Vanden Berghe E., M. Brown, M.J. Costello, C. Heip, S. Levitus and P. Pissierssens (Eds). 2004. p. 227-243
Proceedings of 'The Colour of Ocean Data' Symposium, Brussels, 25-27 November 2002
IOC Workshop Report 188 (UNESCO, Paris). x + 308 pp.
– also published as VLIZ Special Publication 16

A hydrographic and biochemical climatology of the Mediterranean and the Black Sea: some statistical pitfalls

Michel Rixen[1], Jean-Marie Beckers[2], Catherine Maillard[3] and the MEDAR Group

[1]Southampton Oceanography Center, European Way
SO14 3ZH, Southampton, United Kingdom
E-mail: myr@soc.soton.ac.uk

[2]GHER B5, Institut de Physique, University of Liège
B-4000 Liege, Belgium

[3]SISMER, IFREMER, Centre de Brest
Boîte Postale 70, F-29280 Plouzané, France

Abstract

The aim of the MEDAR/MEDATLAS II project was to archive and rescue multidisciplinary in-situ hydrographic and biochemical data of the Mediterranean and the Black Seas through a wide cooperation of countries and to produce a climatological atlas of 12 core parameters, which include temperature and salinity, dissolved oxygen, hydrogen sulfure, alkalinity, phosphate, ammonium, nitrite, nitrate, silicate, chlorophyll and pH. Gridded fields have been computed using the Variational Inverse Model and calibrated by Generalised Cross Validation, by making usual assumptions on the statistical distribution of data and errors. They have been produced for both entire Mediterranean and Black Seas and several additional sub-basins including the Alboran Sea, the Balearic Sea, the Gulf of Lions and the Ligurian Sea, the Sicily Strait, the Adriatic Sea, the Aegean Sea, the Marmara Sea and the Danube shelf area at climatic, seasonal and monthly scale when relevant. Inter-annual and decadal variability of T/S for both basins has been computed as well. The resulting atlas is made available free of charge at http://modb.oce.ulg.ac.be/Medar and on CD-Rom. We review here the different biases that might occur when one of the statistical hypotheses is not satisfied, and suggest further possible improvements to the analysis method.

Keywords: Objective analysis; Statistical pitfalls; Mediterranean Sea; Black Sea.

Introduction

The Mediterranean and the Black Sea have been studied actively for several decades (e.g. Brankart and Pinardi, 2001; Karafistan *et al.*, 2002), generating huge amounts of data that remain in danger of being lost if not archived. New multivariate sensors, in growing number and spatio-temporal resolution produce crucial information which require careful but efficient integration. This information is becoming an essential part of the numerous toolkits used by oceanographers, engineers, managers, navies and authorities to continuously monitor the ocean state and variability, to infer possible climate changes, etc. (e.g. Beckers *et al.*, 2002).

The aim of the MEDAR/MEDATLAS II project is to archive and rescue multidisciplinary in-situ hydrographic and biochemical data of the Mediterranean and the Black Sea through a wide cooperation of countries. The growing interplay between scientific disciplines requires multidisciplinary integration of this information which includes, among many other parameters, temperature and salinity, dissolved oxygen, hydrogen sulfur, alkalinity, phosphate, ammonium, nitrite, nitrate, silicate, chlorophyll and pH.

The project was divided into several tasks. First, a global inventory was compiled using existing data sets of the core parameters and their cruise reports from the WDC-A, the WDC-B, the ICES and the MEDATLAS I inventories. Duplicates were eliminated by careful cross-checking. Secondly, the data sets were assembled regionally for the Western, Central and Eastern Region and the Black Sea sub-areas, transcoded in the common ASCII human readable MODB/MEDATLAS format and quality checked by regional experts. Finally, a global integration ensured the assembling and consistency of the regional data sets. The vertical profiles were interpolated on 25 standard vertical levels, chosen according to the vertical distribution of the data (Maillard *et al.*, 2004).

Raw in-situ datasets are difficult to interpret and higher level products are required to offer a more complete and synthetic view of the Mediterranean and Black Sea hydrographic and biochemical systems. The computation of climatological or gridded fields is also justified by the need to provide initial or boundary conditions to numerical models (e.g. Beckers *et al.*, 2001) and is subject to the choice of an adequate analysis method.

In the following sections, we give some details on the objective analysis method, we review potential problems that may appear when the statistical hypotheses of the analysis are not fully satisfied and we provide some insight into the results of the new climatology. Finally, some conclusions and perspectives are drawn.

Methodology

Instead of using the classical objective analysis scheme (OA) (e.g. Bretherton *et al.*, 1976), gridded fields have been computed using the Variational Inverse Model (Brasseur, 1991; Brasseur *et al.*, 1996; Brankart and Brasseur, 1998; Brankart and Pinardi, 2001; Rixen *et al.* 2001), shown to be statistically equivalent to OA.

