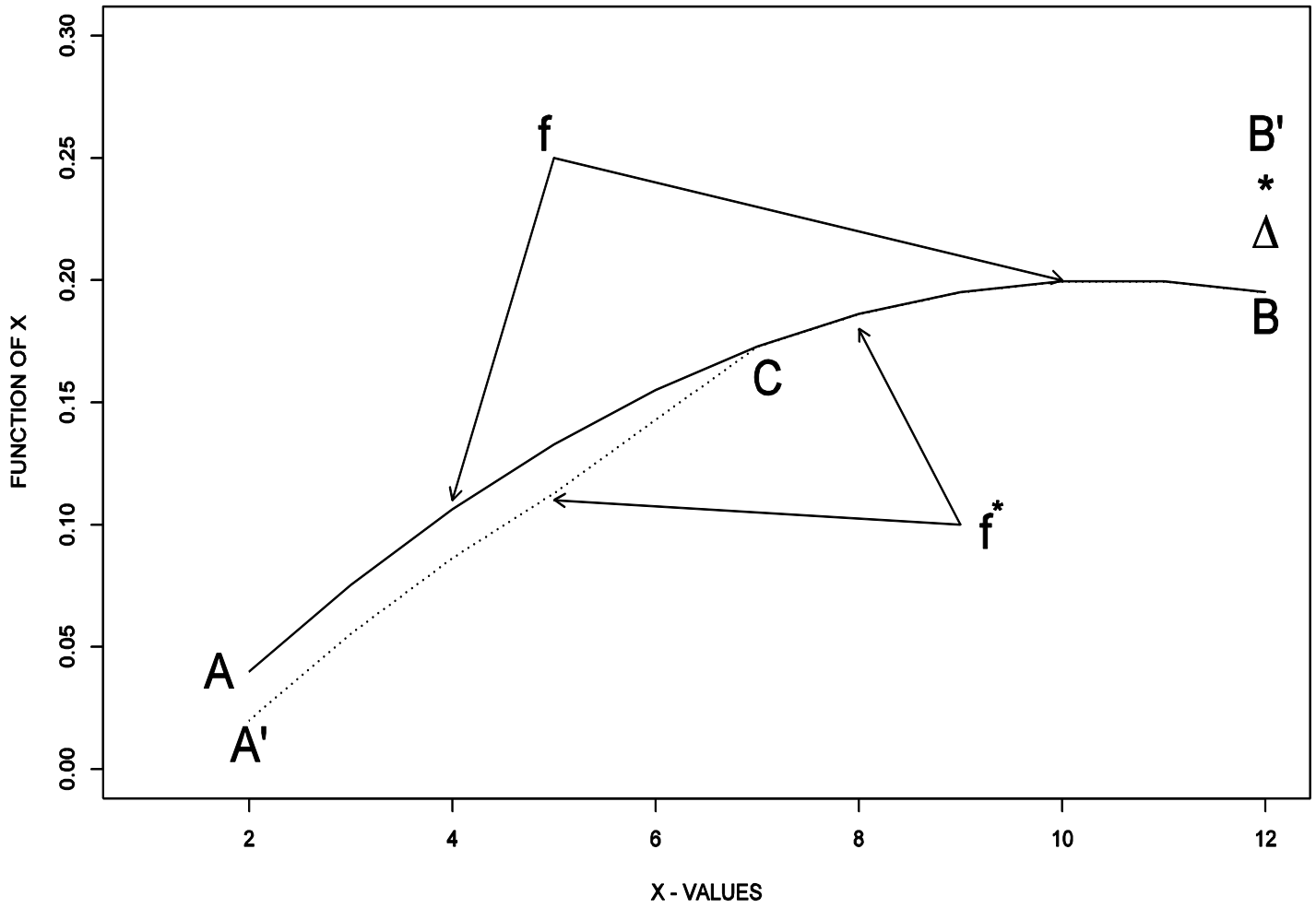

Figure 10. The situation in which the smoothing method is used to fill in the empty sector.

we can correct this to get

$$g* = g + \Delta\frac{x}{x_b} + \partial$$

$$= g' + \Delta\frac{x}{x_b}$$

$$= f + \Delta\frac{x}{x_b},$$

over [C, B] meaning that, the corrector and smoothing methods should match exactly over [C, B] thus proving the coincidence of both results.

