Geostatistical Validation of a Marine Ecosystem Model Using In Situ Data

Nansen Environmental and Remote Sensing Center

http://www.nersc.no

by Marielle Inizan

Technical Report S-435
Centre de Géostatistique, Ecole des Mines de Paris
35 rue Saint Honoré, F–77305 Fontainebleau, France
http://cg.ensmp.fr

July 2002

Acknowledgement

I would like to thank my supervisor at the Centre of Geostatistics, in the École des Mines de Paris, Dr. Hans Wackernagel, for all his precious help and advice, and my supervisors at NERSC, Dr. Lasse H. Petterson, Dr. Dominique Durand, and Prof. Geir Evensen, for their support and helpful comments on my results. I would also like to thanks all the people from Nansen Center, for their warm and friendly welcoming in Bergen.

Contents

Introduction

Marine Ecosystem Models have been developed throughout the last few years by the Nansen Environmental and Remote Sensing Center to answer the growing concern about understanding and monitoring the evolution of oceans and seas all over the world. One of the major stakes of this development is to validate the ability of the model to give a likely image of what reality is, or must be. This ability has been largely improved by using data assimilation techniques coupled to the physical and bio-chemical modeling of the oceans. One usual way to confirm the validity of the model is then to confront the model results with a pre-established simulation of reality that it should be able to reproduce.

Another way to test a model is to compare its results in real conditions with *in situ* or satellite measurements. However, in most of the cases, this comparison can not be directly drawn, and transforming the measurement data in order to fit the model format is a necessary step of the analysis. Furthermore, there is a real need for precise quantitative tools to compare properly the model results and the observations extracted from the reality. My last year practical work in Geostatistics at the École des Mines de Paris, in collaboration with the Nansen Center, aimed at defining through a practical study a way to perform those two analysing steps with the tools and techniques developed by statistics and geostatistics. Therefore I focused on the comparison between the results of a model running on the north of the Atlantic Ocean, and developed by the NERSC in the European Community project DIADEM, and *in situ* data extracted from the OMEX project, which built a complete physical and bio-chemical dataset for precise zones in the North-Atlantic.

This study is divided into 3 parts. The first part consists in a description of both the DIADEM and the OMEX I datasets, insisting on their owns specificities, which make the comparison between numeric model results and *in situ* data more difficult. The second part is a presentation of the geostatistical method established to transform the OMEX data in order to fit them with the DIADEM results, called the Averaged Geostatistical Simulations method, and includes both a theoretical description of the method and the practical application on the OMEX in situ data. Finally, in the third part can be found a quantitative comparison between the results of these Averaged Geostatistical Simulations (AGS) and the results of the DIADEM model.

Chapter 1

The OMEX Dataset and the DIADEM Model Results

This practical study is based on two distinct datasets that totally differ one from the other, which make the comparison between both difficult. We will first present the DIADEM model, how it works and what kind of results will be used from it. Then we will describe the OMEX dataset, and the data we shall extract from it in order to perform our study. We will then shortly sum up the main differences between the two dataset obtained, and that we will have to overcome to perform the analysis.

1.1 The DIADEM Model Results

1.1.1 General description of a Marine Ecosystem model

A marine ecosystem model is usually composed of two coupled model : a physical one and a biochemical one. The physical models describe the dynamics of the ocean, using complex physical equations to model the interaction between parameters such as temperature, salinity, ice, currents, etc., and taking the atmospheric forcing into account. In order to run this type of models, the ocean is divided into vertical layers according to the depth, either regularly ($z-$coordinate model), or proportionally to the total depth ($\sigma-$coordinate model), or according to the potential density (isopycnic model).

The biochemical model is usually divided into a certain number of compartments, and integrates the continuity and transport equations on each of these compartments on a given grid, describing the relationship between compartments, and taking into account the forcing by physical fields. The coupling between the physical and the biochemical model is then done by determining those fields from the results of the physical model. For example, a fairly simple model is the one developed by Evans and Parslow (1985) and describe in (Breuillin, 2000 [3]), were you can find 3 compartments: Nutrients (N), Phytoplankton (P) and Herbivores (H); the equations for this model then traduce the fact that the growth of P depends on the photosynthetic activity, which depends on physical parameters such as light, and is limited by N concentration; moreover the concentration in P decrease because of herbivore grazing on phytoplankton, and this very graze has a growth effect on H; H decrease is due to mortality and to carnivores; N is absorbed by the phytoplankton during the photosynthesis, but is regenerated from the bottom of the sea, and there again physical parameters have a great part to play. The outputs of the biochemical model usually give the concentration of the related parameters in each cell of the calculation grid.

These two models can be improved by using the techniques of Data Assimilation, implemented at NERSC by G. Evensen (see [9] for example). Assimilating data consists in taking into account in the model external observations of the reality. The coupled model exposed above produces a estimation

state, or a forecast of the oceanic system at a given instant t ; if at the same instant t an observation of this oceanic system is available, we can improve the estimate state statistically by integrating this observation, which can be partial on the domain where the model's results are defined. Assuming that both the model and the observations are false, each one including a statistical error, this improvement is done by choosing an new estimate that minimize a penalty function that would correspond to the sum of the normalized errors squares for the model and the observation in a simple 0-D case.

This very short description of such coupled models is sum up on figure 1.1 ; the aim of this section is not to explain the structure of such complex models exhaustively, but to give a short outline about their functioning, in order to be able to give wiser comments on the results of the comparison obtained afterwards. To get more information about those models, please refer to the bibliography.

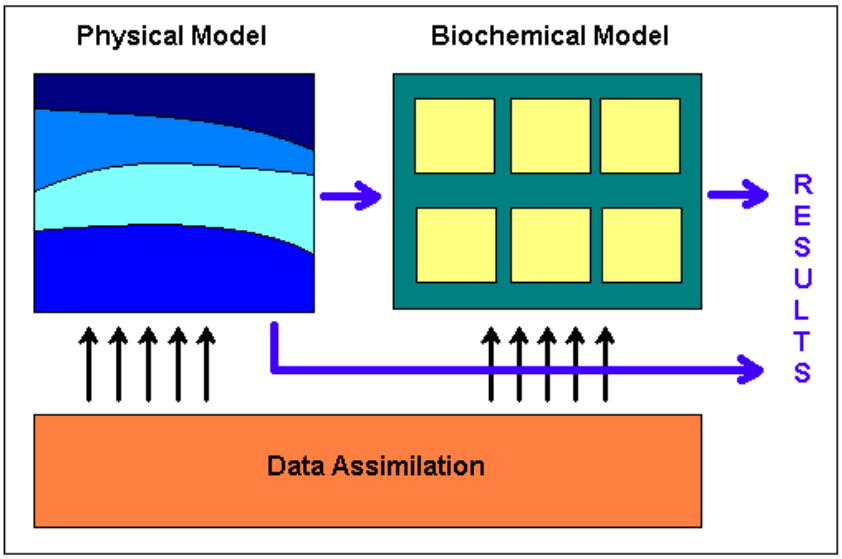


Figure 1.1: Global scheme of an Marine Ecosystem Model

1.1.2 The DIADEM Model

The model used in the DIADEM project follows exactly the previous scheme. It is a coupling of the physical model MICOM (Miami Isopycnic Coordinate Ocean Model), by Bleck *et al* (1992) [1], and an 11 compartments biochemical model, which describes the ecosystem of North Atlantic (Drange, [7] and [8]). Moreover, the model results have been improved by assimilating chlorophyll data from the Sea WiFS ocean colour data.

The physical model is an isopycnic model, that is to say that the vertical layering follows surfaces of constant and predefined potential density, which don't match with the iso$-z$ surfaces (where z would be the vertical Cartesian coordinate). Although this gridding has a lot of advantages in terms of physical model, the grid on which the model is defined is then not regular at all in the vertical direction.

The bio-chemical model is defined on the 11 following compartments :

- Phytoplankton
- Zooplankton
- Bacteria
- Nitrate

- Ammonium
- Dissolved organic nitrogen
- Dissolved organic carbon
- Detritus ; particulate organic nitrogen
- Detritus ; particulate organic carbon
- Alkalinity

A more precise description of the equations which govern the evolution of these different compartments can be found in (Natvik, [12]). We have described the vertical layering used for the physical model above ; however, this layering induces a relatively thick mixed surface layer, which doesn't allow to describe precisely the biochemical phenomena taking place mainly in the 100 upper meter of the total oceanic depth. Therefore the surface layer of the physical model has been splitted into 2 layers for the biochemical layer. We obtain then 17 physical and 18 biochemical vertical layers for the coupled model.

As far as the horizontal gridding is concerned, the coupled model was set up on an orthogonal curvilinear $140 * 130$ grid generated by conformal mapping. To perform this mapping, the North and South pole have been moved onto earth, in order to avoid singularities in the model domain, and have been located close to each other, in order to have a fine resolution grid in between, that is to say on the very North part of the Atlantic. This grid is described on figure 1.2 . One can noticed that it is not regular from a carte Sian point of view, which is the one of geostatistics, and that the more south one goes, the wider the grid cells are.

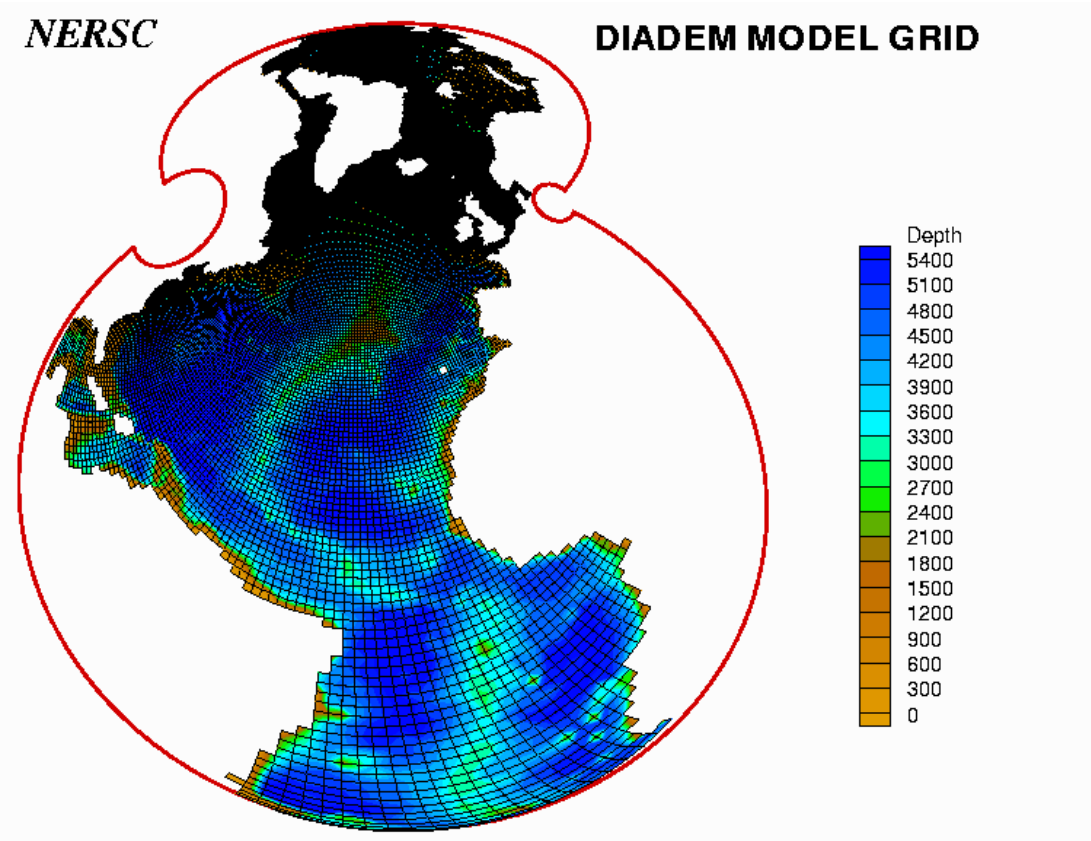


Figure 1.2: The DIADEM model grid

Before performing data assimilation on the model, one have to initialize it. This spin up period should last at least 10 to 15 years for the physical model to reach a stable state, and 2 or 3 years for the

biochemical one. In the case of the DIADEM model, the physical model was spun up on a $280 * 260$ grid (*i.e.* double resolution from the normal grid) for about 10 years (1985-1995). Then the biochemical model was coupled to it and run for 3 years (1995, 1996 and 1997) on the same doubled grid before the data assimilation process started, in 1998, on the normal grid described above.

1.2 The OMEX I dataset

The Ocean Margin EXchange (OMEX) project is a research project supported by the European Community aiming at gaining a better understanding of the physical and biochemical processes on the oceanic margin. Therefore it launched a series of measurements campaigns on three distinct zones : the northern Norwegian margin, the Goban Spur in the Celtic sea, and the Iberian margin, involving 40 principal investigators from 10 different European countries. From these data a model was developed for a marine ecosystem on the oceanic margin.

We concentrated mainly on the first part of the project, OMEX I, and on the data collected during this phase, which took place between 1993 and 1996, and focused on the Goban Spur area, in the northern part of the Gulf of Biscay (see location on figure 1.3). This area is characterized by a broad continental shelf, allowing the vertical mixing of nutrient rich deep ocean water with the surface mixed layer. The investigations were held through 47 research cruises involving vessels from 9 European countries, and were divided into 5 main fields : Physics, Biology, Biogeochemistry, Benthics and Air-Sea Interactions. The results of these investigations consist in 600 datasets gathering measurements from the air, the water columns and the sediments; over 95% of these data are found on the OMEX I Data Set CDROM, edited by the British Oceanographic Data Center (B.O.D.C.) (1997) [4], where over 800 parameters are listed according to a complex table organization. A parameter corresponds to one specific physical, biologi-

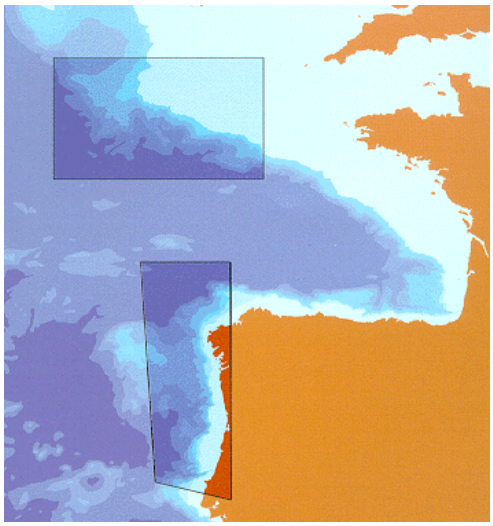


Figure 1.3: Mapping of the zones investigated by the OMEX project ; our zone of interest is located on the Celtic Sea (northern part)

cal, or chemical quantity, measured using a precise experimental protocol. More over, each parameter value is recorded in terms of events. An event is defined as “an action that results in the generation of oceanographic data” [11] ; this means that to each parameter value is attached not only the coordinates in space and time, but also very useful and accurate comments like the cruise during which the datum was obtained, the quality of measurement, the organization which performed it, and the originator of the datum. It allows the user of the dataset to rely on it even if it is not used as foreseen by the OMEX project team (as we did in this study).

The spatial repartition in the OMEX dataset is highly conditioned by the way the measurements where made, *i.e.* during oceanographic cruises. This implies a finer resolution in the vertical direction (*e.g.* a data can be found every 10 meters) than in the horizontal ones (a data can be found every 5 *km*!). Moreover, the Goban Spur area was covered more or less regularly during the different measurement campaigns, so that some part of the zone are better informed than others for certain parameters.

For more information on the OMEX project, please consult their website :

www.pol.ac.uk/bodc/omex/omex.html
A very useful description of the Project and the Dataset is also given in the documentation delivered with the CD-ROM [11].

1.3 The data used in our study

The aim of our study is to compare the results of The DIADEM model to *in situ* measurements extracted from the OMEX dataset. Therefore we have to define precisely which data will be used from both of the dataset presented above.

1.3.1 The variables of interest

First we shall choose the variable of interest of our study. Regarding the DIADEM model, it will be of high interest to analyze physical, biological, and chemical variables. Besides, we shall choose variables corresponding to well-informed parameters in the OMEX dataset, that is to say parameters with a good spatial repartition, and with a sufficient amount of points with a reliable value. Moreover, in order to get a dataset as complete as possible, we shall use variables for which the different measurement methods used can be mixed without any problem, that is to say that the measurement methods shall not induced a too big bias in the value of the parameters studied.

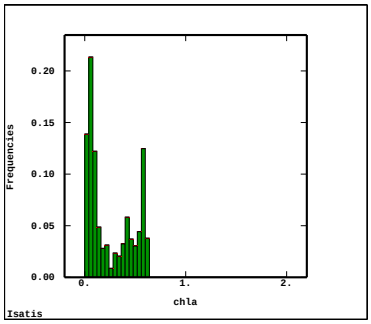
Taking these remarks into consideration, we shall focus on three different variables : the phytoplankton, measured in the OMEX dataset by the chlorophyll-a parameter, the nitrate, and the temperature.

The phytoplankton is one of the main variables for marine ecosystem modeling. Indeed it is very representative of the biological activity in the ocean, and it is most commonly used in the experiments aiming at validating the models (see [12] and [3] for example), also because it is probably one of the most difficult parameter to model, as far as its behavior is quite hard to predict accurately. However, phytoplankton could not be directly measured during the OMEX campaigns ; it is in fact linked to the chlorophyll-a parameter; as far as it is always a major variable when it comes about understanding oceanic systems, it was also scrupulously studied during OMEX I, so that chlorophyll-a is a very well-informed parameter in the database.