The basic idea of variational analysis is to determine a continuous field approximating the data and exhibiting small spatial variations. In other words, the target of the analysis f is defined as the smoothest field that respects the consistency with the observed values over the domain of interest. It is also referred to as a spline interpolation method. Expressed in mathematical terms, the analysis is obtained as the minimum of a variational principle (in a two-dimensional, horizontal space):

$$J[\phi] = \sum_{i=1}^{N_d} \mu_i \left[d_i - \phi(\mathbf{x}_i) \right]^2 + \left\| \phi - \phi_b \right\|^2$$
$$\left\| \varphi \right\|^2 = \int \left(\nabla\nabla\varphi : \nabla\nabla\varphi + \alpha_1 \nabla\varphi \bullet \nabla\varphi + \alpha_0 \varphi^2 \right) d\mathbf{x} \qquad (1)$$

The integral extends over the whole domain. The first contribution in J represents the distance between the data and the target field at the exact position of the observations. The weights mi are determined according to the confidence in the data. In principle, the weights could be adjusted to every observation individually, but in practice it is difficult to decide whether one observation is more reliable than another.

The second contribution in J is a measure of the smoothness of the target field. The coefficients a1 and a2 fix the weights of the lower derivatives in the smoothing operator. In practice, all observations at a given level are selected to perform the minimization in a horizontal plane. The reconstruction of a three-dimensional scalar field is then obtained as a superposition of several analyses at different standard depths. This reduction is made possible because the data profiles describe the vertical structure of the sea reasonably well and do not produce hydrostatic instabilities. For bio-chemical data, the structure of profiles can however be more complex.

The reference field ϕ_b (or background field) has been computed by a semi-normed analysis (leaving out the underived term in J) (Brankart and Brasseur, 1996). It has indeed been shown that in data-void areas this method produces more realistic fields than using a simple constant value or a linear regression of the data. The variational principle is solved using a finite element technique. The main advantage is a numerical cost almost independent of the number of data analyzed (Rixen et al, 2001), as compared to optimal interpolation. The mesh also easily takes into account the complexity of the basin geometry by automatically prohibiting correlations across land barriers (Brankart and Brasseur, 1998).

By converting the dimensional weights of the variational principle into non-dimensional quantities, and imposing that the correlation function be the Modified Bessel function of order 1, the calibration of the free parameters m, a1 and a2 reduces to the choice of a characteristic length scale L and a signal-to-noise ratio (S/N=e2/s2), which represents the ratio of the standard deviation of the signal e to the standard deviation of observational errors s. A Generalized Cross-Validation (GCV) technique has been implemented to compute the statistical parameters L and S/N out of the raw data (Brankart and Brasseur 1996). The method consists in successively eliminating one measurement from the full database and performing analyses with the remaining data. The variance of the misfits between these reconstructed fields and the corresponding eliminated data is then considered as a statistical indicator of the quality of the analysis, with respect to S/N and L. In practice, for climatological analyses, the quality of the analysis is not very sensitive to the correlation length and has thus been fixed a priori, according to a reasonable choice of features to be represented in the domain of interest.

By analogy with OA, the statistical error field is obtained by using the stiffness matrix K of the analysis and two transfer operators T1 and T2 (Rixen *et al.*, 2001)

$$e^2(x) = \varepsilon^2(x) - T_1(x)K^{-1}T_2(x)c(x) \qquad (2)$$

where c is the correlation function, i.e. the Modified Bessel function of order 1.

Statistical pitfalls and related solutions

The variational principle in eq. (2) relies on several statistical assumptions. Among others, it assumes that the frequency distributions of the noise and the signal are gaussian [assumption 1], that the statistics are homogeneous and isotrope [assumption 2] and that the noise is uncorrelated [assumption 3]. Asssumption 1 is generally speaking related to the amount of available information as an observed phenomenon will usually show a gaussian frequency distribution as long as the amount of data is sufficient. Assumption 2 is generally valid except the fact that hydrodynamics usually stretch the features along circulation paths. This problem can be easily circumvented by adding an advection constraint into eq. (2) for example (e.g. Brankart and Brasseur, 1996). This problem occurs frequently in coastal areas and has to be dealt with carefully. Assumption 3 is also generally valid unless they are systematic biases in the measurements that can be more easily detected.

Problems related to assumption 1

In Figs. 1 and 2, we show, respectively the distribution of temperature data in the Alboran basin and in the Ionian Sea at 200m over a 1°x1° square box. Note that even at 200m the data distribution is not strictly gaussian for the Ionian case. This problem may then lead to biases in the background field and hence in the signal itself and subsequently in the solution.

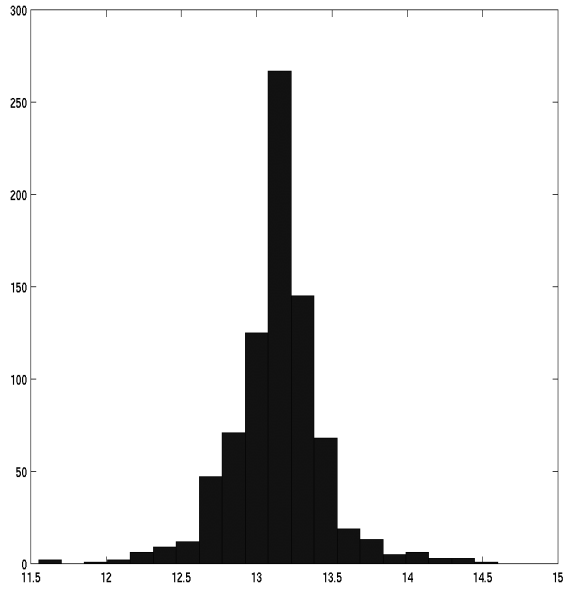

Fig. 1. Distribution of temperature data in the Alboran basin at 200m over a 1°x1° square box. Note the gaussian shape of the frequency distribution.