VALIDATING THE RESULTING BLENDED FIELDS

In order to compare the accuracy of prediction resulting from the blended fields of the *corrector factor* and the normal blending technique, a validation study was done. To do this, twenty (20) randomly selected subsets representing *validation data* of size 100 from the *in situ* field were taken. The remainder of the *in situ* field was then introduced to the blending process alongside the satellite field. The resulting blended fields were then tested to see how closely they can each predict the values in the validation data set. During the selection process, care was taken to ensure that missing values were not included in the sample. The mean squared differences between the validation data and the predictions from the two methods were then calculated. From the validation study, it was seen that at least 80% of the time these methods were used, the *corrector factor* will provide a better estimate of chlorophyll concentration than the original blending method.

Figure 12 is a box plot of the mean squared differences between predictions from the blended fields and the values from the validation data sets. The tolerance margin was set to the mean squared difference between the observed satellite and *in situ* values. In the case of the entire data fields for the second quarter (April to June), this value was calculated to be 0.989.

From the box plots of the mean squared difference, it is evident that predictions from the *corrector factor* method

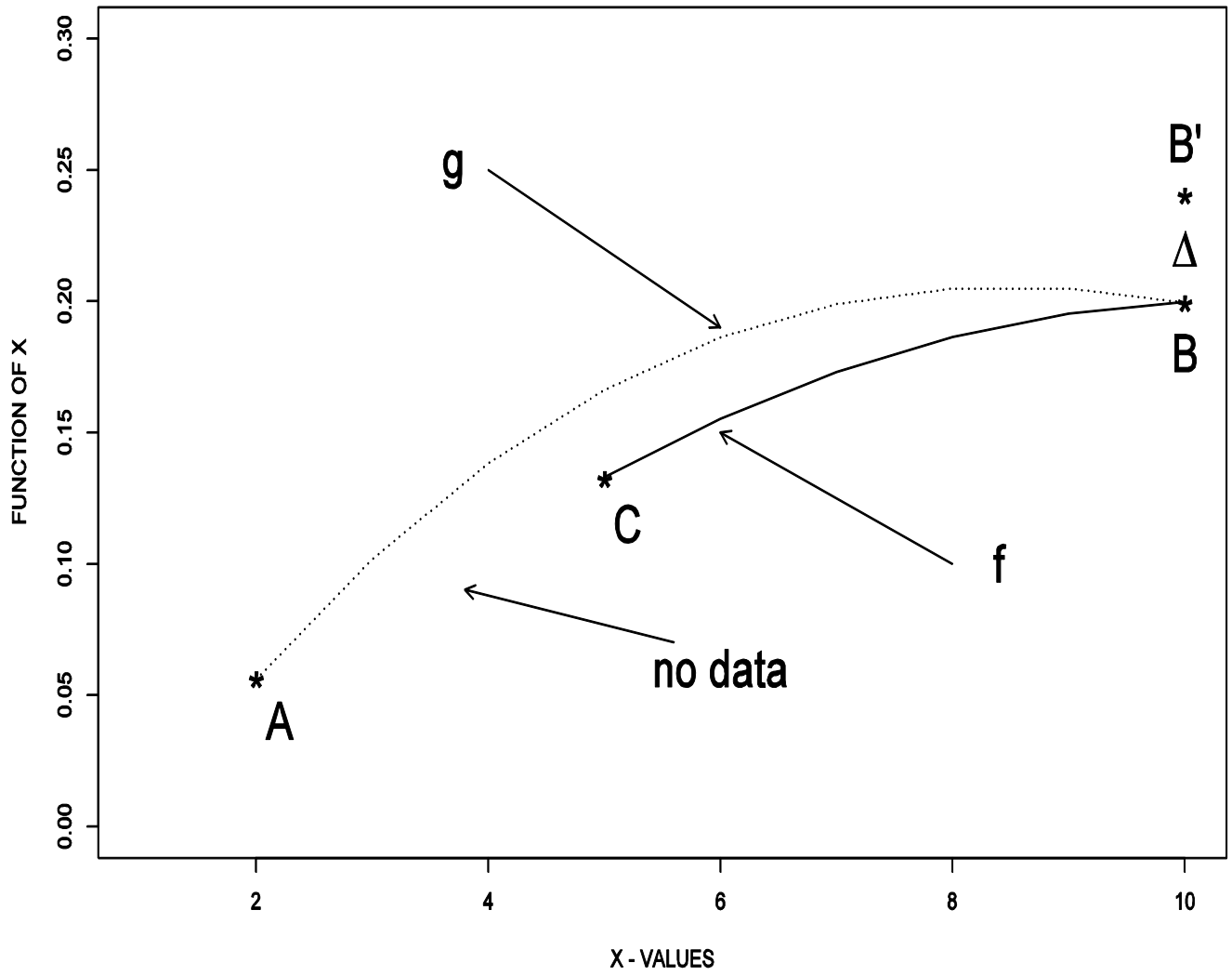

Figure 11. Inconsistent situation in which the missing segment is not filled with values. The derivatives at these points are set to zero; this could represent the land area in the real problem.