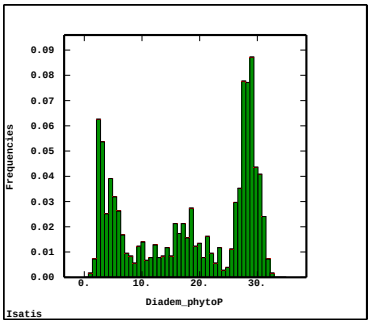
The very difficulty associated with the use of phytoplankton and chlorophyll-a in our study is that we have to establish a link between the two variables. Our first idea was to transform the phytoplankton data from the DIADEM model into chlorophyll-a. This transformation had already been made in [12] in order to assimilate chlorophyll-a data from satellite images into the biochemical model. It is based on an empirical function modeling the carbon to chlorophyll-a ratio (by weight), extracted from [13], and given below (1.1) :

$$\frac{C}{Chl_a} = \rho_{max}\frac{Chl_a}{Chl_a + K_{1/2}} \qquad \text{where } \rho_{max} = 90 \qquad \text{and } K_{1/2} = 0.477 \tag{1.1}$$

However this equation was initially made to convert phytoplankton carbon rate into chlorophyll-a, and not the inverse, and furthermore presents the disadvantage of concentrating the chlorophyll-a value in comparison of the carbon ones ; this can be seen on figure 1.4 : the distribution of the chlorophyll-a variable inferred from the phytoplankton data in the DIADEM model is far more concentrated than the distribution of the initial phytoplankton variable (that is to say phytoplankton expressed in mg C.m^{-3}). Observing figure 1.4(a), one could even think that there is a cut-off on the chlorophyll value at 0.7, whereas the observation of figure 1.4(b) shows that it is not the case. Therefore we shall finally study the phytoplankton in terms of mg C.m^{-3}, a unit which is proportionnal to the one used in the DIADEM model. The DIADEM model outputs, given in mmol N-m^{-3}, are then simply transform using a Redfield



(a) Histogram of the DIADEM chlorophyll-a deduced from the phytoplankton



(b) Histogram of the DIADEM phytoplankton (in mg C.m^{-3})

Figure 1.4: Comparison between the DIADEM chlorophyll-a and phytoplankton distribution

carbon to nitrogen ratio and the molar mass of Carbon, according to the following equations (1.2, 1.3, and 1.4).

$$\frac{C_n[mmol.m^{-3}]}{N_n[mmol.m^{-3}]} = \rho_{CN} = 6.625 \tag{1.2}$$

$$C_n[mmol.m^{-3}] = \frac{C_m[mg.m^{-3}]}{M_C[mg.mmol^{-1}]} \qquad \text{where } M = 12,011 \quad g.mol^{-1} \quad (\text{or } mg.mmol^{-1}) \tag{1.3}$$

$$\text{therefore} \quad C_m[mg.m^{-3}] = M_C[mg.mmol^{-1}].\rho_{CN}.N_n[mmol.m^{-3}] = 79.57N_n[mmol.m^{-3}] \tag{1.4}$$

The OMEX chlorophyll-a data were transformed using directly equation 1.1, which is valid for C and Chl_a given in mg.m^{-3}.

The nitrate variable was chosen because it is one of the biochemical variable that is defined on the total depth of the ocean, unlike the chlorophyll-a that is only present in the upper layers. Moreover, it is a direct output from the DIADEM model, and was pretty well surveyed by the OMEX cruises. Finally, it is also an interesting variable as far as it is directly linked to the phytoplankton compartment in the DIADEM model (and, hopefully, in reality!) : the phytoplankton growth rate is limited by the amount of nutrients, among which nitrate,available in the ocean ; reversely, the presence of phytoplankton is the only sink in the nitrate model.

The temperature is the variable extracted from the physical part of the DIADEM model. It is therefore supposed to be more reliable than the chlorophyll-a variable, whose calculation is not as easy. Besides, it is one of the "best" variable in the OMEX dataset, as far as almost each measurement done on a water sample is accompanied by a temperature measurement.

1.3.2 Temporal and spatial repartition of the data

Having defined the variable on which we shall focus, we have now to check if they are defined on the same area and for the same period of time.

Concerning the temporal definition of our data, the values in both dataset are supposed to be "punctual", that is to say that they correspond to a precise moment defined in time, which enables us to draw a comparison between them. However, time arise here two main problems. We have first to find the most appropriate period for the comparison, when both dataset are defined. As we already said, the OMEX I Project ran from 1993 to the beginning of 1996, whereas the DIADEM model was started in 1985 ; the physical model was spun up till 1995, and the biochemical model spin up was performed for years 1995,

1996 and 1997 ; data assimilation was only made from year 1998 on. It is then obvious that we shall not be able to work on the assimilated model. Nevertheless, we would like to work with a DIADEM model as stable as possible, and with as many OMEX data as possible. Therefore we shall choose the end of year 1995 ; we will focused on August, September and October 1995, where the number of OMEX data is consequent enough to perform geostatistical calculations. The second problem comes from the very distribution of both dataset in time. The OMEX dataset results of cruise measurements. Considering one variable among the three defined above, it means that for a precise instant (for example Sept. 1^{st}, 1995, 8:00 am), we have one and only one OMEX value at a precise point. We will then have ten minutes later a new value at the same place, but 5 m deeper, and one hour later a new value at the same depth, but a few kilometers further on the oceanographic ship road. The DIADEM results are defined for precise days, and are available every 60 days in 1995. In our case day, we shall focus on day 240, *i.e.* August 29^{th}, 1995. To be able to draw any comparison between the 2 datasets, we shall then assume that the OMEX data we selected, *i.e.* from August to October 1995, are equivalent to a set of measurements made at the same points in the same time during day 240. This assumption is of course not so clearly acceptable (it overcomes, for example, all the daily variations that can be observed in the oceanic system), but it is the only way to get a proper dataset from OMEX ; that should be kept in mind for future interpretation of the difference between the DIADEM results and the selected OMEX data.

As far as the spatial area of interest is concerned, the choice was mainly dictated by the OMEX dataset. Indeed this dataset is defined on a very precise area in the Celtic Sea, whereas the DIADEM model has been run on the whole North Atlantic. We shall then focused on the Goban Spur area surveyed by the OMEX cruises.The points retained for both the OMEX and the DIADEM datasets are plotted on figure 1.5.

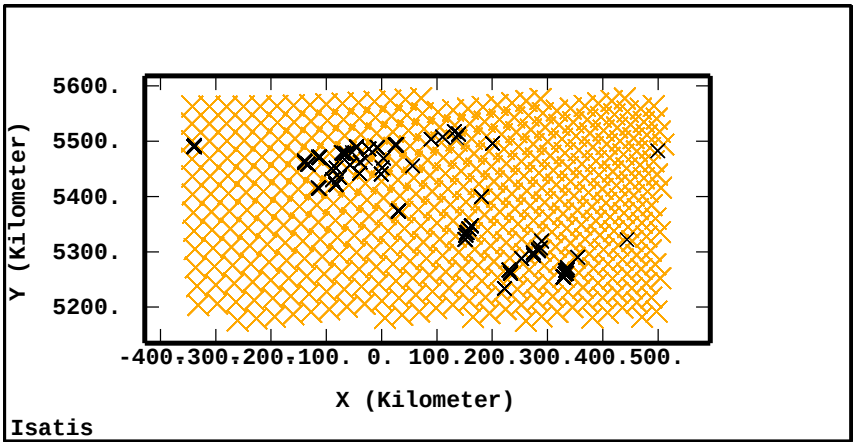


Figure 1.5: Basemap of the OMEX and DIADEM data used in our study. The DIADEM points are plotted in orange, the OMEX points are plotted in black

The differences in the spatial distribution of both datasets can be noticed on figure 1.5. The DIADEM points are located on a grid that is pseudo-regular in the horizontal directions, with approximatively 30x30 km cells, whereas the OMEX points are irregularly distributed on the area of interest, with a very well informed zone in the center (only 2 or 3 km between some points), and no data in the periphery. In the vertical direction, the difference is even worse : whereas the DIADEM grid is made of 18 layers with changing thickness when moving laterally, the OMEX points are defined on a very fine scale, but only at the location where the cruise ships stopped to make measurements. The main characteristics of both our datasets are sum up in table 1.1. To be able to compare the DIADEM results with the measurements made during the OMEX campaigns, we shall then find a way to relate the spatial distribution of the OMEX data points to the DIADEM grid, while in the same time conserving the main statistical characteristics of the variable studied (first and second order moment, histogram, spatial distribution ...). This is the aim of the geostatistical model : we shall performed Averaged Geostatistical Simulation (AGS) on the

OMEX data, a method that will be exposed in Chapter 2.

1.3.3 Identifying homogeneous populations in the OMEX dataset

In order to perform any geostatistical analysis, we have to ensure that the data we use are homogeneous. Indeed geostatistics consider a dataset as one realization (called the regionalized variable z) at some points of the space of a random field Z, with a given spatial distribution, and given statistical properties such as expected value, variance, covariance, etc. Therefore, the ensemble of data must be homogeneous enough to be taken as representative of one and only one particular spatial distribution. It can then be very sensible to separate a given dataset into several sub-sets which obviously don't follow the same spatial distribution. However, as far as the geostatistical investigation is based on the geographical properties of the data, this separation must also correspond to a geographical splitting of the data.

A first analysis of the OMEX dataset, using histograms, reveals that for each variable we can distinguished different groups (or populations) of homogeneous data that we shall separate to perform a consistent geostatistical work.

As far as the phytoplankton is concerned, we can identify on the histogram two distinct modes (*i.e.* two values possessed by a great number of points in the dataset). Plotting the points belonging to one or the other mode on a vertical basemap of the dataset, it clearly appears that the difference in the values of the data is linked to the depth of the data points : the surface point corresponds to the highest values of phytoplankton, whereas in the deepest part of the ocean the phytoplankton concentrations are smaller. We can then define two different population in the phytoplankton dataset:

- the points above 48.5 m depth, with great phytoplankton values
- the points between 48.5 and 105 m, with small phytoplankton values

Under 105 m, no chlorophyll-a measurements were carried by the OMEX surveys, simply because there is no more phytoplankton at such depth, so that we shall perform the analysis only for the 2 populations mentioned above.

As far as the temperature variable is concerned, we can observed three modes on the histogram, and once more they are linked to the depth; we shall then defined the three following populations:

- the points above 70 m depth, with warm temperature
- the points between 70 and 1300 m, with colder temperature
- the points below 1300 m

However, below 1300 m there are not enough data points to expect to perform any kind of geostatistical analysis, so that we shall forget about the third population further on in the study.

As far as the nitrate variable is concerned, there is not such a depth distinction as for the previous variables. However, we shall distinguished the shallower part of the zone, *i.e.* the continental shelf, from the deeper part, corresponding to the oceanic plain. Indeed, considering the mechanism of the renewing of nitrate in the ocean, depending on the particle fluxes on the oceanic margin, we cannot expect to have a similar distribution of our variable on these two parts of the studied area. Moreover, as far as these phenomena are mainly occurring in the vertical direction, there is an obvious change of scale in their physical or geostatistical description on the continental shelf and above the oceanic plain. This is why we shall perform distinct analysis on the shallow part of the zone and on the deep part of the zone.

We shall then try and find the corresponding zones in the DIADEM model. This part will be developed in Chapter 3.However, as far as the study is only aiming at establishing the method to perform a

sensible comparison between numerical models and *in situ* measurements, and as it is limited in time, we shall perform a complete analysis only on the phytoplankton variable, and focus on the upper part of the ocean for the temperature variable, and on the continental shelf for the nitrate variable. Nonetheless, we would like to perform quick statistics on both DIADEM and OMEX dataset in order to remind their main characteristic ; we have then to be aware of the fact that the phytoplankton variable is not present on the total oceanic depth. This biological phenomena is well traduce by the DIADEM model too, so that all the values of phytoplankton above 175 m depth are equal to zero. In order to get comparable statistics from both our dataset, we shall then take only into account, for the phytoplankton variable, the points above 175 m. Besides, we know that the OMEX dataset has not been defined under 105 m, but it does not necessarily mean that all the values under where null, so that wee also computed statistics on the DIADEM dataset only for points above 105 m. These statistics are summed up in table 1.1 ; concerning the phytoplankton, the first value correspond to the statistics above 175 m, whereas the value into brackets corresponds to the statistics above 105 m.

	DIADEM data			OMEX data		
Variables	PhytoP.	Nitrate	Temp.	PhytoP.	Nitrate	Temp.
Spatial distribution	• Pseudo-regular grid in the horizontal direction, with approximatively 30x30 km cells. • 18 vertical layers : - iso-potential density layering - the layer thickness vary a lot when you move horizontally, from 20 to 1650 m thick.			• Irregular horizontal distribution, following the oceanographic ships road : - very dense distribution in the central area (about one data point every 5 km) - fewer data points in the periphery. • Regular distribution in the vertical direction (one value every 5 m), but only along the measurements profiles.		
Spatial distribution	• Output from day 240 in 1995 (*i.e.* August $29_t h_i$), 1995. • The physical model's spin up has been totally performed (*cf* Temp.). • The biological model is spinning up (*cf* PhytoP. and Nitrate). • No data have been assimilated yet.			• Measurements carried out during August, September, and the beginning of October 1995. • We shall assume that they are equivalent to a dataset for day 240 in 1995.		
Statistics : Nb. of points	3969(2694)	7452	7452	302	750	891
Min.	0.18(1.24)	0	-1.2	0	0	0
Max.	32.77	22.84	19.2	90	23.7	20.04
Mean	13.04(18.7)	8.87	10.13	16.15	8.38	11.81
Standard Dev.	11.58(9.90)	9.03	5.65	16.72	6.79	4.25

Table 1.1: Main characteristics of the DIADEM and OMEX dataset in our study

Chapter 2

Averaged Geostatistical Simulations

We would like to be able to compare easily the OMEX *in situ* data to the DIADEM model's results. To do so, it would be very useful to define a value for the OMEX dataset at each node of the DIADEM grid ; this will be achieved by using geostatistical models. Moreover, we want to check the model ability to reproduce reality, and correctly reflect the spatial and statistical variablity of reality on the model support. This definition of our needs reveals in fact two different requirements that we shall try to fulfill. First, we would like to know how reality looks like on the DIADEM grid ; the most appropriate tool to obtain this kind of results is the simulation. Then we would like to take into account the support of the DIADEM data in our study, *i.e.* the fact that the results of the model are given for cells with a volume of about 30 x 30 km x 100 m, knowing that this volume can vary a lot with the layer thickness. This is why we shall use Averaged Geostatistical Simulation (AGS), that is to say the average of simulations that will reproduce the variability observed on the OMEX dataset on a regular grid, on each of the DIADEM grid cells. This method and its statistical and geostatistical properties are exposed thereafter, as well as the practical study of the OMEX phytoplankton, nitrate, and temperature.(For more information about geostatistical methods, please refer to [10], [5], [14] or [6].)

2.1 Theoretical Presentation of the method

2.1.1 Geostatistical simulations

General properties of the geostatistical simulations

The first step of the AGS method is to model the variable on a regular grid, using the *in situ* data, and trying to respect as much as possible their statistical and geostatistical properties. This is done using geostatistical simulations.

Considering a random function $Z(x)$, a *simulation* $Z_s(x)$ of Z is by definition a random function which admit the same spatial distribution as Z. This is a very strong property, and in particular Z_s has the same expectation, variance, covariance, histogram and monovariate distribution as Z. Yet all these properties combined together, although necessary, are not sufficient for the two random functions to have the same spatial law. In geostatistics, what is considered as a random function Z is the variable of interest. We consider then that "the reality", called in geostatistics regionalized variable z, corresponds to a particular realization of Z, corresponding to a particular event ω ($z = Z(\omega)$), and that we know the value of this realization on a few points $\{x_\alpha, \alpha = 1 \dots n\}$ corresponding to the data points. A simulation of the regionalized variable z will then be a realization z_s of the simulation Z_s. As far as we want our simulation to resemble as much as possible the regionalized variable, *i.e.* to the data, we would like it to be equal to the data values at each data point. This bring us to define the *conditional simulation* Z_c

: considering the data $Z_{\alpha_1}, \ldots, Z_{\alpha_n}$, Z_c is a conditional simulation of Z his distribution is the spatial distribution of Z conditionally to $Z_{\alpha_1}, \ldots, Z_{\alpha_n}$. A realization z_c of Z_c is then a conditional simulation of the regionalized variable z, and at each experimental point we have $z_c(x_\alpha) = z(x_\alpha)$. Besides, this definition of a conditional simulation imposes far more than just the respect of the *in situ* value at the data points. In particular it insures that the connection with the data value around the data points does not create any artefact.

Whereas the regionalized variable z is only known at the data points, we would like to be able to build up simulations in any points of the space, in particular on a regular grid. This would be possible if we knew completely the spatial distribution of the random function Z. Yet the only information we have on Z are the data points, which are limited in number and often irregularly distributed in space. Those data only allows us to infer the bivariate spatial distribution of Z, *i.e.* the distribution of the couple (Z(x), Z(x+h)). And even this distribution can be inferred for only a few values of h. We shall then admit that Z_s (respectively Z_c) is a simulation (respectively conditional simulation) of Z if Z_s (respectively Z_c) and Z have the same bivariate spatial distribution (respectively the same bivariate distribution conditionally to the data).