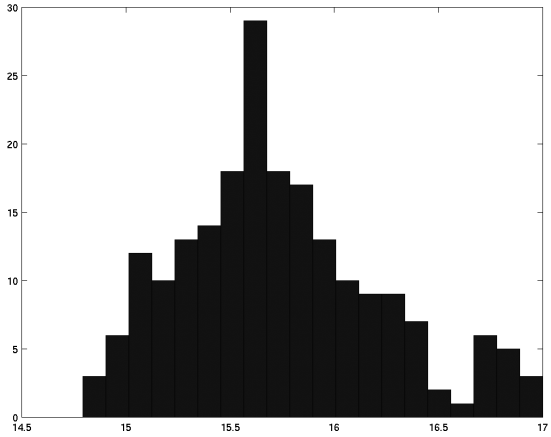

Fig. 2. Distribution of temperature data in the Ionian basin at 200m over a 1°x1° square box. Note the bi-modal shape of the frequency distribution.

A second problem that can appear is the weak spatial coverage in certain areas, where the solution field has to be extrapolated either from adjacent available stations or eventually from upper/lower available values when a 3D interpolation is considered. For the salinity for example, although the coverage including all seasons is very good (see Fig. 3), there might be areas void of data for specific months, even at the surface. In the case of silicate for example (see Fig. 4), the distribution is very patchy and concentrated in areas of interest, like the Gulf of Lions, the Ligurian Sea, the Gibraltar Strait, … but the areas in between have to be extrapolated from these patches.

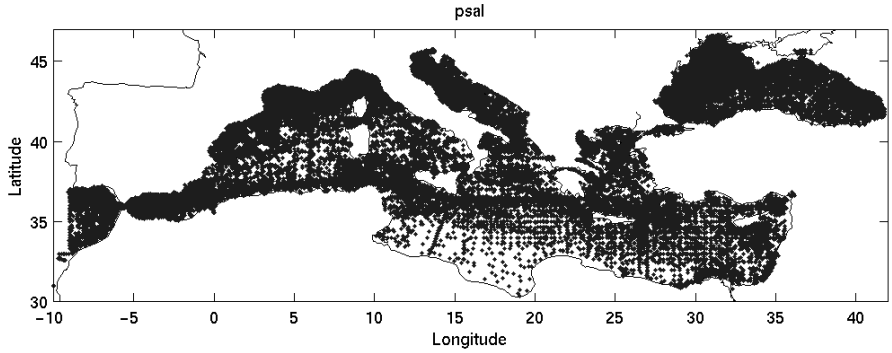

Fig. 3. Station distribution of salinity. Note the nice data coverage except along the Tunisian coast. The monthly distribution (not shown) is however more patchy.

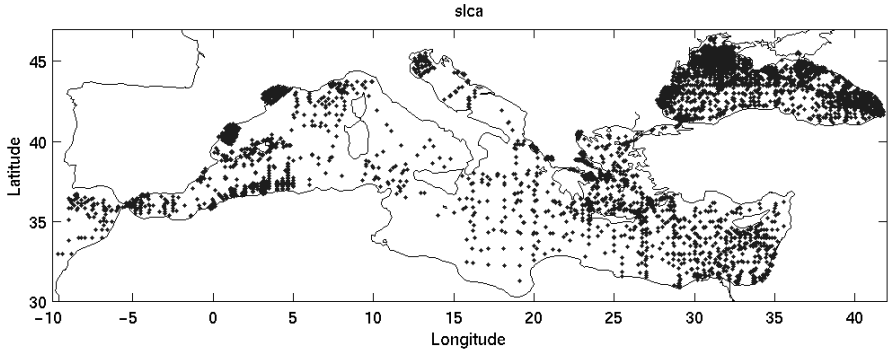

Fig. 4. Station distribution of silicate. Note the patchiness of the distribution and some large areas void of data where extrapolation will be needed in the analysis.

On the vertical, the availability of data rapidly decreases from the surface towards the bottom as seen from Fig. 5 for temperature for example. As the typical correlation length in the Mediterranean varies between 100 and 500km, a homogeneous distribution would require at least ~1200 to 50 data for each horizontal level. As the geometry of basin is very complex, these figures are however surely too optimistic.

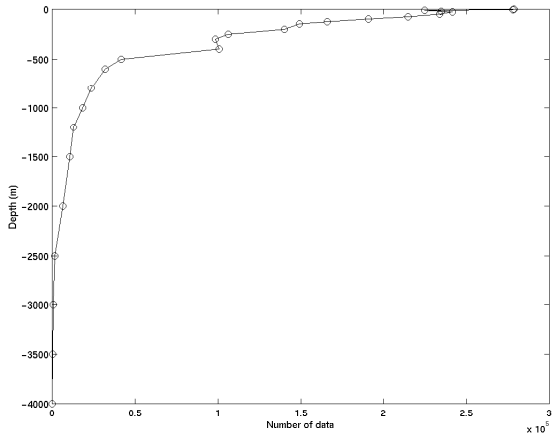

Fig. 5. Number of temperature data as a function of depth. Note that although the number of stations at the surface is impressive, it rapidly decreases towards the bottom.