are closer to the the real values than those from the normal blending method. In addition, the heights of the boxes show that range from the *corrector factor* is quite small. This indicates that the predictions are close to the observed values as described by Agresti and Finlay (1997).

Conclusions

It is evident that the blending technique requires some modification in order to provide improved results when used for ocean chlorophyll calibration. This could be attributed to the wide range in chlorophyll values and because of the vastly reduced sampling of the *in situ* field. The constraints of transformation by taking naturallogarithm in order to reduce the effect derived from the extreme data range, the definition of seasons (Winter, Spring, Summer, Autumn) following the northern hemisphere convention, and the use of data from the recently launched sea-WIFS as imposed by Gregg and Conkright (2001) definitely improved the results, but did not solve the distortion problem encountered as one approaches the coastline.

In this article, causes of the distortion were identified, alternatives methods of modification to solve the problem were suggested and discussed; some by violating some underlying principles of the process, others by overlooking mathematical theories and principles. One of these methods herein termed the *corrector factor* method in which the original principle of the process was maintained, but the process performed twice, happened to correct the problem and provided the expected results in the simulation studies. When applied to the real data, the results were reasonably better than any of the other alternative methods. This method is further backed by a

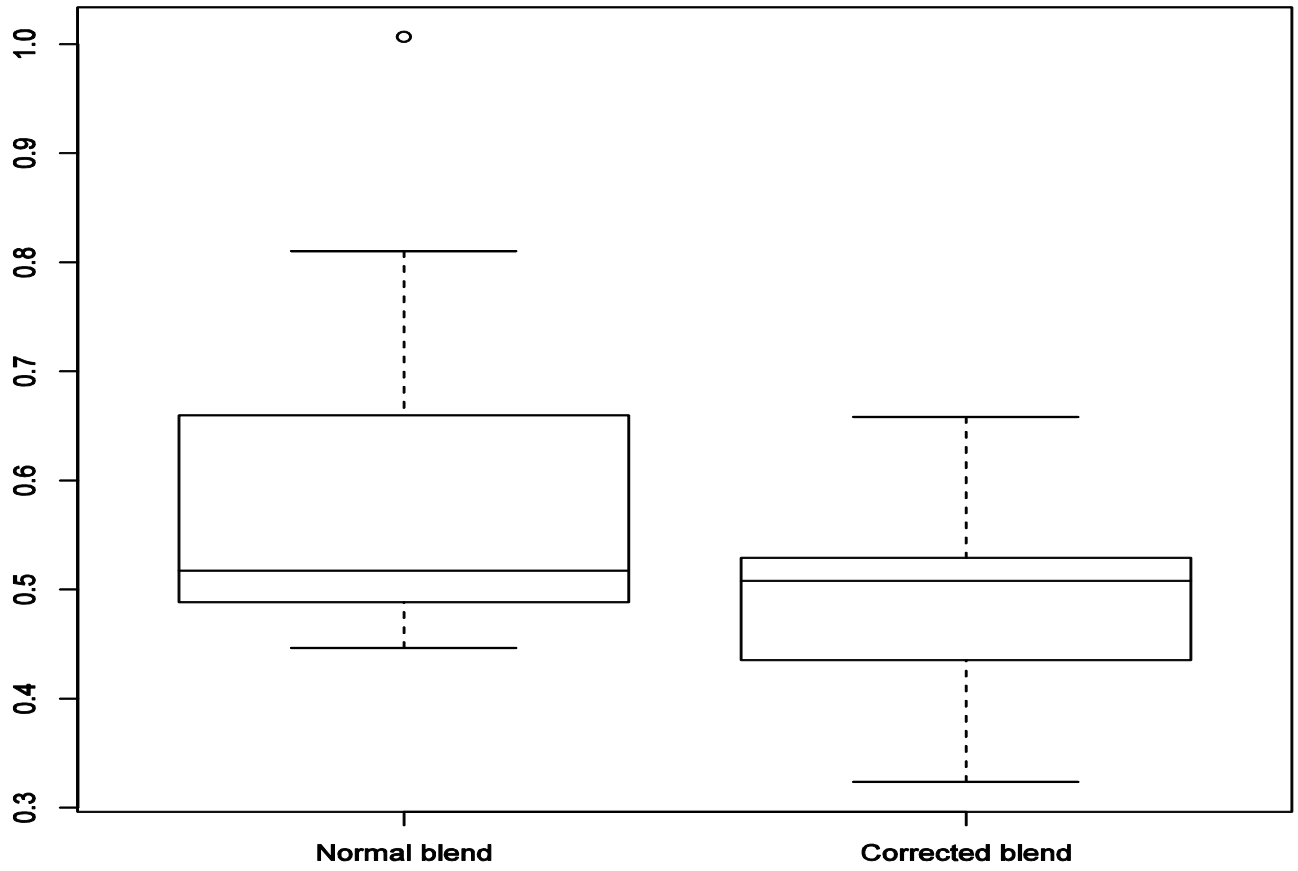

Figure 12. A plot of the mean squared difference between predicted and observed *in situ* values from both the normal and the corrector factor blending methods.

standard statistical procedure which produces identical results to those obtained by using this technique even though the two methods differ in structure. This suggests that this technique has some mechanism in its mode of operation which handles noisy data better than some parametric statistical methods when it comes to analyzing sparse data such as the type encountered in this research.

It is thus conclusively ascertained that, the *corrector factor* blending method herein developed should be adopted as a better tool for calibrating remotely sensed data because of its reliability, ease of implementation and minimal or no need for additional programming.