It is important here to underline the differences between simulations and an estimation of the variable z ; this will explain why we shall use the ones rather than the other. The estimation technique of a regionalized variable usually used in geostatistics is the kriging technique. Let us remind here briefly how it is implement in the case of a stationary random function with a known mean (this is usually the case, as far as we assume that the mean of the variable values at the data points is the mean of the random variable) ; we will consider that we know Z in n experimental points x_α, and that the mean is $m = 0$ (if not, we just consider $Z - m$). The kriging estimator $Z^*(x)$ of Z at some point x is then built as a linear combination of the data points values :

$$Z_x^* = Z^*(x) = \sum_{\alpha=1}^{n} \lambda_x^\alpha Z_\alpha \qquad \text{where } Z_\alpha = Z(x_\alpha) \tag{2.1}$$

We shall minimize the variance of the estimation error:

$$\begin{aligned} var(Z_x - Z_x^*) &= var(Z_x - \sum_{\alpha=1}^{n} \lambda_x^\alpha Z_\alpha) \\ &= var(Z_x) + \sum_{\alpha,\beta=1}^{n} \lambda_x^\alpha \lambda_x^\beta C_{\alpha\beta} - 2\sum_{\alpha=1}^{n} \lambda_x^\alpha C_{\alpha x} \qquad \text{where } C_{\alpha\beta} = cov(Z_\alpha, Z_\beta) \end{aligned}$$

This draws us to write the following system :

$$\begin{cases} \sum_{\beta=1}^{n} \lambda_x^\alpha C(x_\alpha - x_\beta) = C(x_\alpha - x) \qquad \forall \alpha \in \{1 \ldots n\} \end{cases} \tag{2.2}$$

This system always have solutions for $\{\lambda_x^\alpha, \alpha = 1 \ldots n\}$, and this solution is unique as far as the covariance matrix $C_{\alpha\beta}$ is regular ; this solution gives us the value of the estimator at point x. The kriging technique produces an exact interpolator, *i.e.* $Z_{x_\alpha}^* = Z_{x_\alpha}$ at each data point ; furthermore, the estimator is unbiased ($E(Z_x^* - Z) = 0$). However, this estimator presents two main "defects". First, the variance of Z^* is smaller than the variance of Z. Indeed we have :

$$Z_x = Z_x^* + (Z_x - Z_x^*) \tag{2.3}$$

Moreover, at any data point α,

$$\begin{aligned} cov(Z_\alpha, Z_x - Z_x^*) &= cov(Z_\alpha, Z_x - \sum_{\beta=1}^{n} \lambda_x^\beta Z_\beta) \\ &= C_{\alpha x} - \sum_{\beta=1}^{n} \lambda_x^\beta C_{\alpha\beta} \\ &= 0 \quad \text{according to the kriging sytem 2.2} \end{aligned}$$

Therefore the kriging error is not correlated to the data Z_α, and then is not correlated to any linear combination of the data, in particular *the kriging error $Z_x - Z_x^*$ is not correlated to the kriging Z_x^**. Then from (2.3) we gather :

$$var(Z_x^*) = var(Z_x) - var(Z_x - Z_x^*) \leq var(Z_x) \tag{2.4}$$

Therefore the result of the kriging is smoother than the original variable ; kriging is not the relevant tool to get a picture how reality should vary in space. Moreover, Z_x^* doesn't admit the same covariance as Z_x ; the spatial distribution of Z^* is then not the same as the one of Z. This is why we shall focus on the simulations techniques : we want indeed to get a picture of how reality may look like, to check if the DIADEM model is able to reproduce this allure. However, we shall be aware of the fact that unlike kriging, which is an estimator that give one and only one result at a given point x, with a value that one can consider as representative of reality at that point, the results of simulations are not supposed to be values very close to reality at each point (except the data points if we consider conditional simulations). Indeed, it is very possible to get several simulations of the same variable at the same points (this is simply done by generating several independent realization of the random function Z_s (or Z_c)), and the different simulations can give pretty different values for the same points, especially if we are located far away from the conditioning data points. This is also why we shall perform several simulation before drawing any comparison between the results of our geostatistical simulations and the results of the DIADEM model, in order to be able to be sure that any observed difference is not linked to a particular drawing, but to the allure reality should have in any case. One could also propose to take an average of several simulations to avoid this kind of problem ; this is absolutely not the thing to do, because by averaging different realizations of the random function Z_s (or Z_c), one will just lose this very variability reflecting reality that is one of the main property of simulations ; indeed, the mean of a great number of conditional simulations calculated from a certain dataset tends to the kriging of the variable using the same dataset.

Generating geostatistical simulation

Our aim in this section is to obtain one or several realization of a conditional simulation of a random function Z on a set of points, usually regularly distributed on a grid (called further on "the simulation grid", even if it is not necessarily a regular grid). We shall assume that Z is a stationary random function, whose expectation is known, and supposed to be $m = 0$ (otherwise, as usual, we can just perform the same calculation using $Z - m$). A stationary random function, in geostatistics, usually means a second order stationary random function, *i.e.* $\forall x, \forall h, Cov(Z(x), Z(x++h))$ is defined, and does not depend on the point of support x. We can then defined the covariance function K as $K(h) = Cov(Z(x), Z(x+h))$ for any x.

Let us first assume that we are able to perform a non-conditional simulation of Z, called S, and that S and Z are independent. Then we can obtain a realization of S at any point of the simulation grid and at each data point. Besides, by definition, we have a realization of Z at each data points. How can we perform a conditional simulation of S ?

Considering any point x on the simulation grid, we can then perform two different kriging : the kriging Z^* of Z at point x using the data values, and the kriging S^* of S at point x, using the simulated values of S on the data points. These two kriging write :

$$Z(x) = Z^*(x) + (Z(x) - Z^*(x)) \tag{2.5}$$
$$S(x) = S^*(x) + (S(x) - S^*(x)) \tag{2.6}$$

Let us define a new random function :

$$W(x) = Z^*(x) + (S(x) - S^*(x)) \tag{2.7}$$

A realization of W, corresponding to the realization of S we have calculated, is then known at each point of the data grid. What are the statistical properties of W, in particular its expectation and covariance ?

$$\begin{aligned} E(W(x)) &= E(Z(x)) + E(S(x) - S^*(x)) \\ &= 0 + 0 = 0 \qquad (2.8) \\ K_W(x, x+h) &= cov(W(x), W(x+h)) \\ &= K^*(x, x+h) + K_{\epsilon_S}(x, x+h) \qquad (2.9) \end{aligned}$$

because S and Z are independent therefore S and Z_α are and S and Z^* are

$$\begin{aligned} \text{Besides} \quad K(h) &= cov(Z(x), Z(x+h)) \\ &= cov(Z^*(x) + (Z^*(x) - Z(x)), Z^*(x+h) + (Z^*(x+h) - Z(x+h)) \\ K(h) &= K^*(x, x+h) + K_\epsilon(x, x+h) \qquad (2.10) \end{aligned}$$

because the kriging and the kriging error are independent

$$\text{In the same way} \quad K_S(h) = K^*_S(x, x+h) + K_{\epsilon_S}(x, x+h) \qquad (2.11)$$

Moreover Z and S have the same covariance according to the definition of a simulation, and the data points considered in the krigings 2.5 and 2.6 are located at the same place, so that the kriging weights $\{\lambda_i, i = 1 \dots n\}$ defined by 2.2 are the same for $Z^*(x)$ and $S^*(x)$. We can then write:

$$\begin{aligned} K(h) &= K_S(h) \qquad (2.12) \\ \text{and } K^*(x, x+h) &= K^*_S(x, x+h) \qquad (2.13) \end{aligned}$$

From 2.9, 2.10, 2.11, 2.12, 2.13, we gather :

$$K_{\epsilon_S}(h) = K_\epsilon(h) \qquad (2.14)$$

$$\begin{aligned} K_W(x, x+h) &= K^*(x, x+h) + K_\epsilon(x, x+h) = K(x, x+h) \\ K_W(h) &= K(h) \qquad (2.15) \end{aligned}$$

Besides, at each data point we can write :

$$W(x_\alpha) = Z(x_\alpha) + 0 = Z(x_\alpha) \qquad (2.16)$$

Let us assume now that Z is a Gaussian random function, *i.e.* :

$$\forall n \in \mathbf{N}, \forall x_1, \dots, x_n, \forall l_1, \dots, l_n \qquad \sum_{i=1}^n l_i Z(x_i) \quad \text{is gaussian}$$

Then S is also Gaussian, and $(S - S^*)$ too ($\forall x$, $S^*(x)$ is a linear combination of $S(x_\alpha)$). In the same way, Z^* is a Gaussian random function. Therefore W, which is the sum of two independent Gaussian random functions is itself a Gaussian random function. Having the same expectation and covariance as Z (see 2.9 and 2.15), *S has the same spatial distribution as Z*. Moreover, we have seen (equation 2.16) that *S coincides with Z at each data point*. We have then managed to build a conditional simulation of a Gaussian random function from a non conditional simulation.

Now we shall find a way to build a non conditional simulation of our Gaussian random function.

Building a non-conditional simulation of a given Gaussian random function (ie with a given covariance $K(h)$ in the stationary case, and as far as we consider the restricted definition of the simulation, see above) in one dimension can be done using various methods : the spectral method, the dilution method, the migration method, etc. It will be too long to explained them in details here ; for more information please refer to [2] and [10]. The choice of the method, theoretically free, is usually made in order to optimize the generation of this or that specific model of covariance.

Generating a 3-dimensional simulation from a 1-dimensional one can be done by using the *Turning Bands Method*. Considering the 1-dimensional random function Y(x), with covariance K(h), simulated on a line, we can define in the 3-dimensional space the function (see also figure 2.1) :

$$\begin{aligned} Z(u,v,w) = Z(P) &= Y(x) \\ &= Y(\overrightarrow{OP}.\overrightarrow{s}) = Y(< P.s >) \end{aligned} \tag{2.17}$$

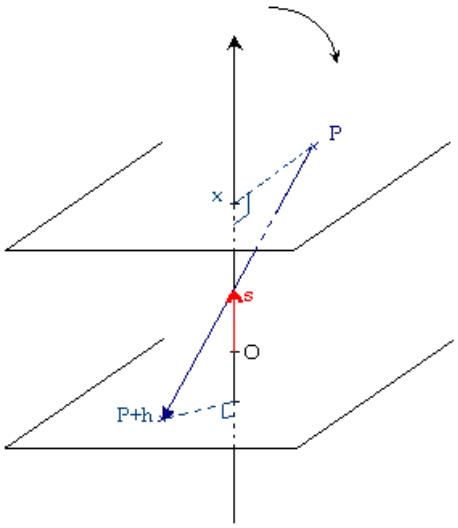


Figure 2.1: The Turning Bands Methods

From 2.17 we gather :

$$\begin{aligned} E(Z(P).Z(P+h)) &= E(Y(< P.s > .Y < (P+h).s >) \\ &= K(< (P+h).s > - < P.s >) \\ &= K(< h.s >) \end{aligned} \tag{2.18}$$

However, according to 2.18, such a function Z is anisotropic, and depends on the choice of the orienting vector $\overrightarrow{s}$. To get rid of this anisotropy, let us consider a random vector $\overrightarrow{S}$ uniformly distributed on the unity sphere in $\mathbf{R}^3$. Then we obtain :

$$\begin{aligned} Z_S &= Y(< P.S >) \\ E(Z_S(P).Z_S(P+h) &= E_S(E(Z_S(P).Z_S(P+h))|S=s) \\ &= E(K(< h.S >)) \end{aligned}$$

As far as the covariance K is an even function, s only need to browse half of the unity sphere. We can then calculate the covariance $C_3(h)$ of our 3-dimensional random function Z :

$$\begin{aligned} C_3(h) &= \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_0^{\pi/2} K(h\sin\phi)\cos\phi\, d\phi \\ \text{writing} \quad u = h\sin\phi \quad &\text{we get} \quad du = h\cos\phi\, d\phi \quad \text{and} \\ C_3(h) &= \frac{1}{h}\int_0^h K(u)du \\ i.e. \quad K(h) &= \frac{d}{dh}hC_3(h) \end{aligned} \tag{2.19}$$

Using this method, we can then perform a 3-dimensional non-conditional simulation of a Gaussian random function thanks to the 1-dimensional methods mentioned above, and further condition it to the measurement values at the data points. The last step to perform is to obtain geostatistical simulations of a random function Z with any distribution. This is made using an *anamorphosis* function Φ, which transforms a Gaussian variable Y in a new variable Z with any distribution, called raw variable. Having performed a conditional simulation of the Gaussian random function Y, we can get a conditional simulation of $Z = \Phi(Y)$; as far as the anamorphosis function is reversible, the conditioning on the Gaussian variable from data values inferred from the data thanks to Φ results in the correct conditioning of the raw variable Z.

2.1.2 Averaged Geostatistical Simulation

From our punctual geostatistical simulations, we would like to get one value per cells of the DIADEM grid, in order to draw a proper comparison between the numeric model's results and the *in situ* OMEX data. One solution would be to perform the geostatistical simulations directly on the DIADEM grid, as we have seen that the simulation grid does not need to be regular. However, this is not a satisfactory solution. Indeed, the results of the geostatistical simulations reproduce the reality variability inferred from the data at some chosen points of the simulation grid; it means that we will get at each node of the DIADEM grid a value corresponding to what could be the concentration of phytoplankton or nitrate in a few liters of oceanic water collected at that precise point, or the sea temperature at that precise point ; besides, these simulated value will have a statistical distribution, and in particular a variance, corresponding to the one calculated on the data points, *i.e.* on points very close to each other. Yet it is proven [14, 6] that the distribution of a regionalized variable depends on the support on which it is defined (*i.e.* in our case the volume which the data is representative of); it can be clearly seen on the figure 2.2: the distribution on a small support (for example the OMEX data, or a punctual simulation of these data) is larger than the distribution on a block support (*e.g.* the DIADEM model grid).

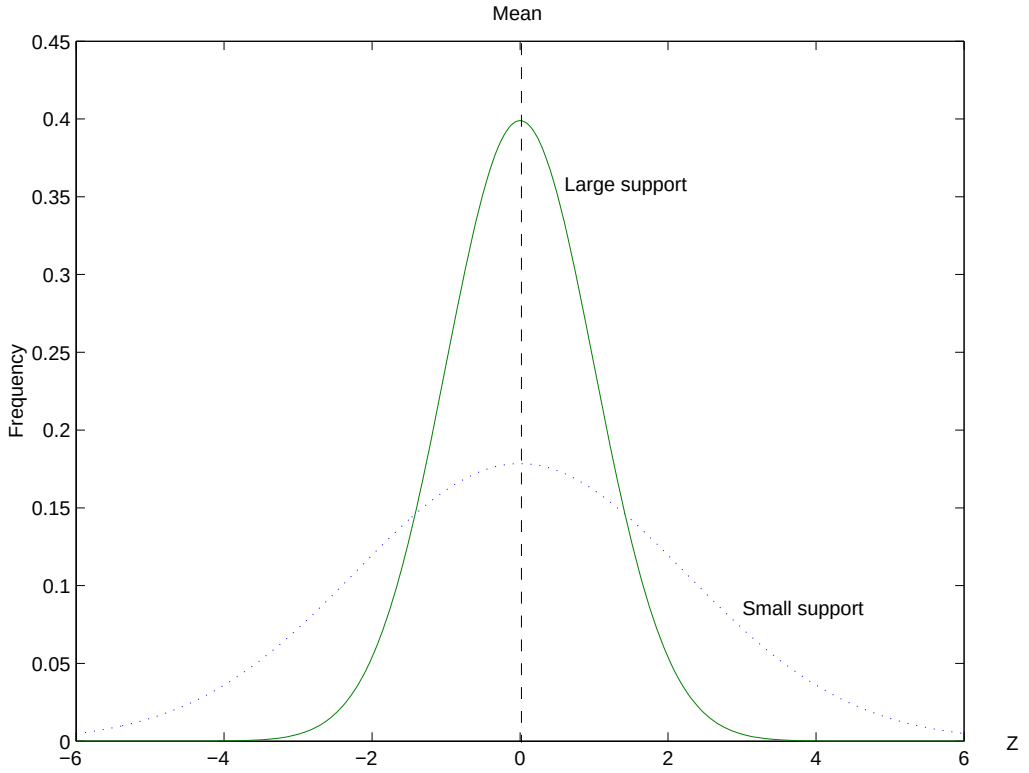


Figure 2.2: Distribution of a regionalized variable on a large and on a small support

This can be simply understood : if consider values representative of large domains, you won't expect them to present the extrema that can be found at some precise points of space ; therefore their distribution should be more concentrated around the mean value than in the case of values taken on smaller domains. Geostatistics explained this by what is called the Krige's relation (2.20): let us consider, on the domain of study D of a regionalized variable, two nested supports: D is divided into large volumes V, themselves divided into smaller volumes v (see figure 2.3). Then, denoting $\sigma^2(v|V)$ the dispersion variance of a small volume v in a larger volume V, we can write :

$$\sigma^2(v|D) = \sigma^2(v|V) + \sigma^2(V|D) \tag{2.20}$$

This means that the variance calculated on a large support (corresponding to the large volumes V) on

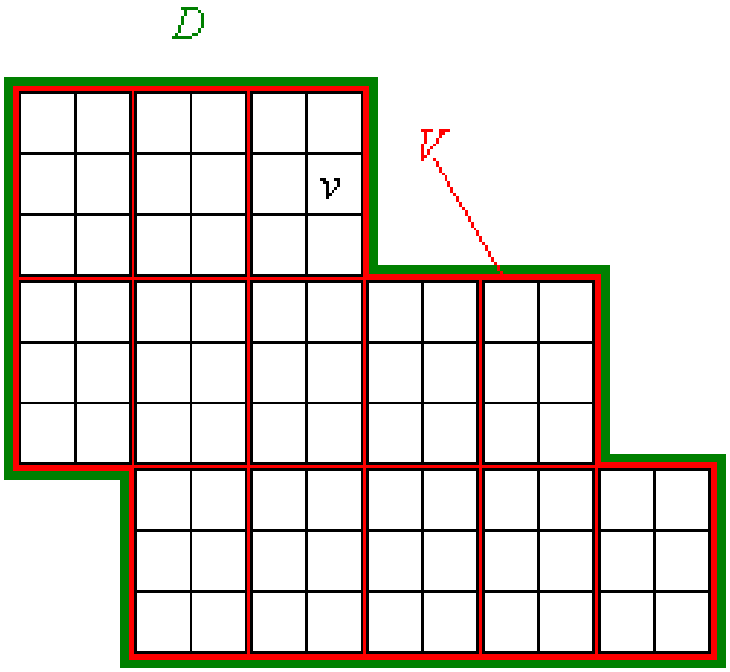


Figure 2.3: The domain D is divided into large volumes V, themselves partitioned in smaller volumes v

the studied domain D is smaller than the variance of a smaller support (corresponding to the smaller volumes v), because the variance calculated on the "v values" in a V volume is usually not null, and not negligible.

We have therefore to find a way to get representative values of the OMEX data on the DIADEM grid, starting from the results of geostatistical simulations calculated on a fine regular grid. As far as the phytoplankton and nitrate variables are concerned, the model output is given in terms of concentration on the grid cell; therefore if we obtain values of phytoplankton and nitrate concentration on a regular grid (each cell having the same volume), the best way to get the corresponding concentration on the DIADEM grid would be to average all the values located in each DIADEM cell. As far as the temperature is concerned, it is also sensible to assume that a representative value of the temperature on the whole DIADEM cell will be the average temperature on an ensemble of points regularly distributed in the cell. We can then defined the Averaged Geostatistical Simulations (AGS) of the OMEX data on the DIADEM grid : it consists in averaging the results of punctual simulations on a fine regular grid on the DIADEM model grid, *i.e.* for one given cell of the DIADEM grid taking the average of all the values at the nodes of the simulation grid located in this very cell.

2.2 Practical geostatistical study of the OMEX data

The practical study of the OMEX phytoplankton, temperature, and nitrate variables was led following four main steps : we have first to calculate the anamorphosis function which will transform our raw variables into Gaussian ones, and reversely ; then we have to perform the variographic analysis of the "gaussianized" data ; the third step consists in generating punctual geostatistical conditional simulations using the Turning Bands method, and transforming the Gaussian variables back into raw variables ; and finally we have to average the simulations on the DIADEM grid. The 3 first steps were computed thanks to the geostatistical software ISATIS, developed by Geovariances in collaboration with the Center of Geostatistics at the École des Mines de Paris. The last step, as it is a new and original way of dealing with support problems in the case of irregular grids such as the DIADEM one, was done using more usual data manipulation tools.