The MEDAR database includes data over the 1900-2002 period, but most of the data have been collected after 1945 (see Fig. 6). A climatic analysis will then not properly represent the mean state of the century but most likely a state close to the period referring to the second half of the century.

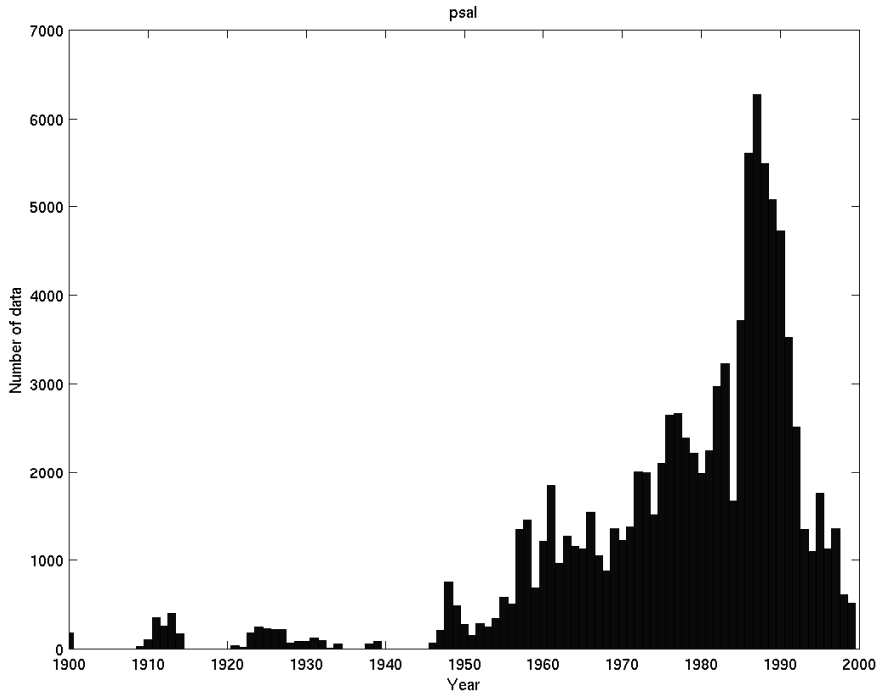

Fig. 6. Yearly distribution of salinity profiles from 1900 to 2000. Note the data-void periods during the two world wars. Note also the constant increase in available data during the second half of the century and the decrease of the availability from 1990 on due to the delay of releasing the information.

For some parameters, the distribution over months is not uniform, and seriously biased towards high summer frequencies and low winter frequencies (see Fig. 7 for example). A climatic temperature analysis for example may then bias the mean towards higher temperature locally.

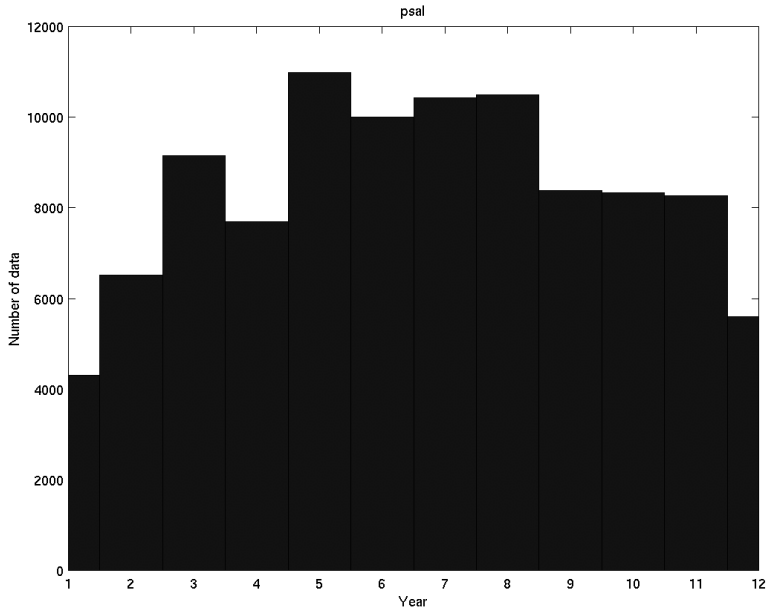

Fig. 7. Monthly distribution of salinity profiles. There are more data available in summer than in winter.

For the Ionian sea, temperature has been sampled during the following years mainly during the following months, where '1' stands for 'January', '2' for February, etc.:

- 1980 Months 2
- 1986 Months 3, 4, 9
- 1988 Months 7
- 1990 Months 10, 11
- 1992 Months 5
- 1994 Months 1

Similarly, for the Levantine basin, the available data are:

- 1984 Months 10
- 1986 Months 8, 9, 10, 11
- 1988 Months 3, 8, 9
- 1990 Months 7, 10, 11
- 1994 Months 1, 2

In both cases, the available data could result in the analysis in an increase of temperature for the period 1980-1988 followed by a cooling period during 1990-1994.

A final possible bias is illustrated in Fig. 8 where we show a vertical section for ammonium, a parameter whose values are at the limit of instrumental errors and for which very few data are

available. In this case the 2D horizontal analysis results in a 'layered' solution, although the original profiles are smoother.