This research also suggest that the availability of more in-situ observations will improve ocean chlorophyll calibration and portends the hope that these results will significantly improve analysis on primary productivity and management of the marine environment since chlorophyll is one of the most important components in the formation of the ocean life cycle.

ACKNOWLEDGMENTS

Appreciations go to the sources of the data provided for the research and especially to Steve Holmes who extracted and prepared the *in situ* data used in this research from the World Ocean Data base.

REFERENCES

Agresti A, Finlay B (1997). Statistical Methods for the Social Sciences, third edn, Prentice Hall International, Inc.

Bowman AW, Azzalini A (1997). Applied Smoothing Technique for Data Analysis, The Kernel Approach with S-plus Illustrations, Oxford University Press.

Clarke E, Speirs D, Heath M, Wood S, Gurney W, Holmes S (2006). `Calibrating remotely sensed chlorophyll-a data by using penalized regression splines.' J. Royal Stat. Soc., 55(3): 331-353.

Edgar SL (1992). FORTRAN for the '90s Problem solving for Scientists and Engineers. Computer Science Press, USA.

Gregg WW, Conkright ME (2001). 'Global seasonal climatologies of ocean chlorophyll: Blending *in situ* and satellite data for coastal zone colour scanner era.' J. Geophys. Res., 106(c2): 2499–2515.

Gregg WW, Conkright ME (2002). 'Decadal changes in global ocean chlorophyll.' Geophys. Res. Lett., 29(15): 14–15.

Gregg WW, Conkright ME, O'Reilly JE, Patt FS, Wang MH, Yoder JA, Casey NW (2002). 'Noaa-nasa coastal zone color scanner reanalysis effort.' Appl. Optics, 41(9): 1615–1626.

Oort AH (1983). 'Global atmospheric circulation statistics.' NOAA Prof. Nat. Oceanic Atmos. Admin., (Silver spring. Md.), 14: 180

O'Reilly JE, Maritorena S, Mitchell BG, Siegel DA, Carder KL, Garver SA, Kahru M, McClain C (1998). 'Ocean colour chlorophyll algorithms for seawifs.' J. Geophys. Res., 103(c2): 24937–24953.

Paciorek C (2006). SpectralGP: Approximate Gaussian processes using the Fourier basis. R Package version 1.0.
Press W, Tenkolsky S, Vetterling W, Flanney B (1992). Numerical recipes in C programming Language, CUP.
Reynolds RW (1988). 'A real-time global sea surface temperature analysis.' J. Clim., 1(c2): 75–86.
Venables W, Smith DM, R Development Core Team (2003). R: A programming Environment for Data Analysis and Graphics, 1.8.0 edn.
Wand M, Jones M (1995). Kernel Smoothing, 1st edn, Chapman and Halls.