However this analysis can only be performed on homogeneous data, following the same spatial distribution. It was then performed separately on each populations defined on our variables in Chapter one:

- As far as the Phytoplankton variable is concerned :
  - Above 48,5 m depth
  - Between 48,5 and 105 m depth
- As far as the Nitrate variable is concerned :
  - On the continental shelf
- As far as the temperature variable is concerned :
  - Above 70 m depth

2.2.1 Calculating the anamorphosis function

We have seen (section 2.1.1, page 19) that the conditional geostatistical simulations can only be performed on a Gaussian variable, and that it is easy to obtain afterwards simulations of a raw variable using a anamorphosis function. The three variables on which our study focuses do not follow a Gaussian distribution. Therefore we shall use a anamorphosis for each of them, and in fact for each of the homogeneous group defined among them. Calculating these functions is the first step to perform in such an analysis ; indeed, as far as the anamorphosis function is supposed to be reversible, it will enable us to calculate at each data point the value of the Gaussian variables to be used in the simulation process. These "gaussianized" data (which we shall call further on the Gaussian data, phytoplankton, nitrate or temperature) are necessary to model the spatial bivariate distribution of the Gaussian variables, and later to condition the simulations.

The anamorphosis function can be written as a polynomial expansion using the Hermite Polynomials H_i:

$$\Phi(Y) = \sum_{i=0}^{+\infty} \Psi_i H_i(Y)$$

In practice, we can only model Φ with a finite number of polynomials in the sum ($n \leq 100$ in Isatis):

$$\widehat{\Phi}(Y) = \sum_{i=0}^{n} \psi_i H_i(Y)$$

We then try to fit this model to a discrete curve plotting Y in function of Z, and calculated from the data values.

However, the anamorphosis function, which is invertible, must be strictly increasing with Y ; it is not the case if the polynomial expansion stops at a given order ; therefore, in order to get a proper result while using $\widehat{\Phi}$,we must define the practical interval of definition of $\widehat{\Phi}$, which is the largest interval on which $\widehat{\Phi}$ is increasing with Y, and is delimited by the 2 points : $(Y_{p_{min}}, Z_{p_{min}})$ and $(Y_{p_{max}}, Z_{p_{max}})$ (see figure 2.4).

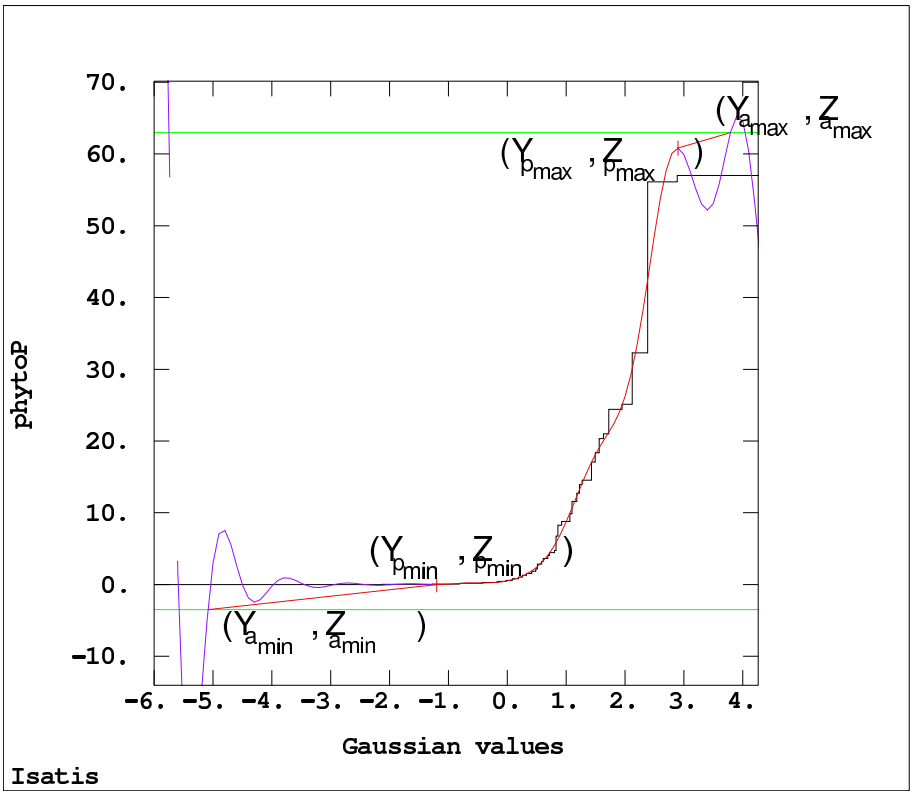


Figure 2.4: Calculation of the anamorphosis function : the experimental curve is plotted in black, the polynomial expansion $\widehat{\phi}$ is plotted in purple, and the final anamorphosis to be used is plotted in red

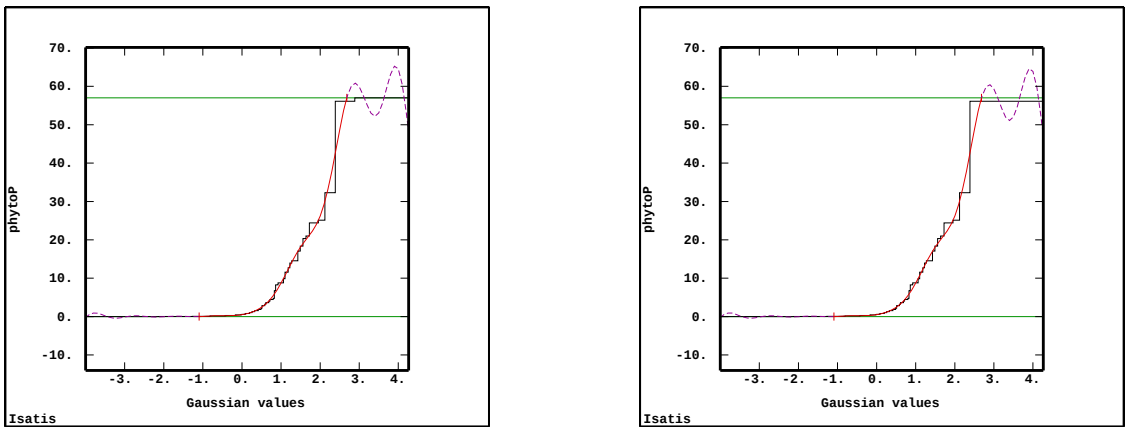
Furthermore, whereas a Gaussian variable is definite on whole $\mathbf{R}$, *i.e.* it can take any value between $+\infty$ and $-\infty$, our variables of interest are limited : all of them are positive, and we won't expect, for example, the ocean temperature to reach $50\,^{\circ}\mathrm{C}$ in the middle of the North Atlantic. Therefore we have to define an authorized interval on the raw variable, delimited by $Y_{a_{min}}, Y_{a_{max}}$. The intersection of $\widehat{\Phi}$ with the two horizontal lines defined by these limits, $(Y_{a_{min}}, Z_{a_{min}})$ and $(Y_{a_{max}}, Z_{a_{max}})$, are the bound of what is called the absolute interval of definition of Φ. It represent the interval in which the final modeled anamorphosis function will take its value.

If the absolute interval of definition is larger than the practical one, we cannot use the polynomial expansion on the extreme parts of it ; we will then defined the modeled anamorphosis by drawing a line between $(Y_{a_{min}}, Z_{a_{min}})$ and $(Y_{p_{min}}, Z_{p_{min}})$, and between $(Y_{a_{max}}, Z_{a_{max}})$ and $(Y_{p_{max}}, Z_{p_{max}})$.

Our anamorphosis is bound to be finally used with a random drawing of a Gaussian variable, in order to get a random drawing of the raw variable ; what will happen if we draw a value out of the absolute interval of definition ? If we simply decide to reject it, and draw another value, we take the risk of perturbing the Gaussian distribution of the drawing by systematically rejecting extreme value. A better solution in that case is to give the limit authorized value to the raw variable.

We then completely model the anamorphosis by the function plotted in red on figure 2.4. Nonetheless this calculation is only a model, and we shall remember the artefacts it might introduce for the analyzing part in Chapter 3 ; in particular, the treatment of great values drawn on the Gaussian variable can produce too many points with the extreme authorized value on the raw variable. The parameters on which the modeler can play to obtain the best modeled anamorphosis are the number of Hermite polynomials used in the expression of $\widehat{\phi}$ and the limit values of the raw variable.

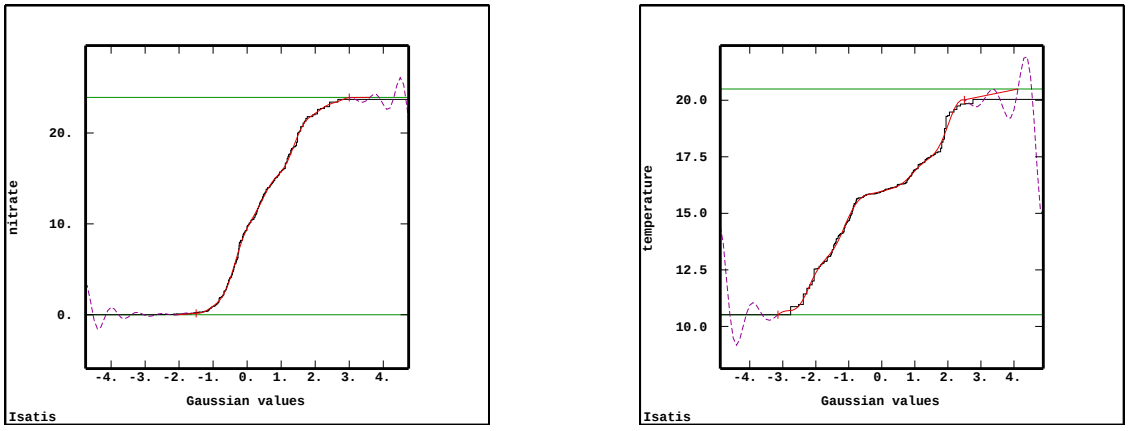
Figures 2.5, 2.6(a) and 2.6(b) shows the anamorphosis functions used in our study for the 3 variables : phytoplankton, temperature, and nitrate, and the corresponding authorized interval of definition for the raw variable. The calculation were of course performed separately on each of the populations of points reminded for each variable in the introduction of this chapter.



(a) Above 48.5 m

(b) Below 48.5 m

Figure 2.5: Anamorphosis functions used for the phytoplankton variable



(a) Nitrate on the continental shelf

(b) Temperature above 70 m

Figure 2.6: Anamorphosis used for the nitrate and temperature variables

Variographic modeling

In order to perform the simulations using the Turning Band method (see section 2.1.1), we must determine the covariance function K for each of our variables. This is done using the variogram function, defined for any vector h by :

$$\gamma(h) = \frac{1}{2}E((Z(x+h)-Z(x))^2) \tag{2.21}$$

Denoting, in the stationnary case, (2.22)

$$K(h) = cov(Z(x), Z(x+h))$$

we get :

$$\gamma(h) = \frac{1}{2}(E(Z(x+h)Z(x+h)) + E(Z(x)Z(x)) - 2E(Z(x)Z(x+h)))$$

$$= \frac{1}{2}(K(0)+K(0)-2K(h))$$

$$\gamma(h) = K(0) - K(h) \tag{2.23}$$

Knowing the variogram and the variance K(0) for a stationary random function Z is then equivalent to knowing its covariance.

In practice, the variance K(0) is simply calculated from the dataset using the usual estimator $\frac{1}{n}\sum_{\alpha=1}^{n}(z_x^\alpha - m)^2$ where $\{z_x^\alpha, \alpha = 1,\dots,n\}$ are the data value at the data points x_α and m is the mean of the data. The variogram is estimated for a given vector h by the experimental variogram $\widehat{\gamma}$:

$$\widehat{\gamma}(h) = \frac{1}{2p}\sum_{\alpha,\beta}(z_x^\alpha - z_x^\beta)^2 \tag{2.24}$$

where the couples (x_α, x_β) are chosen in such way that $x_\beta = x_\alpha + h$, and p is the number of data points verifying this very relation. It is quite obvious that there are vectors h for which we won't be able to find enough points x_α, x_β in our data set to perform a satisfactory calculation of the variogram.

The first solution to this problem is to consider that h is no more a vector, but only represent the distance between two points. We can then more easily calculate what is called the omnidirectional experimental variogram. In practice, we can only perform this calculation for some values of h. We shall choose these values to be regularly distributed ; the value between two consecutive h is called the lag, and the number of lag is limited by the extension of the studied domain. Moreover, to be sure to have a sufficient number of points to calculate the mean 2.24, we shall accept a certain tolerance ϵ on h, *i.e.* we shall take into account all points x_α, x_β such as $|x_\alpha - x_\beta| \in [h-\epsilon, h+\epsilon]$. Usually the tolerance is expressed in terms of proportion of the lag value. When the tolerance is equal to half of the lag value, the class of values accepted for h for two consecutive calculations are contiguous.

On another side, it can be very interesting to conserve the directional aspect of vector h, as far as in many cases the distribution of the variable depends on the direction you consider. In this case, you have to select several direction on which to perform the calculation using the previous lag method for $|h|$. Here again, we shall choose these directions to be regularly distributed in space. We shall also define a tolerance on the direction ; this tolerance is a given angle θ, and a couple of points x_α, x_β will be retained in the calculation if x_β lies in the cone with an opening equal to θ, centered on the line following the chosen direction and passing through x_α (see figure 2.7).

In the case of the OMEX dataset, the experimental variograms are calculated for the Gaussian phytoplankton, nitrate and temperature for the different defined populations. We choose to perform a directional calculation, as far as we cannot expect our variables to vary in the same way horizontally and vertically, and as we can even assume that, knowing the environment of the oceanic margin, the variation in two different horizontal direction can change a lot. We finally calculated the experimental variogram in the vertical direction, and in one or two horizontal direction. The calculation in only one horizontal

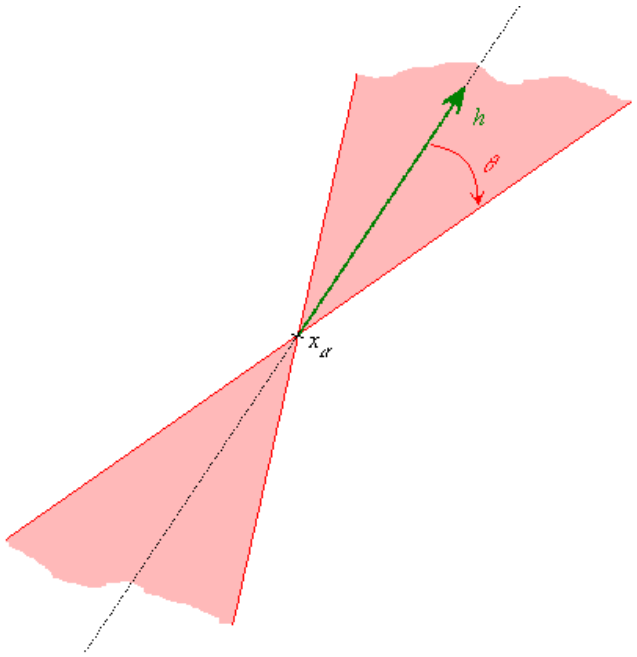


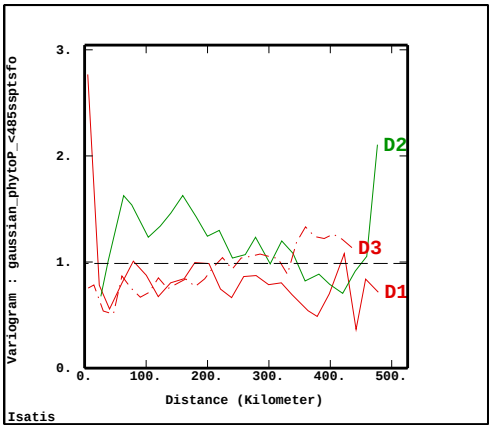
Figure 2.7: Representation of the tolerance angle θ when calculating an experimental variogram along the direction h

direction was performed with a $180\,^\circ$ tolerance angle, but a limitation on the slicing height, that is to say that the vertical distance between two points of a retained couple of points is limited, so that it is equivalent to an omnidirectional experimental variogram calculated in horizontal plans. To choose the two directions in which to perform calculation in the other case, we just tried several and find the two perpendicular ones for which the experimental variograms obtained were the most different. The results of these calculations are displayed on figure 2.8.

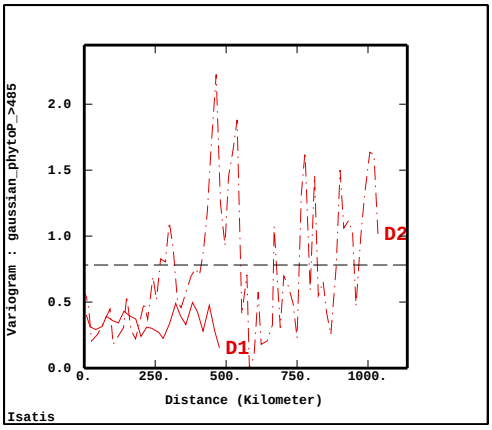
The experimental variogram provides us with a certain amount of points on which we have inferred the value of the variogram. This is however not sufficient to perform simulations, which require the knowing of the entire covariance function K. Therfore we shall fit a model on the experimental variogram. This variogram model must be as close as possible to what we have calculated from the data, but can also take into account qualitative information about the data.

For the phytoplankton variable above 48.5 m, we shall fit on the experimental variogram a spheric model, with a geometrical and a zonal anisotropy (see figure 2.9). The geometric anisotropy translates the fact that the phytoplankton vary in the same way along the vertical direction and along the horizontal direction making a $+120\,^\circ$ angle with the East-West line (mathematical definition of the angles), but that there is an obvious difference of scale between the two phenomena. The zonal anisotropy between the these two directions and the third one ($+30\,^\circ$ from the E-W line) translates the fact that the variability of phytoplankton is smaller in this direction.