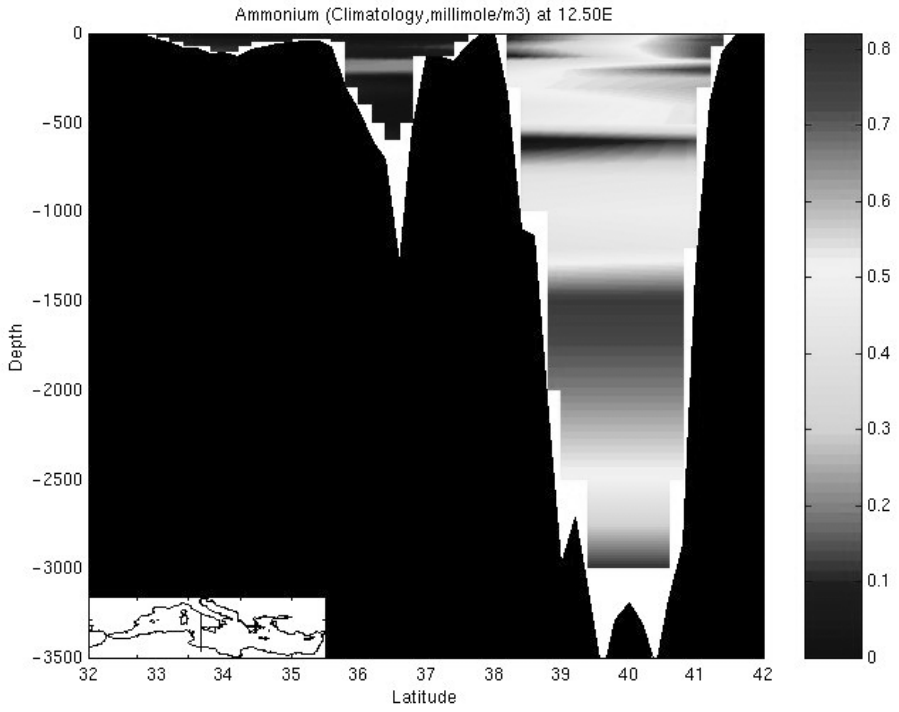

Fig. 8. Vertical section of ammonium in the Tyrrhenian Sea: an illustration of a possible bias due to the horizontal analysis approach. An obvious improvement would be to extend the analysis scheme to a 3D method.

Problems related to assumption 2

For the preliminary version of the climatology, a preprocessing quality check (QC) was applied to the raw data in order to eliminate values out of 3 standard deviations on bins 5-10 times bigger than the analysis grid. This method provided a final QC on the data and ensured that most potential bulls' eyes were removed in the analyses when the statistical distribution of data was typically gaussian. This was surely the case for parameters like T or S at levels close to the surface. However, at deeper levels or for parameters less frequently sampled, this pre-processing was rejecting a significant amount of useful information when the statistical distribution was not strictly gaussian. As the major discrepancies were only found for coastal areas or areas with shallow waters, this preprocessing was not considered anymore.

Coastal data are sometimes the only information available but they might be significantly affected by coastal processes, including resuspension of sediments and anthropogenic activities. In preliminary versions of the climatology, coastal data have been used and have led to misinterpretation of the resulting gridded fields, especially regarding nutrient parameters. There

is thus a difficult compromise to be made between the use of all data and non-coastal data only. Coastal data closer than 15 km from the coast or in areas shallower than 50m were rejected. These values have been chosen a priori but correspond roughly to the horizontal and vertical radius of influence of land and bottom respectively, when focusing on basin and regional processes. It was effectively shown that this preprocessing was far more efficient in removing artificial features in the analyses.

In Fig. 9, we show the station distribution for phosphate at 30m depth. The 'big dots' stations (3317) have been rejected by the pre-processing either because they are closer than 15km from the coast, either because the associated sounding depth is shallower than 50m. The 'small dots' stations (12261) are those that are effectively used in the analysis. In this example, rejection occurs mainly along Algeria, Spain, Egypt and Israel, around the Balearic islands, in the Aegean and Alboran Sea and in the Northern Adriatic. The impact of this pre-processing is shown in Fig. 10. On the top figure, the phosphate field using all available data shows clear biases in the analysis with high values extending anomalously far from the coast. Indeed, the interpolation 'spreads' these errors off-shore by the correlation function influence. The analysis of phosphate without the rejected data (middle figure) shows a clear reduction of these biases. The differences between the two fields (bottom figure) are located in areas where the data have been rejected. The corrections are particularly visible in the Almeria-Oran area, in the Adriatic, the Aegean, around the Balearic island, and along Egypt/Israel and Spain. But one notes also significant changes in the open ocean. For temperature or salinity, the artificial influence of rivers or other local processes is also reduced. This pre-processing thus prevents coastal data from creating artificial features far from the coast.

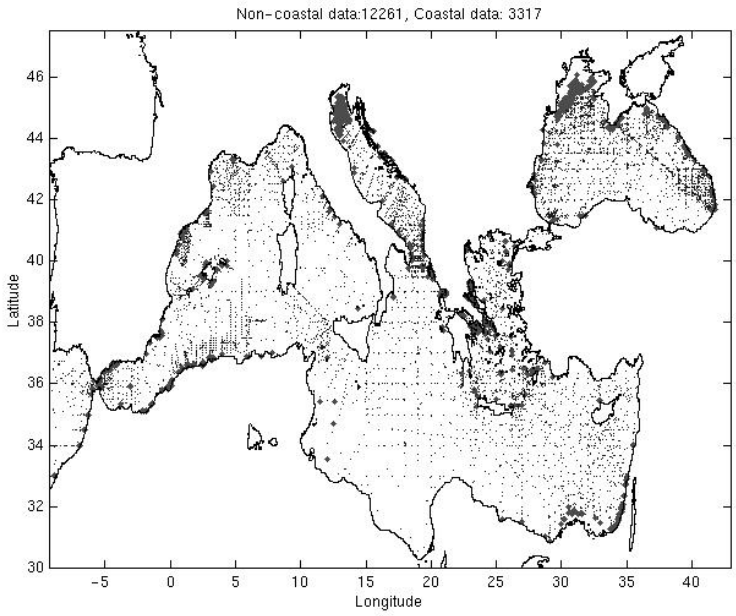

Fig. 9. Station distribution of phosphate at 30m depth used in the analysis (small dots). Coastal data and data with sounding shallower than 50m (big dots) have been rejected by the pre-processing.

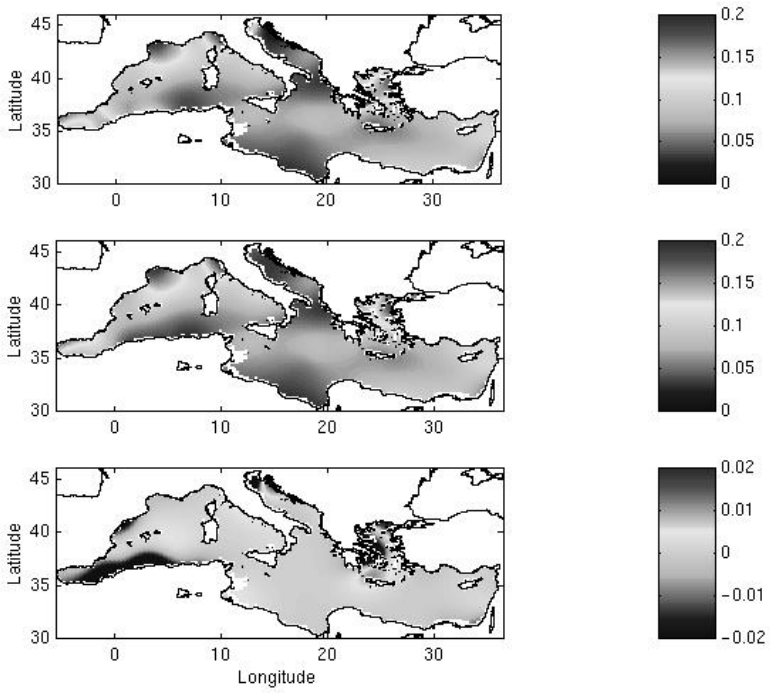

Fig. 10. Phosphate field (millimole/m3) at 30m depth. Top: all available data are used. Middle: analysis without coastal data. Bottom: difference between both analyses.

Problems related to assumption 3

Obvious instrumental error can sometimes be easily detected as seen for example in the T/S diagram (Fig. 11) for the Ionian Sea in the area 36-37ºN, 19-20ºE. Unless these data are flagged properly, that will induce serious biases in the analysis.

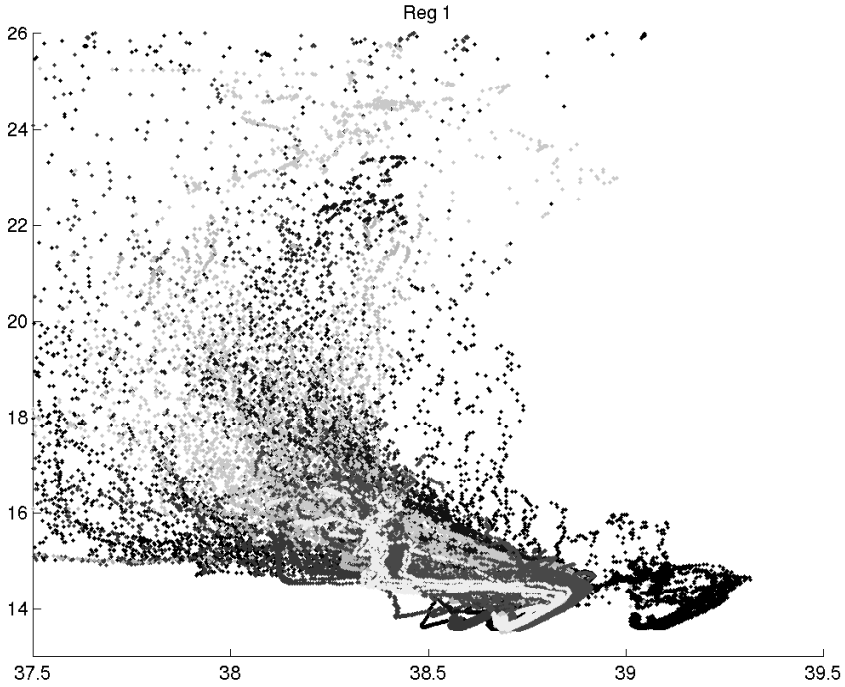

Fig. 11. T/S diagram for the Ionian Sea in the area 36-37 ºN, 19-20ºE. The different gray scales correspond to different field experiments. Note the clear shifts in salinity due to calibration problems. Most of these problems have been identified and corrected for the final MEDAR database.