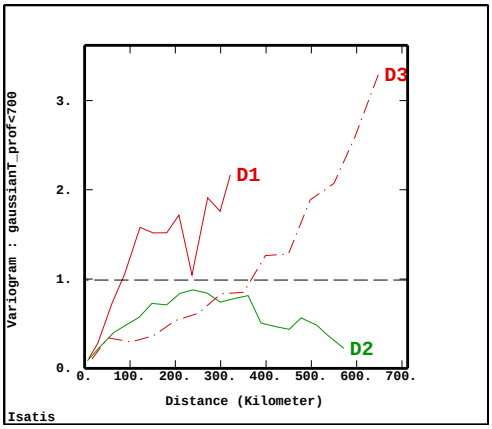
On the experimental variogram plotted for the temperature above 70 m, we have to combine several schemes to be able to fit a acceptable variogram model. This time, we shall choose to fit in the best possible way the experimental curves rather than to link the model to the physical properties of our variable. Therefore we shall combine a cubic model in the first horizontal direction, with two nested model in the second horizontal direction (a Gaussian and a J-Bessel model). In the vertical direction, we shall use first a Gaussian model for the small scale phenomena, nested with a cubic model for the larger scales. In this direction, it seems in fact that the variable is not stationary ; however, the simulations can only be performed on stationary random function, so that we shall plot here a cubic model with a very large scale, larger than the area of interest, so that it would not have reach it sill into the studied domain.



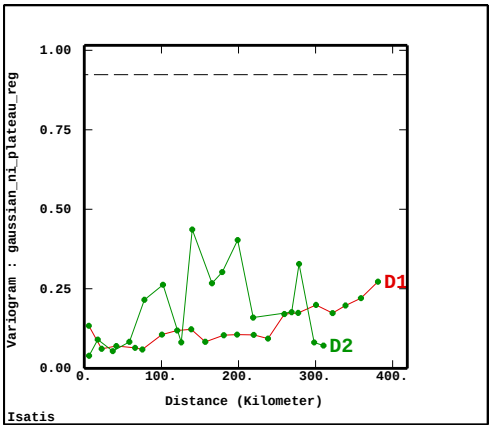
(a) Phytoplankton above 48,5 m ; global anisotropy : $+30\,^{\circ}$



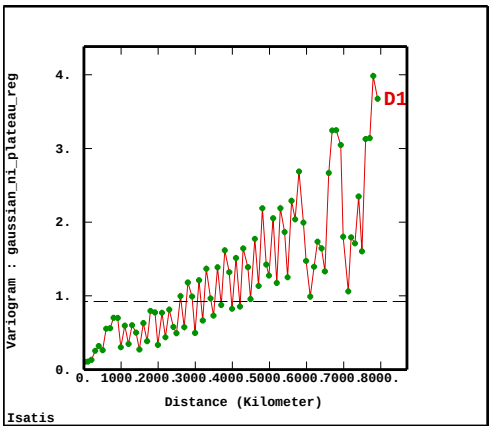
(b) Phytoplankton below 48,5 m ; only 1 direction in the horizontal plan



(c) Temperature above 70 m ; global anisotropy : $+70\,^{\circ}$



(d) Nitrate on the continental shelf, horizontal direction ; global anisotropy : $-15\,^{\circ}$



(e) Nitrate on the continental shelf, vertical direction

Figure 2.8: Experimental variogram for the different variable of our study. The variogram value is represented on the vertical axis and the values of h are represented on the horizontal axis

It will thus correctly fit the data, at leat for smaller values of $|h|$, while remaining stationary, but will obviously not give satisfactory results far away from the data points.

For the nitrate variable on the continental shelf, in the horizontal directions, we shall choose the cubic model with a small geographic anisotropy and a large zonal anisotropy ; it is to be noticed that the zonal anisotropy translates a larger variability in the direction crossing the global direction of the continental margin, which coherent with what we know of the nitrate behavior in the ocean. In the vertical direction, we have to combine a spheric and a cubic scheme with different scale and sill, in order to model nested small and large scales phenomena.

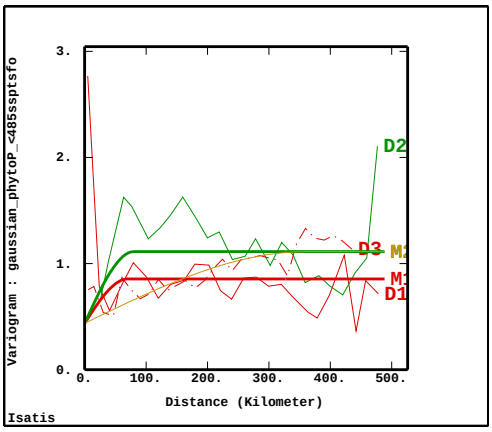
Performing the geostatistical simulations

We first have to define the grid on which we shall perform our simulations. The simulations are due to be representative of the reality measured in the OMEX dataset, so that the grid should be theoritically as fine as the distribution of the points in the dataset. However, we shall set up a not too fine grid, in order to avoid too long calculation times. We also have to define the spatial extension of the grid : the simulations, as most of the geostatistical modeling, cannot be trusted far away from the data points ; therefore we should only draw our grid on the zone prospected by the OMEX campaigns. As far as the phytoplankton variable is concerned, we shall build a very fine grid in order to perform an accurate study, with a 2.5x2.5 km grid in the horizontal plan, and a 25 m thick layering. For a matter of calculation time, the temperature and nitrate shall be coarser : 5x5 km x 25 m for the temperature grid, 10x10 km x 40 m for the nitrate grid.

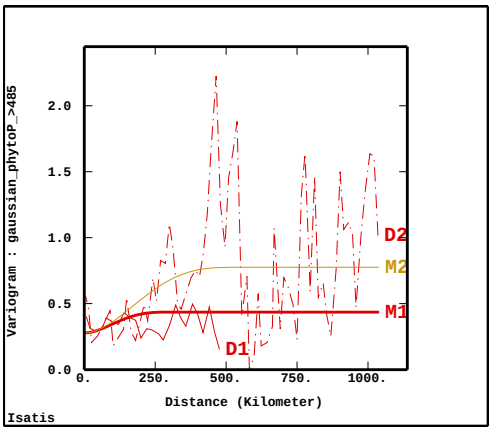
We have then to determine 3 parameters for the simulations. First, we shall choose the number of bands used in the Turning Band method. Indeed, in this method (see section 2.1.1), we have seen that we have to randomize the 1-dimensional vector s in the 3-dimensional space ; in practice, this randomization is only done on a finite number of directions. This can generate visible artefacts on the simulation results. We shall choose a number of turning bands (that is to say of "s"-directions) big enough to avoid these artefacts ; in our case, we shall use 500 Turning Bands.

Then we have to define the neighborhood used in the conditioning process. This neighborhood is defined around one given simulation point, and includes all the datapoints that will be used in the kriging procedures aiming at conditioning the simulation (see section 2.1.1). One one hand, it is important not to take to much points in our neighborhood ; indeed, the variogram model is usually better fitted on the smallest values of h than on the larger one (it is in particular the case for us when we have modeled a variable that does not seem stationary with a stationary model with a very large scale), so that considering the covariance between remote points would not be a good idea. On the other hand, if the number of points in the neighborhood is to small, the kriging processes will not give consistent results. As far as the variographic analysis revealed an anisotropy in the covariance function, we would like to take into account in the conditional simulations points in every direction ; therefore we shall divide the neighborhood into angular sector, and limit the number of empty angular sectors to validate the simulation at one given point; moreover, we shall use an anisotropic neighbor, larger along the oceanic margin than across, according to the spatial repartition of the OMEX points.

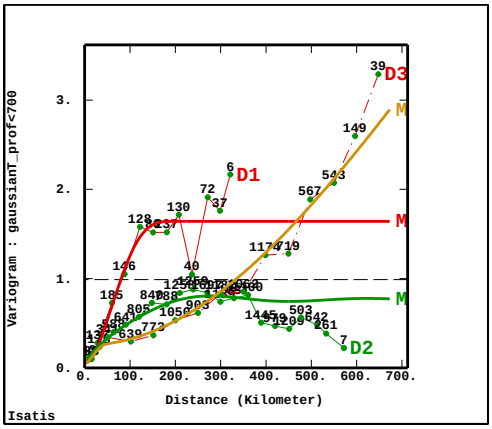
Finally, we have to define the number of simulations to perform. Here we shall perform ten simulation for each of the populations of phytoplankton, nitrate, and temperature. This should provide us with a set of results representative enough of all the simulations that could have been realized. We limited this number for a matter of calculation time, but a further investigation should have led us to perform far more simulations (up to 100), in order to be able to know if the particularities observed on each of our simulations are very representative of what reality should be or not (*i.e.* if the particularities observed are present on a large number of simulations or only on few ones). Furthermore, before conditioning to the data points, we must check the quality of the simulating procedure on the non-conditional simulations. This checking consists in verifying that the simulating process reproduce correctly the histogram and



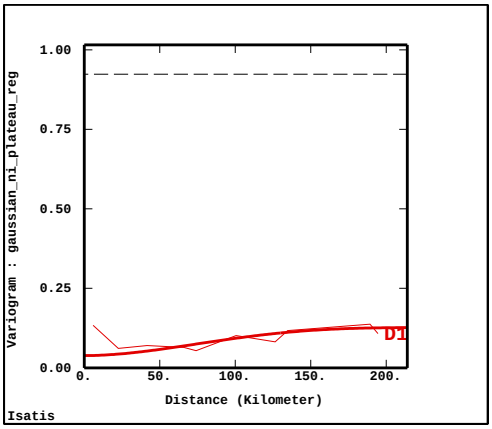
(a) Phytoplankton above 48,5 m ; global anisotropy : $+30\,^{\circ}$



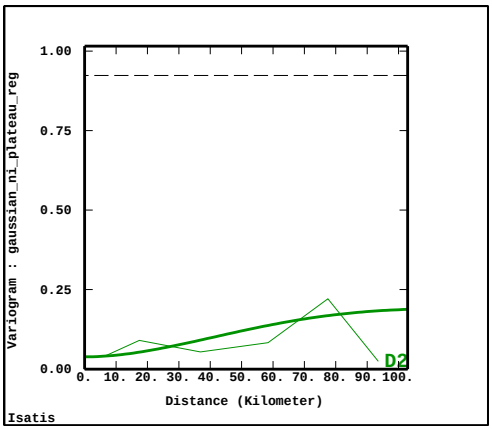
(b) Phytoplankton below 48,5 m ; only 1 direction in the horizontal plan



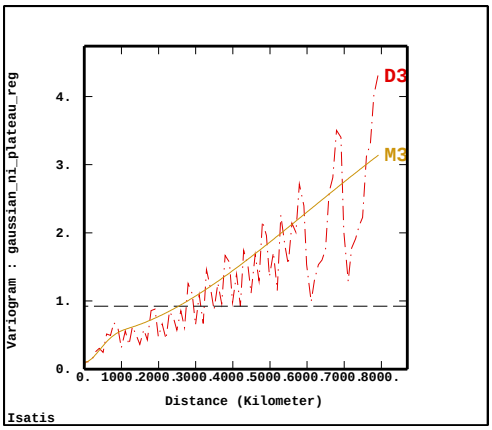
(c) Temperature above 70 m ; global anisotropy : $+70\,^{\circ}$



(d) Nitrate on the continental shelf, horizontal direction 1; global anisotropy : $-15\,^{\circ}$



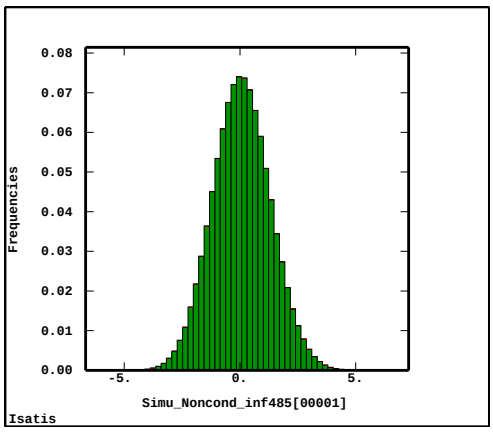
(e) Nitrate on the continental shelf, horizontal direction 2; global anisotropy : $+75\,^{\circ}$



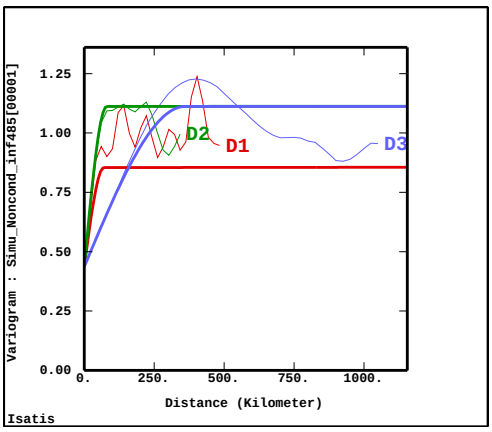
(f) Nitrate on the continental shelf, vertical direction

Figure 2.9: Model variogram for the different variable of our study. The variogram value is represented on the vertical axis and the values of h are represented on the horizontal axis

variogram calculated on the data. It is done on the non conditional simulation because, given that the conditioning is based on kriging, which produces smooth results (see section 2.1.1, page 17), it will not exactly reflect the variability seen on the data. The following figures (2.10) show the verification in the case of the phytoplankton variable above 48.5 m. We can see that the histogram obtained is very close to the Gaussian distribution, and that the variogram model is pretty well reproduced, especially for horizontal direction 2.



(a) Histogram of the non conditional simulated variable



(b) Variogram of the non conditional simulated variable

Figure 2.10: Checking of the simulations for the phytoplankton variable above 48,5 m

The final step in generating geostatistical conditional simulation of the phytoplankton, nitrate, and temperature variables is to transform the simulations of the Gaussian corresponding variables into simulation of the raw variables. This is simply done by using the anamorphosis function defined in the first step of the analysis (2.2.1, page 24).

Generating the Averaged Geostatistical Simulation

Generating the Averaged Geostatistical Simulation seems to be an easy job in comparison with the simulation process. However, even if the mathematical part of it is quite evident, its practical implementation is not so easy. Indeed this approach is quite original in such a study, and therefore no particular function has been developed for it into Isatis. The difficulty comes mainly from the fact that the DIADEM grid is very irregular, especially in the vertical direction, so that it is very hard to determine in which cell a point is. We shall then look at the problem from another point of view : we shall consider that one cell in the DIADEM grid is influenced by the simulations points that are the nearest from it center. For each point of the simulation grid, we can reversely determine the nearest DIADEM cell center, and thus define the DIADEM cell it influences. This operation can easily be carried into the geostatistical software. We shall finally get the AGS value in each DIADEM cell by averaging all the values of the simulated points that are considered as influencing this cell, which can be done using any common data manipulation tool. We shall do it manually in this methodological study, but further applications of the method would require a more automated protocol.

Having performed all these steps in order to obtain Average Geostatistical Simulations of our three variables phytoplankton, nitrate, and temperature on the whole domain of our study, we have then to compare this results with the output of the DIADEM ecosystem model. This is the points of Chapter 3, where we shall try not to forget the way how we calculated our geostatistical modeling of reality, all the assumptions we made and the various artefacts they might engender.

Chapter 3

Comparison between the OMEX AGS and the DIADEM Model Results

Having of the Omex AGS on the one hand, and of the DIADEM results on the other hand, we would like to perform a quantitative comparison between these two dataset, using precise statistical or geostatistical tools, in order to point out the main characteristics of the numeric model DIADEM. This will be done using three main tools : the histograms, the scatter diagrams, and the proportion curves, whose main properties are reminded in the first section of this chapter. However the observations drawn from one of this tool is often confirmed by the others, and the analysis should combine the use of the three simultaneously. This is what have been done in the second section, where the practical comparison between the OMEX AGS and the DIADEM results is presented. It is yet important in this step not to take the values given by the AGS for granted ; indeed, they are only the results of one other kind of model, this time statistical rather than bio-physical. Therefore we shall keep in mind all the steps that we implemented to obtain the averaged simulations, and do not hesitate to use the intermediate steps, such as the anamorphosis function or the punctual simulations, to better inform the results of our comparison.

3.1 A few useful quantitative tools...

We choose in this study to concentrate on three main tools : the histograms, the scatter diagrams, and the proportion curves. In order to draw sensible conclusion from the observations noticed, we shall then have a good knowledge of their main properties.

3.1.1 Histograms

The histograms plot a discrete representation of the distribution of a given variable in a dataset: the interval of definition of the variable is spitted into a chosen number of classes, and for each class the proportion of points in the whole dataset taking their value in the class is plotted as a bar. We can perform two main types of observation on an histogram.

First, an histogram can be characterized by its modes. A mode is defined as a maximum of the density function of a random function. The different modes of the random function are then most probable values taken by this function, and are represented by the peaks of the histogram (see figure 3.1). An important point is then to check if the DIADEM model correctly reflects the modes observed on the OMEX AGS, *i.e.* if it give the same most probable value as the statistical study of the measured reality.

Furthermore, the histogram is a very good tool to evaluate the way the model deals with the support

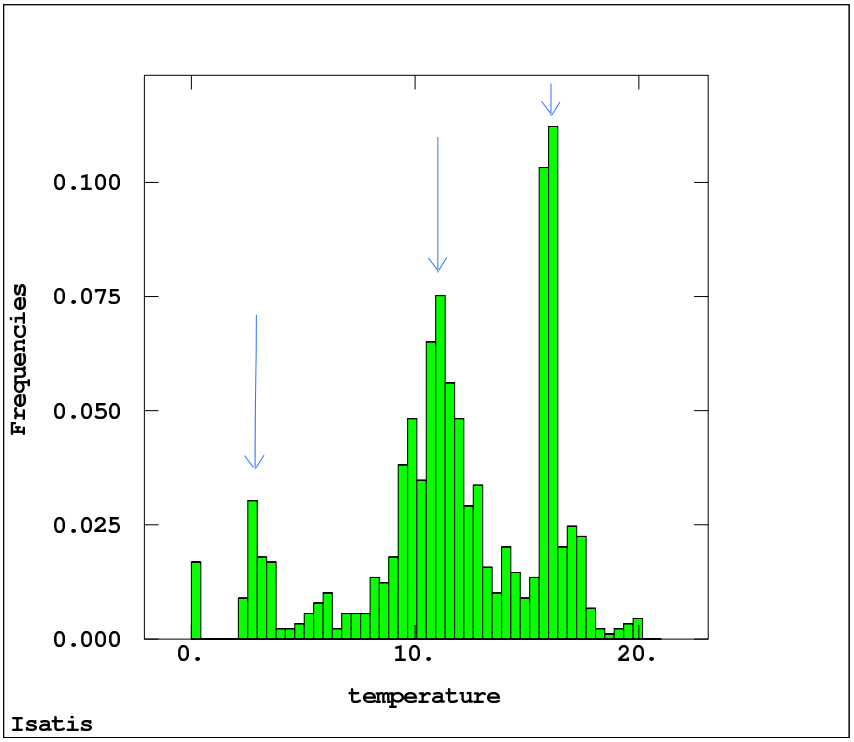


Figure 3.1: Example of histogram with 3 modes (blue arrows)

problems evoked in section 2.1.2, page 22. It is indeed a representation of the frequency plotted on figure 2.2, so that we would expect the histograms of the DIADEM results and of the OMEX AGS, both calculated on large support variables, to be narrower and higher than the histogram of the OMEX variables, or than the one of OMEX punctual simulations. It will then be very interesting to superpose the histograms of the OMEX punctual simulations, AGS, and the DIADEM histogram. It will allow us first to check the support effect on the AGS in comparison with the punctual simulations, and then to see if the DIADEM model expresses the same effect or not.

3.1.2 Scatter diagrams

A scatter diagram can be drawn using two variables defined at the same points. In our study, it is the case for the OMEX AGS amd the DIADEM results.For each cell of the DIADEM grid, we shall then plot the value given by the OMEX AGS against the value given by the DIADEM results. We shall thus obtain a cloud of points in the plane (DIADEM results, OMEX AGS). If both the numeric and the geostatistical model were perfect, and gave us the “true value” of phytoplankton, nitrate and temperature, all the plotted points would be located on the first bisector line of our graphic ; obviously it is not the case, so that we usually get a cloud of points spread around this idealistic line (see for example figure 3.2). It is then very interesting to study the shape of the cloud of points, and in particular for the points that are the most faraway from the bisector line. This reveals on which kind of values (high, low, middle, in a precise class, etc.) the numeric model and the geostatistical simulations strongly disagree. Moreover, Isatis allows us to locate on the data basemap the very points that are different from one dataset to the other. We can then check if the difference can be charged rather to the DIADEM model or to the geostatistical analysis. Namely, thanks to the basemap, it is rather easy to see if the points where the OMEX and DIADEM values were found pretty different are close or not from the data points. If not, it means that the conditioning of the simulation by the OMEX data did not influence much on the result, so that the OMEX AGS cannot be trusted, and the DIADEM model’s results appears to be more reliable. But if the disagreement points are located in the center of our domain of interest , where the OMEX data distribution is very dense, the

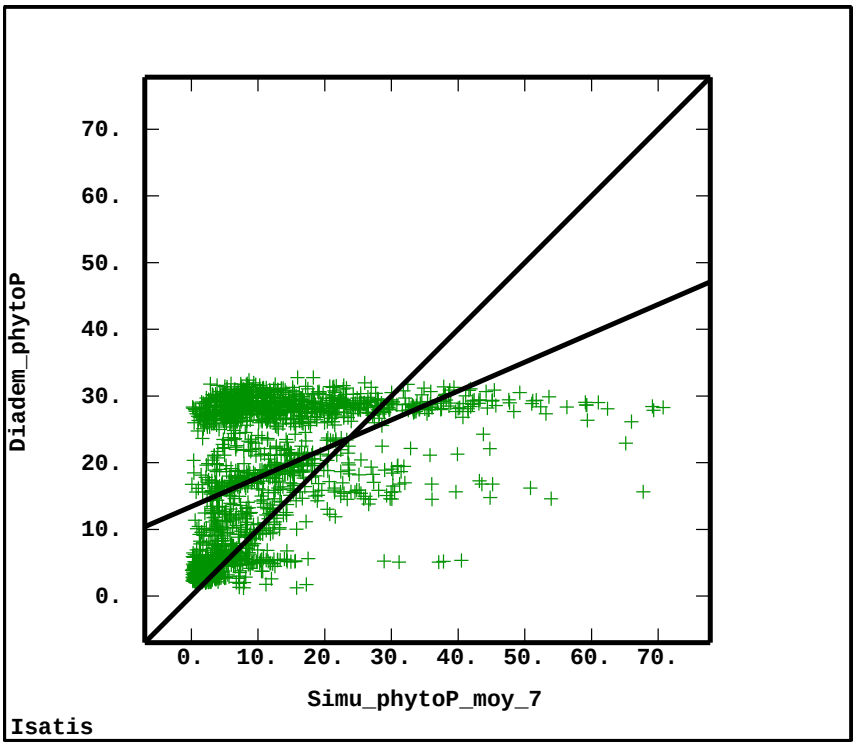


Figure 3.2: Scatter Diagram between the phytoplankton variable modelled by DIADEM and the OMEX AGS #7

simulations at that points are very well conditioned, and can be far more relied on than the DIADEM model.

3.1.3 Proportion curves

The proportion curves are usually used in the mining field, where they are called “Grade Tonnage curves”. However, some of can also have applications environmental such as the one we focus on. We shall especially concentrate on the Proportion above cut-off curves, which plot the proportion of points in a dataset above a given cut-off, in function of this cut-off values. This calculation usually results in such curves as the one plotted on figure 3.3. The comparison between the Proportion curves plotted for the OMEX AGS and the DIADEM results underlines the ability of the model to reproduce faithfully the extreme values in the data set, and the global trend of a model to be too high or too low in comparison to the other one. Moreover, the ability to reflect correctly the number of point above or under a given cut-off is fundamental for a lot of practical application of the ecosystem model. Let us give an example quite close to the mining field : in the mining area, the proportion curves are used to evaluate the economic value of a deposit by estimating the part of it where the content of ore is higher than a given limit value; we can imagine similarly that in the fishing industry, it will be rentable to launch a campaign in area only if there is at least a certain amount of fish likely to move around there, and that the amount of fish in a given area is quite linked to the amount of food that they can find, *i.e.* the concentration in phytoplankton in a given zone. It can then become very interesting to model properly the amount of cells, in a marine domain, where the concentration in phytoplankton (or zooplankton, another variable modeled by DIADEM on which we have not focused) is above a given cut-off related to the fishes. Cut-off values are also very useful in environmental issues, where we often define limit values on certain parameters, above which we consider that we face a more or less serious pollution. Finally, a last example of the cut-off utility can be found in the off-shore oil field where the development of a platform is limited by such physical phenomena as currents.

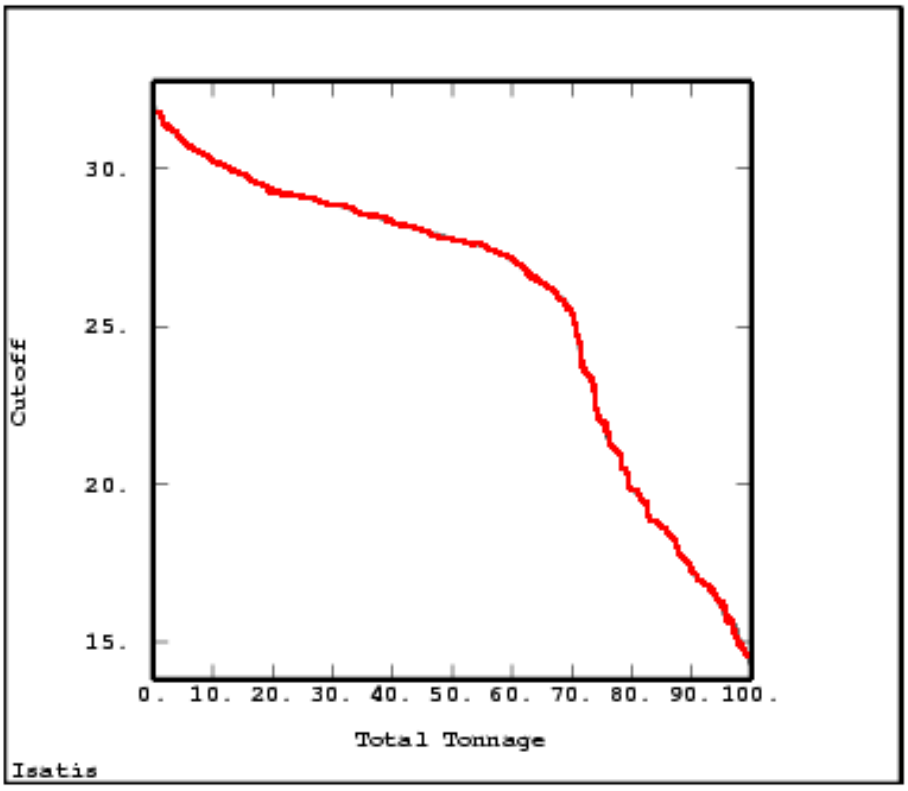


Figure 3.3: Proportion above Cut-off Curve plotted for the Diadem phytoplankton on the 4 upper layers of the ocean

3.2 Practical comparison between the DIADEM results and the OMEX geostatistical study

We shall know present the main results obtained during the practical comparison drawn on the results of the DIADEM model and the OMEX dataset, and the geostatistical study we performed on it. As we already said, the use of the three tools presented in the previous section are complementary, so that we shall expose our results variable after variable, for the phytoplankton, the nitrate, and the temperature. To keep this document comfortable for reading, we only expose here the graphics relevant to our conclusions ; the ensemble of the graphic obtained during our study can be found in the appendix part of the present document.

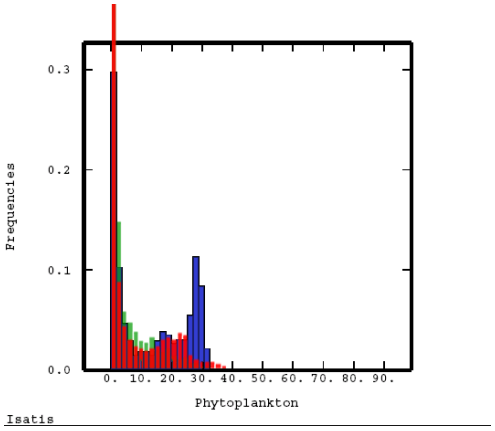
3.2.1 About the phytoplankton variable

The phytoplankton was the prior variable in our study, so that we shall perform an accurate study on it. As we already said, the grid used to calculate the phytoplankton simulations was the finest one, with 2.5x2.5kmx25m cells, and we performed the whole study for the whole area of interest, *i.e.* in this case above and below 48.5 m.

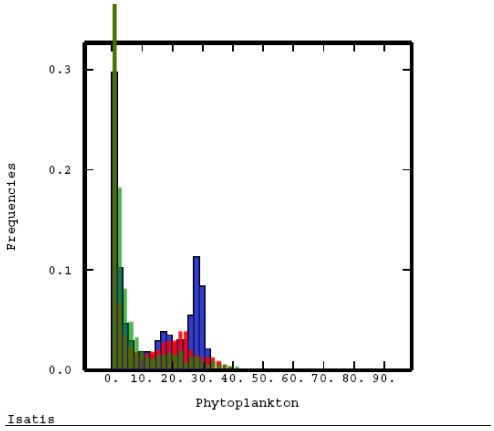
General study on the whole domain

The difficulty we are facing with the phytoplankton variable is that, as we have seen in Chapter 1, the variable is not defined on the all depth. This is the case of course in the OMEX dataset, where the phytoplankton values are only given for points above 105 m, and can be observed on the DIADEM results

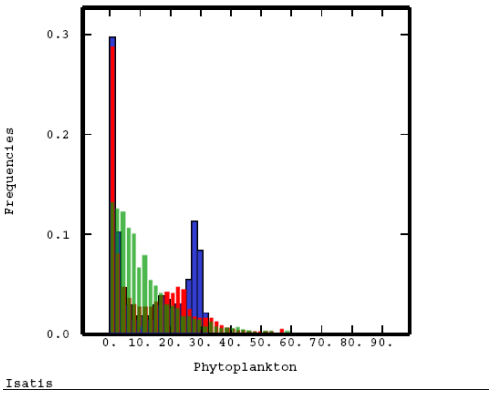
too : under 175m, all phytoplankton concentrations are equals to zero. By superposing the histograms of OMEX AGS or simulations above 105 m with the histograms of the DIADEM results above 175 (see figure 3.4), we can see that the results are quite close, so that we can infer that the DIADEM model simply “extends” the zone located above 105 m for the OMEX points down to 175 m. We shall then perform the study for the whole depth between 0 and 175 m for the DIADEM model outputs when it is possible.



(a) Using OMEX simulation #1



(b) Using OMEX simulation #4



(c) Using OMEX simulation #7

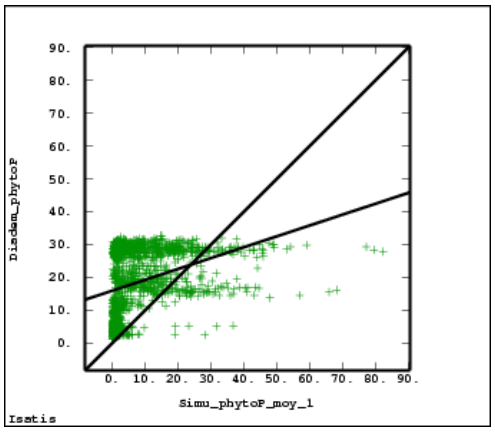
Figure 3.4: Superposed histograms of the DIADEM results (in blue)above 175 m, the OMEX ponctual simulations (in red), and the OMEX AGS (in green) above 105 m

Observing more in detail the histograms, we can notice the DIADEM histogram presents a pretty high peak around 30 mg/m^3, which cannot be seen neither in the OMEX simulations nor averaged simulations. Furthermore, there are very few DIADEM values above this peak (the maximum is set to 32.77), whereas, as the data, the OMEX simulations extend to around 90 mg/m^3, and the averaged simulations can reach up to 70 mg/m^3. We first thought then that there was some kind of cut-off value on the phytoplankton variable in the DIADEM model, which would cause a carry-over of the highest values around 30. However, there is no direct limiting value in the model that would explain this cut-off. Besides; it appears that we are studying a period (end of summer/begin of fall) when the biological activity in the ocean is not very high, so that the model must be able to give far greater values for the phytoplankton in the spring bloom ; as far as this bloom has already been studied for this model (in [12] for example), we know that it is the case, and we shall forget about a direct limit imposed in the model. This is confirmed

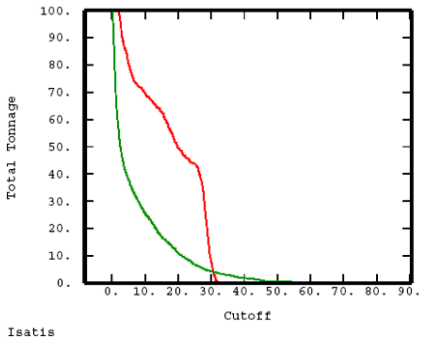
by the fact that even if the frequency of points above 30 falls very quickly, there are some values in the following classes of the histogram, so that the image we have does not really correspond to the one usually given by a cut-off. We shall then assume that this is due to some problem in the model, which is not so easy to determine as far as the equations describing the phytoplankton compartment are very nested and complicated to analyze.

Choosing not to focus on the abnormal peak around 30, we can try and compare the rest of the distribution with the one obtained for the OMEX simulations and AGS.The main mode of the histogram, around small values, is pretty well reproduce by the DIADEM model. However, it seems that the distribution given by DIADEM is somehow closer the one of the punctual simulations than to the one given by the AGS, especially if we consider the small increase in the frequencies for the phytoplankton values around 17-20 mg/m^3, which appears on the punctual simulations but has been erased by the averaging.

The scatter diagrams and proportion curves plotted on the total depth reveals the same characteristics as the histograms : the DIADEM values are not high enough, which is clearly visible on the scatter diagrams (see figure 3.5(a). Moreover, it appears that even on the part where the histograms seem to match, the same values given by DIADEM or OMEX do not always correspond to the same points, as far the global scatter diagram is pretty far from the first bisector line. The Proportion curves (see figure 3.5(b)) show us that the model reproduce correctly the proportion above cut-off for small values of phytoplankton, and the mismatch between the DIADEM results and the OMEX AGS can be easily explained considering what we have already noticed : between 10 and 30 mg/m^3, the DIADEM curve is far too high, due to the presence of a lot points around 30. Then above 30, the DIADEM curve brutally fall to zero, because of the absence of high values in the model results, whereas the OMEX AGS more slowly goes to zero. We shall now perform a further study in order to check if these features can be seen also in the two different domain we used in our analysis, *i.e.* above and below 48.5 m.



(a) Scatter Diagram : DIADEM results (on the vertical axis) against OMEX AGS #4 (on the horizontal axis)

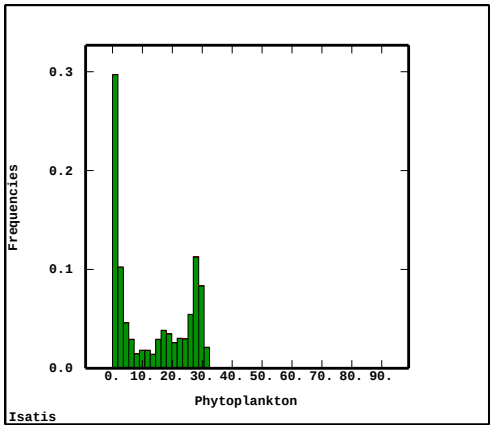


(b) Proportion Curves for the DIADEM results (in red) and the OMEX AGS 1 (in green)

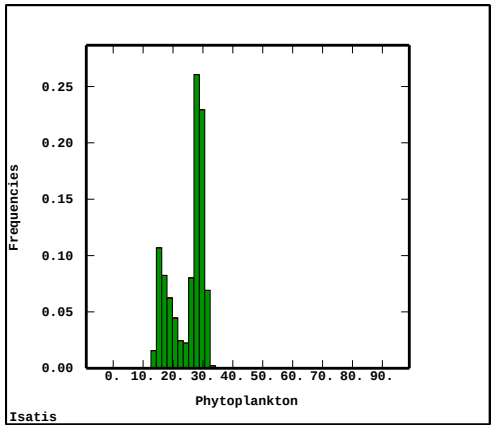
Figure 3.5: Comparison on the whole depth

Comparison according to depth

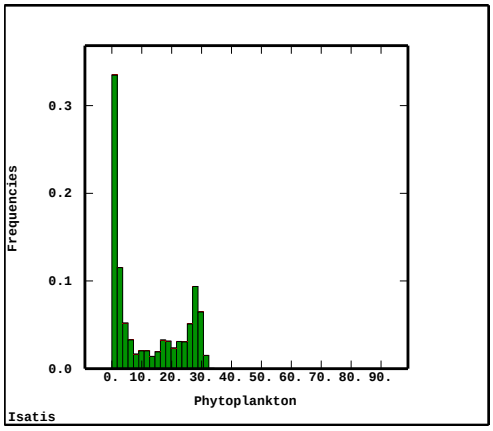
We have distinguished on the OMEX data two different populations of data according to the depth. We shall try and see in these distinction can be made on the DIADEM results too : the DIADEM histogram also reveals two distinct modes, but a simple splitting of the DIADEM points into groups above and below 48.5 m does not give satisfactory results (see figure 3.6). Nevertheless, a further study of the DIADEM



(a) Histogram of the DIADEM results above 175 m



(b) Histogram of the DIADEM results above 48.5m



(c) Histograms of the DIADEM results between 48.5 and 175 m

Figure 3.6: Splitting the DIADEM results in two subsets according to depth

dataset shows that the two modes in the histogram correspond to different layers of the DIADEM model : the highest values in the DIADEM results are almost exclusively corresponding to the 4 upper layers of the model, whereas the lowest values are located in the 14 deepest layers (see figure 3.7). We shall then perform a comparison between the DIADEM results in the 4 upper layer an d the OMEX AGS above 48.5m, in order to check if the model has not once more “extended” the behavior of phytoplankton in the upper part of the ocean to its 4 upper layer; and do the same below 48.5 m and on the 14 deepest layers of the grid.