Results

Gridded fields have been produced for both the Mediterranean and Black Sea basins and several additional subbasins including the Alboran Sea, the Balearic Sea, the Gulf of Lions and the Ligurian Sea, the Sicily Strait, the Adriatic Sea, the Aegean Sea, the Marmara Sea and the Danube shelf area at climatic, seasonal and monthly scale when relevant. Inter-annual and decadal variability of T/S for both basins has been computed as well. It should be noted that for a given temporal scale, all data have been used, but the observational error standard deviation has been modified according to Brankart and Pinardi (2001), by imposing a temporal correlation length: 3 months for a seasonal analysis, 1 month for a monthly analysis, 10 years for a decadal analysis and 1 year for an annual analysis.

In practice, the usual statistical hypotheses of OA are not fully satisfied, either because some (potentially big) areas are void of data, or because the distribution of the signal and/or noise is not gaussian (except for T/S and maybe dissolved oxygen). Therefore, the spatial correlation length has been increased artificially for consistency of the gridded fields at basin and regional scale. As a result, a posteriori, only gridded fields that have been computed by 'robust' data distributions have been selected. They include:

- temperature and salinity at climatic, seasonal and monthly scale
- dissolved oxygen, phosphate and silicate at climatic and seasonal scale
- nitrate, nitrite, alkalinity, pH, ammonium at climatic scale

The resulting atlas is made available free of charge at 'http://modb.oce.ulg.ac.be/Medar' and on CD-Rom, where the fields are found in Netcdf format (compressed with the gzip tool), and figures in Jpeg format. Results are organized in a 'tree' way: first identify the domain, then the parameter, then the relevant period. To illustrate the climatology, some examples of the Mediterranean and the Black Sea are shown below. For the corresponding color figures, the reader is invited to visit the web site.

Figs. 12 and 13 show horizontal sections of climatological temperature and alkalinity at 100m depth in the Mediterranean, with a clear signature of the Mediterranean inflow, the Alboran Western and Eastern gyres, the gradient of temperature from west to east, the Rhodes gyre, the outflow of the Black Sea. The continuous evaporation taking place over the Mediterranean, together with the regional topography and local intermittent formation of deep water and intermediate water are responsible for these features, well reported in the literature.

Fig. 14 and 15 show horizontal sections of climatological salinity and oxygen at 50m depth in the Black Sea, with a clear signature of the typical dipole cyclonic circulation, and the saltier and less oxygenized in the core of these cyclones. This general circulation of the Black Sea and the associated distribution of bio-geochemical parameters are resulting from the atmospheric forcing over the basin.
Provided physical and biochemical processes are sampled adequately, i.e. with enough accuracy in time and space, they will appear in the analyzed fields. The examples shown are of course only a tiny fraction of exploitable results of the climatological atlas that has been produced. The whole climatology currently represents some 2000 fields, 20000 figures and 600 animations.

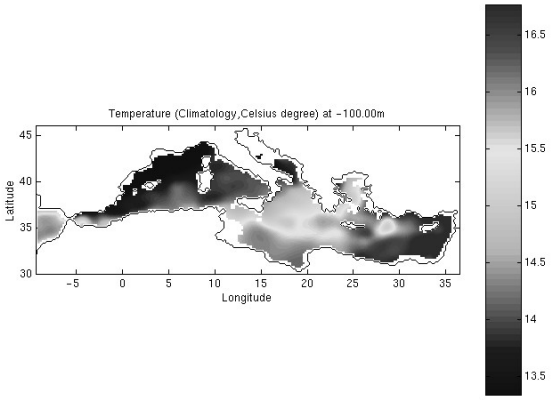

Fig. 12. Climatological temperature in the Mediterranean at 100m depth.

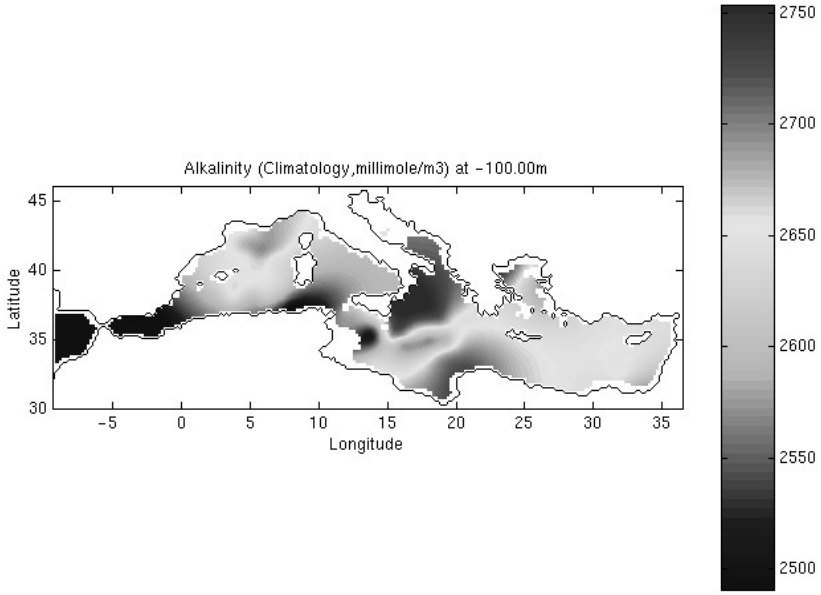

Fig. 13. Climatological alkalinity in the Mediterranean at 100m depth.