Comparing the histograms above 48.5 m and on the 4 upper layers (see figure 3.8), the problem due to the “30 peak” evoked in the previous section is even more flagrant: the distribution of the DIADEM outputs are not as widely distributed as the OMEX AGS, maybe denoting a problem of support, here the results seems to even more smoothed than they should be considering the cells of the DIADEM grid. We should notice anyway that, even if the mode of the histogram, located at 30 off-course in the DIADEM case, and lower on th OMEX AGS, is not well reproduce by the model, there is a kind of secondary mode for DIADEM around 17-20 which would better match the OMEX results.

The proportion curves also reveals the same features as observed on the total depth : the DIADEM curve is over-estimated before 30, and fall to zero just after, whether the OMEX curve is smoother. The

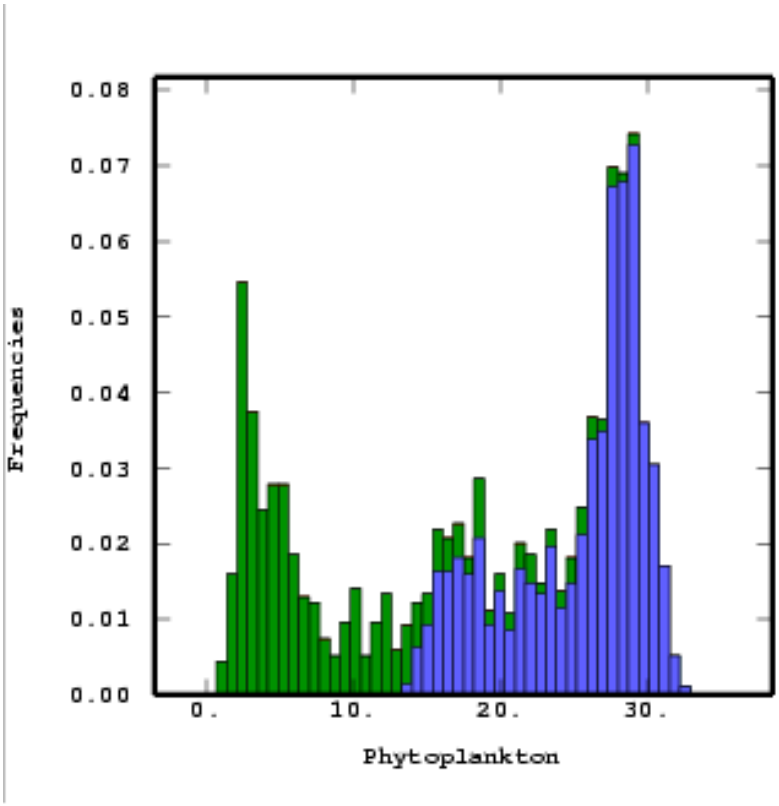
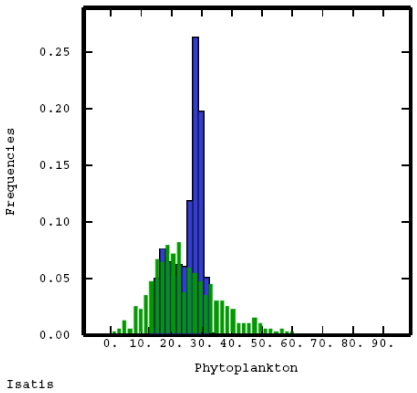
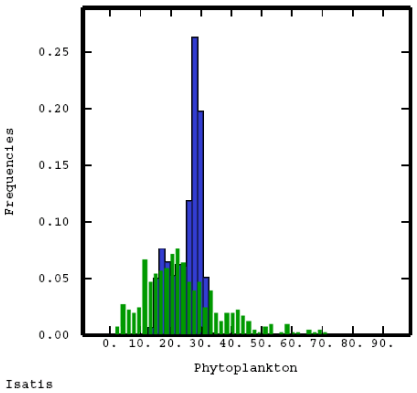


Figure 3.7: Histogram of the DIADEM results above 175 m ; the points belonging to the 4 upper layers of the grid are highlighted in blue



(a) Using OMEX AGS #4

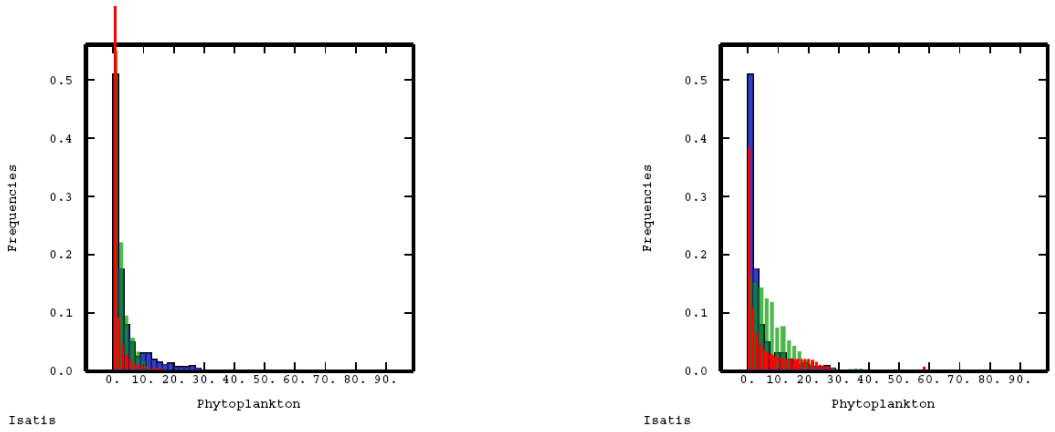


(b) Using OMEX AGS #7

Figure 3.8: Histograms of the DIADEM results on the 4 upper layers of the grid (in blue) superposed with the OMEX AGS above 48,5 m (in green)

scatter diagram obtained is not good at all : indeed we can only plot it for the point above 48.5 m, as far as we need to have the 2 datasets defined at the same points. Further more, if the DIADEM model reproduce the phenomenon observed above 48.5 on a thicker depth, we won't expect it to give the good value at the same points as the OMEX geostatistical model.

The DIADEM model results distribution on the layers 5 to 18 of the grid, above 175 m, pretty resembles the OMEX AGS histogram below 48.5 m (see figure 3.9). The global shape of the histogram is the same, but we can wonder if the support effect is taken into account by the ecosystem model or not. Effectively if you look at simulation #7, you would say that the model reproduces well the average simulation, whereas compared to OMEX AGS #4, it seems to have a distribution corresponding to a smaller support. In fact, for this short example study, we only performed ten simulations per variable, which allows us to have an image of several different "possible reality" given the data variability, but we did not perform enough simulations to be able to distinguish a global trend in the histograms resembling rather to AGS #4 or #7. To conclude firmly on this question, we should have performed about 10 or 20 times more simulations, which was not possible in our case, but should be done in a further practical study using these methods. Besides, we can notice that for each simulation, there are a lot of null or very small value for the OMEX simulations. In fact, this is due to the practical geostatistical model rather than to the data themselves ; indeed, examining the anamorphosis used for the phytoplankton below 48.5 m (see figure 2.5, page 26), we notice that the lower limit on the Gaussian variable practical interval of definition is set to -1.10, which means that any value drawn below -1.10 for the Gaussian simulation (which is not such an improbable value!) will be set to 0 when performing the back transform into a raw variable. This can explain the high peak on zero for the OMEX simulations and AGS, and is therefore not so much reproduced by the DIADEM model.

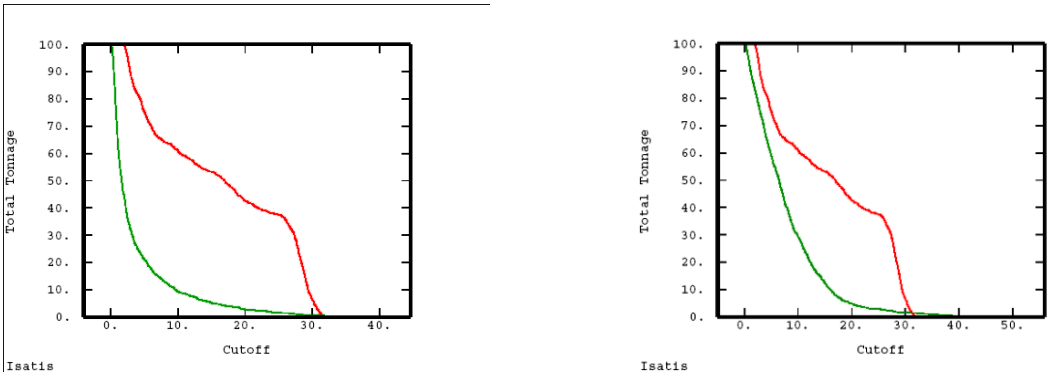


(a) Using OMEX AGS #4

(b) Using OMEX AGS #7

Figure 3.9: Histograms of the DIADEM results on the 14 deepest layers of the grid (in blue) superposed with the OMEX simulations (in red), and the OMEX AGS below 48,5 m (in green)

Here again, the scatter diagram does not make much sense, as far as it could only be defined for the points between 48.5 m and 105 m, and that we won't expect a model "extending" a phenomenon from between 48.5 and 105 m to the 14 deepest layers above 175 m to give the same results at the same points that the geostatistical study of the initial data. The proportion curves, however; are quite good, which confirms that the model can be trusted to test small limit values on the deeper part of the DIADEM grid.



(a) Using OMEX AGS #1

(b) Using OMEX AGS #7

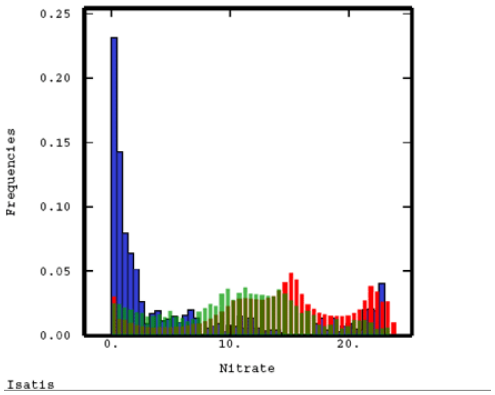
Figure 3.10: Proportion curves of the DIADEM results on the 14 deepest layers of the grid (in red) superposed with the OMEX AGS below 48,5 m (in green)

3.2.2 About the nitrate variable

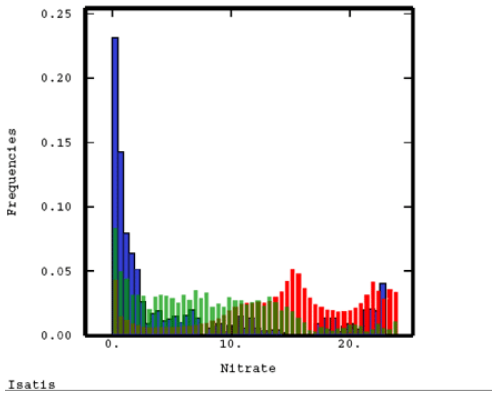
As far as it was the second variable extracted from the biochemical model's results, we didn't focused as much on the nitrate variable as on the phytoplankton. Therefore we only studied the shallow part of area, *i.e.* the continental shelf, with a quite coarse simulation grid (10x10 km cells in the horizontal direction and m thick layer in the vertical one).

The observation of the histograms comparing the DIADEM outputs with the OMEX simulations and AGS distribution shows that there is a main difference between them around the smaller value, with a quite higher and wider peak on the DIADEM results than on the OMEX simulations (see figure 3.11). However, this is present in the OMEX dataset itself, but not as highly as in the DIADEM results, and is well reproduced by the anamorphosis function used for the nitrate, as figure 3.12 shows. If not considering this main difference, the DIADEM distribution is quite concordant with the OMEX AGS one, and is even closer from the AGS model than from the punctual simulations' one, proving the quality of the model for this variable.

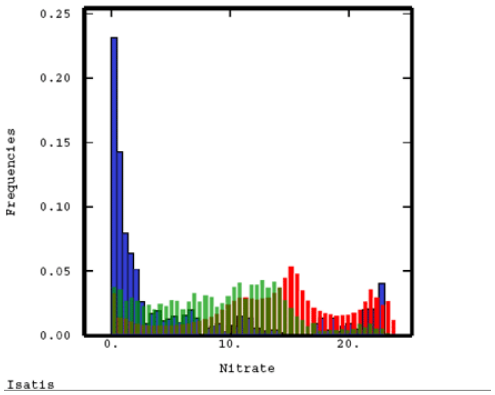
We can then perform a more accurate study using the scatter diagrams : if we plot on a basemap the points for which the concentration of nitrate is set to very small value by the DIADEM model and a larger value by the OMEX averaged simulations, we simply realize that all those points are located on the north-western part of the studied zone, where there is only one profile of measurement to condition the geostatistical simulation (see figure 3.13). We can therefore not rely on the geostatistical analysis in this zone, and shall consider the DIADEM results as more probably correct. Besides, if we just look at the point out of this zone, they are quite close from and well-distributed around the bisector line, which means that not only the ecosystem model give the good range of values to the grid cells, but also does so at the good points, unlike for the phytoplankton, for example, where we have seen that the natural phenomenon were somehow "extended" in thicker area. The DIADEM model can therefore be considered as quite valuable as far as the nitrate variable is concerned, since it produces matching results in the well conditioned simulation area, with a good compliance with the support of the DIADEM grid. Moreover, we have seen that the covariance model chosen for the nitrate variable on our zone was not applicable to the whole zone, revealing that on the north-western part of the domain, there was probably another behavior of our variable, with, as we said, smaller nitrate concentration, which would have probably lead us to using to different model if we had had enough data in this part. This analysis is moreover compliant with the general phenomena occurring on the oceanic margin, where the regeneration of nutrients is ensured by upwards currents on the border of the continental edge, *i.e.* where the data profiles where sampled, whereas the north-western part of the zone, more far on the continental



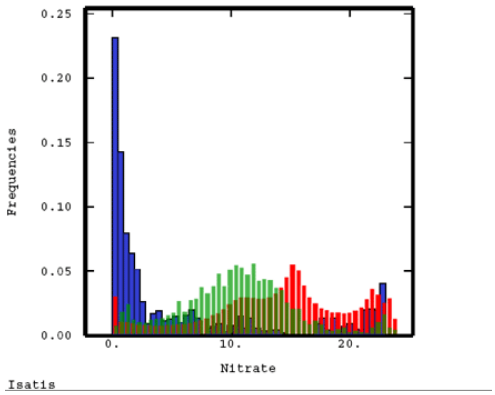
(a) Using AGS #4



(b) Using AGS #8



(c) Using AGS #9



(d) Using AGS #10

Figure 3.11: Histograms of the nitrate variable comparing the OMEX DIADEM results (in blue) with OMEX punctual simulations (in red) and AGS (in green)

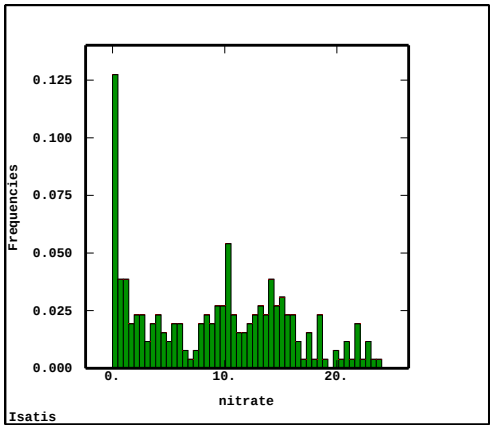
shelf, does not benefit as much of this nitrate input. We shall not make any comment on the proportion curves, as far as the large amount of small value points that did not appear in the OMEX analysis off-course distort a lot the obtained curves.

3.2.3 About the temperature variable

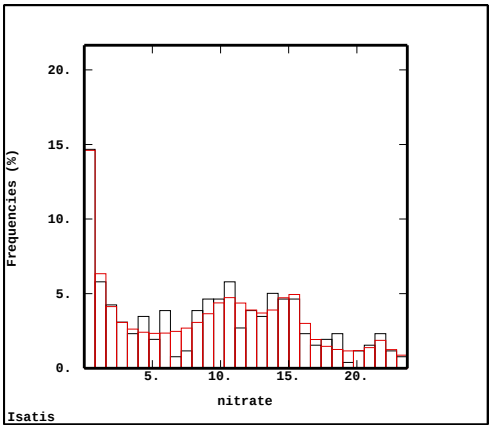
We had only time to compute the study of the temperature variable on the upper zone defined in Chapter one, *i.e.* for the points above 70 m depth. This zone was indeed far smaller than the area between 70 and 1300m, allowing us the set up a quite fine simulation grid : 5x5 km cells in the horizontal direction, and 25 m thick layers.

Both the histograms (figure 3.14) and the scatter diagrams (figures 3.15 and 3.16) shows that the temperature values are far not enough spread between 10 and $20\,^\circ$ in comparison with the OMEX AGS. Moreover, we can see that the mode of the DIADEM histogram is to high, around $17-18\,^\circ$ against 15-16 for the OMEX AGS or simulations.

The further study of basemap associated with the scatter diagram does not give much reliable results



(a) Histogram of the OMEX Nitrate data

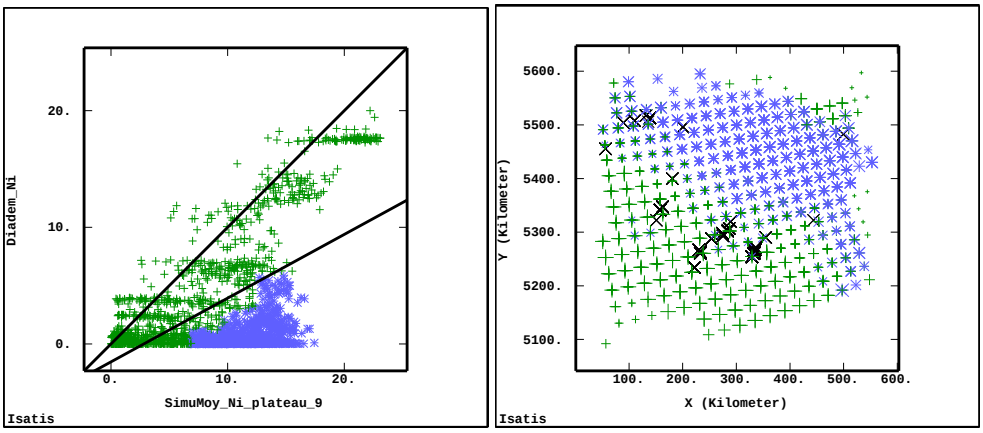


(b) Histogram of the OMEX nitrate data (in black) compared to the one obtained after the back anamorphosis computed on a Gaussian variable (in red)

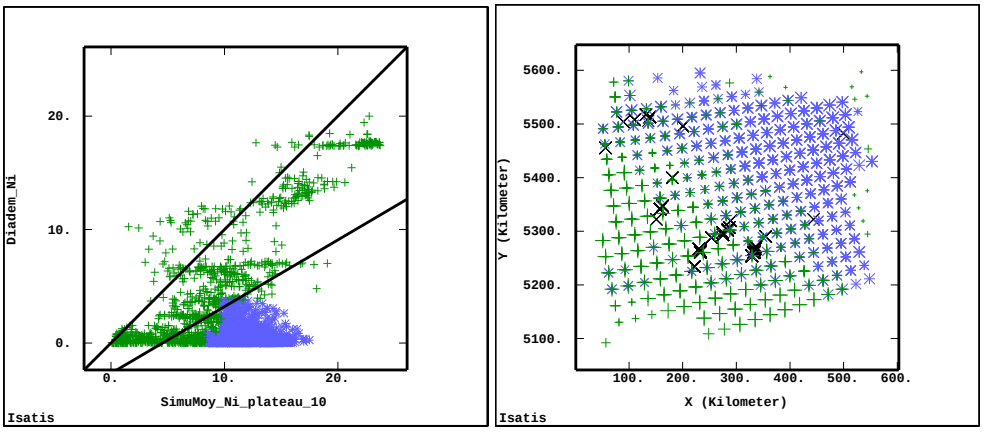
Figure 3.12: The OMEX data distribution is well reproduced by the anamorphosis for the small value of nitrate

: whereas while examining the basemap in figure 3.15, it seems that the points for which the DIADEM and OMEX AGS results are the most different are away from the data points, figure 3.16 indicates that the best correspondence between our two models also occurs far from the conditioning data! In fact, a further information on the model explains that the physical model is only constituted of 17 layers, among which a thick surface mixed layer that has been splitted into 2 layers in the biochemical model, but which is only provided with one value for the physical variable.

This implies that the distribution of the OMEX value is narrower, that is to say “more averaged” than we thought in the upper 70 m of the ocean, so that the DIADEM model correctly implemented the change of support between the physical and biochemical model. Finally, as far as the histogram mode is concerned, we have to take into account the fact that for the studied period, the physical model is just ending its spin-up, and that it has already been noticed that in this case, the modeled temperature are sometimes too high, a defect that is corrected afterwards by assimilating data.

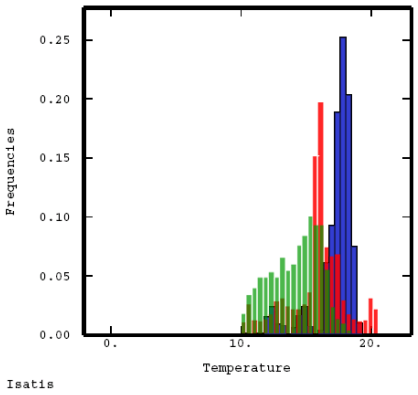


(a) Using AGS #9

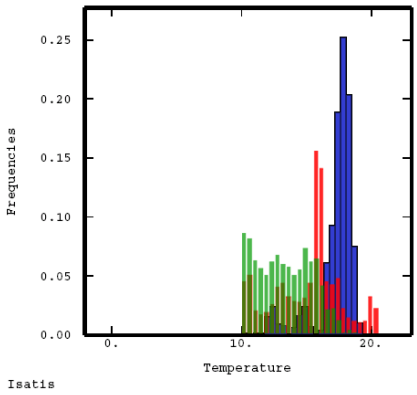


(b) Using AGS #10

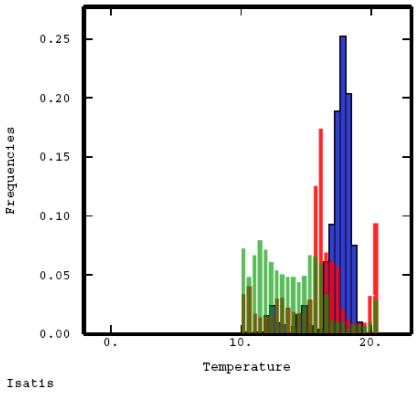
Figure 3.13: Scatter Diagram of the DIADEM nitrate outputs (on the vertical axis) against the OMEX AGS (on the horizontal axis). The higlighted points are plotted in blue on the basemap



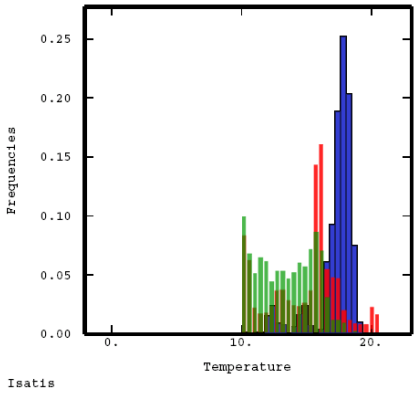
(a) Using AGS #3



(b) Using AGS #6

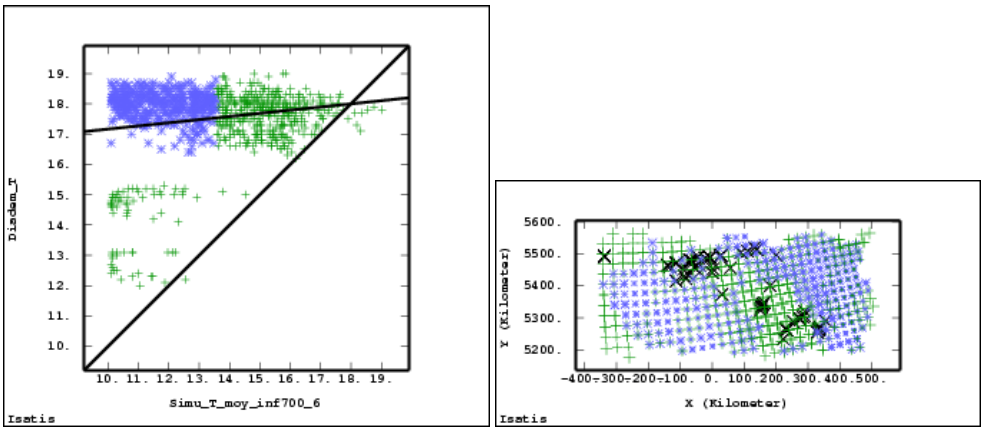


(c) Using AGS #7

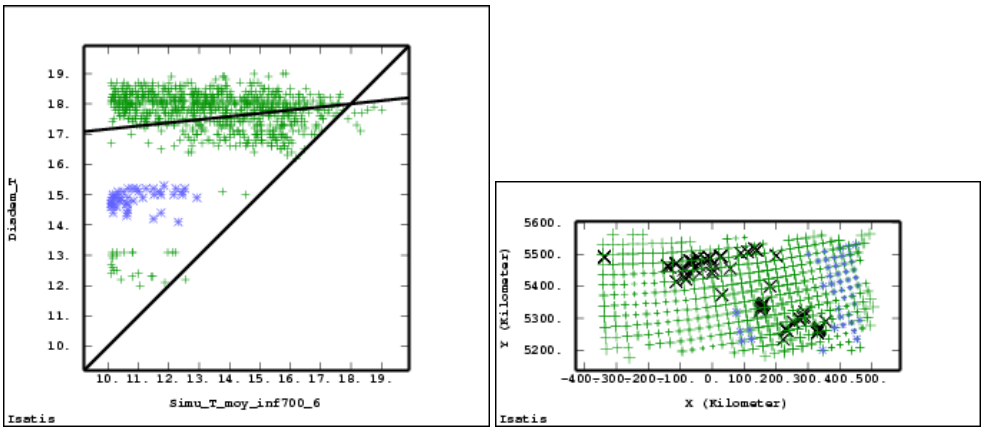


(d) Using AGS #9

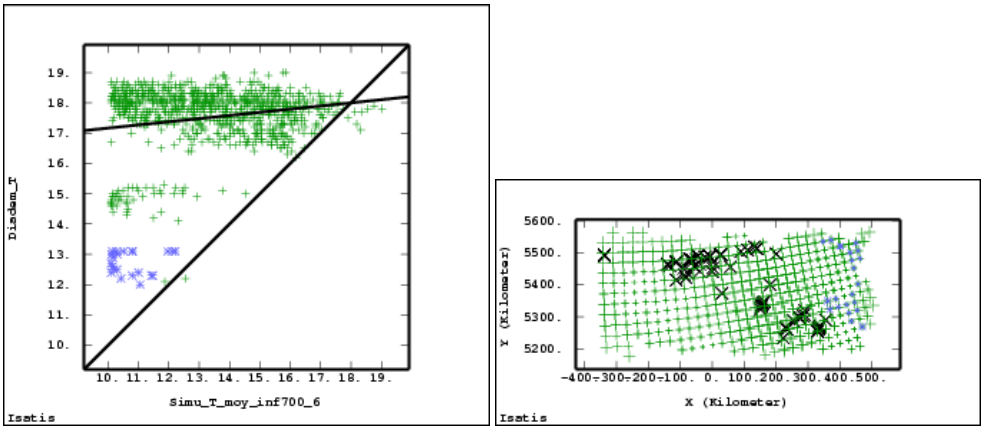
Figure 3.14: Histograms of the temperature variable comparing the OMEX DIADEM results (in blue) with OMEX punctual simulations (in red) and AGS (in green)



(a) Using AGS #6

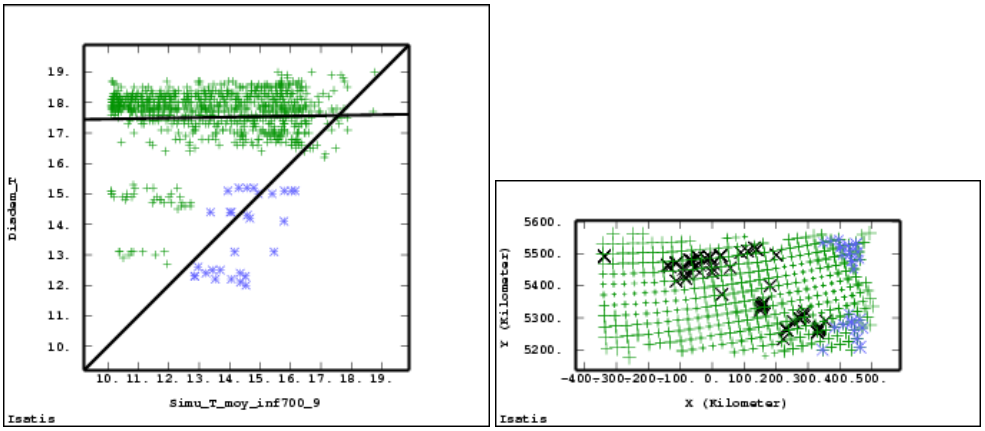


(b) Using AGS #6

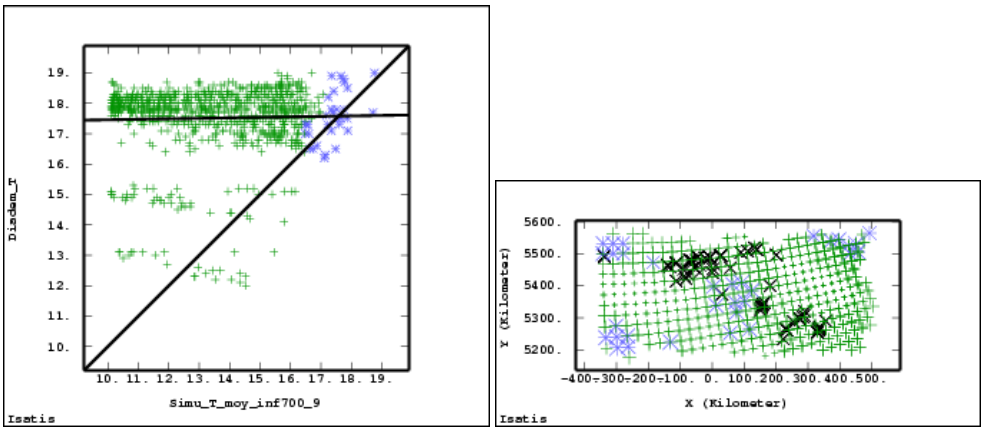


(c) Using AGS #6

Figure 3.15: Scatter Diagram of the DIADEM temperature outputs (on the vertical axis) against the OMEX AGS (on the horizontal axis). The highlighted points are plotted in blue on the basemap



(a) Using AGS #9



(b) Using AGS #9

Figure 3.16: Scatter Diagram of the DIADEM temperature outputs (on the vertical axis) against the OMEX AGS (on the horizontal axis). The higlighted points are plotted in blue on the basemap

Conclusion

This practical study based on the comparison of the results of the marine ecosystem model DIADEM with Average Geostatistical Simulation calculated thanks to the OMEX I dataset aims at establishing a model for further implementation of this original method for validating oceanic models. The case we were facing here, with highly irregular cells in the model grid, and a great difference between the *in situ* data distribution and the model results support, was indeed particularly adapted to the utilization of such geostatistical techniques, and during comparison itself in the last chapter of this report, we came across the main different type of results we could obtain : non-reliable geostatistical analysis far away from the data points, non-satisfactory model results due to the implementation of the model itself, support problem, good correspondence between the model and the geostatistical calculation . . . , always keeping in mind that we were not comparing a model to a reality that we shall never know anyway, but only comparing to different and complementary model techniques in order to improve both of them. We also relied a lot on the knowledge we had of the numeric ecosystem model and of the physical and bi-chemical environment we were studying to get a better understanding of the results of our comparison.

This methods should now be further investigated by applying it to new practical study and exploring more deeply all its possibilities, like using more simulations as we already said, or combining more quantitative analysis tools to perform a more accurate comparison between the averaged geostatistical simulations and the model results. It would also be very interesting to try and use it with other model, in particular the new combination between the biochemical model used in DIADEM and the physical oceanic model HYCOM, where the top layering of the ocean is no more isopycnic, but regular, so that the modeling of the mixed layer should be far more accurate than here. Furthermore, it would be interesting to test the effects of data assimilation on the model outputs in terms of support, as far as for the moment the fact that the data used in assimilation are usually defined on a small support is not taken into account, and as it will be worth validating an assimilated model which such *in situ* data that are not used in the assimilation process.

Bibliography

[1] R. Bleck, C. Rooth, D. Hu, and L. T. Smith. Salinity-driven thermocline transients in a wind- and thermohaline-forced isopycnic coordinate model of the North Atlantic. *J. Phys. Oceanogr.*, 22, 1992. 7

[2] Catherine Bleines, J. Deraisme, François Geoffroy, N. Jeannée, S. Perseval, Frédéric Rambert, Didier Renard, and Y. Touffait. Isatis software manual, November 2001. 20

[3] Clotilde Breuillin, Geir Evensen, and Mette Eknes. Data assimilation methods applied to marine ecosystem models. Technical Report 187, Nansen Environmental and Remote Sensing Center, Bergen, Norway, 2000. 6, 11

[4] British Oceanic Data Center. OMEX I data set, 1997. 10

[5] Pierre Chauvet. *Aide-mémoire de Géostatistique Linéaire*. Les Presses de l'École des Mines de Paris, Paris, 1999. 17

[6] J. P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York, 1999. 17, 22

[7] Helge Drange. *An Isopycnic coordinate carboncyclemodel for the North Atlantic: and the possibility ofdisposing of fossil fuel CO_2 in the Ocean*. PhD thesis, Dep. of Mathematics, University of Bergen/Nansen Environmental and Remote Sensing Center, Begen, Norway, 1994. 7

[8] Helge Drange. A 3-dimensional isopycnic coordinate modelof the seasonal cycling of carbon and nitrogen in the North Atlantic. *Phys. Chem. Earth*, 21(5-6), 1996. 7

[9] Geir Evensen. Sequential data assimilation fornonlinear dynamics : The ensemble Kalman filter. In N Pinardi and J-D Woods, editors, *Ocean Forecasting : Conceptual basis and applications*. Springer-Verlag, Berlin, 2002. 6

[10] C. Lantuéjoul. *Geostatistical Simulation: Models and Algorithms*. Springer-Verlag, Berlin, 2001. 17, 20

[11] R.-K. Lowry, R.-M. Downer, and Z. Loncar. OMEX I data set users' guide, 1997. 10, 11

[12] Lars-Jørgen Natvik. *A data assimilation system for a 3-dimensional biochemical model of the North Atlantic*. PhD thesis, Dep. of Mathematics, University of Bergen/Nansen Environmental and Remote Sensing Center, Begen, Norway, 2001. 8, 11, 37

[13] S.V. Semovski and B. Woźniak. Model of the annual phytoplankton cycle in the marine ecosystem-assimilation of monthly satellite chlorophyll data for the North Atlantic and Baltic. *Oceanologia*, 37(1):3–31, 1995. 11

[14] H. Wackernagel. *Multivariate Geostatistics: an Introduction with Applications*. Springer-Verlag, Berlin, 2nd edition, 1998. 17, 22