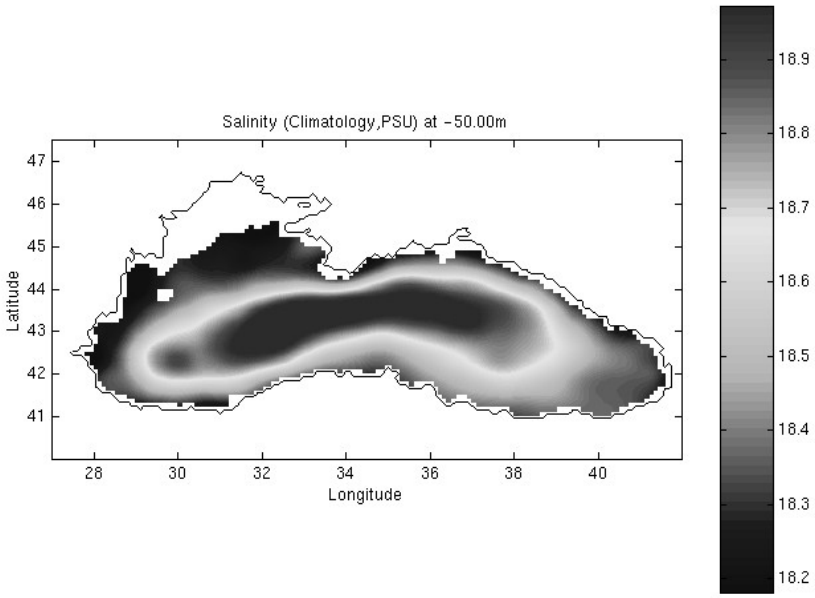

Fig. 14. Climatological salinity in the Black Sea at 50m depth.

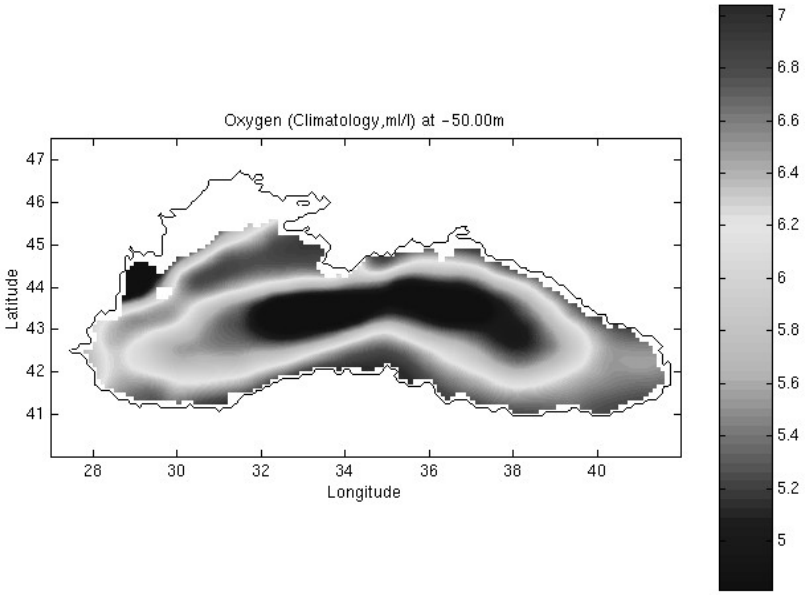

Fig. 15. Climatological oxygen in the Black Sea at 50m depth.

Conclusions

In this paper, we have described the methodology used in the gridding task of the MEDAR/MEDATLAS project. The critical technical choices made during the project have been justified following a constant feedback between data providers, regional centers, the global assembling center and the data analysis center. This feedback loop, actually 3 rounds, has allowed to refine the pre-processing steps and the data analysis technique in order to enhance the quality of the climatology throughout the project.

The quality of the analyses is highly dependent upon the validity of the underlying statistical hypotheses. In the future, more powerful interpolation schemes should investigate the usefulness of multivariate and/or 3D interpolation schemes that effectively take into account the vertical correlation present in the profiles of raw parameters. Systematic decimation of the data has not been possible as this method might result in other biases when the data coverage is very weak.

The issue of coastal data is particularly critical, especially for biochemical parameters, as they reflect anthropogenic activities. The scale of influence of the related processes must thus be correctly mapped. For this purpose, the rejection of coastal data and data with shallow sounding is particularly useful, as shown by the changes reflected in the analyses.

Some results have been presented and we expect that both the datasets and the gridded fields produced by state-of-the-art statistical techniques will generate a new era of scientific

investigation of the Mediterranean and the Black Sea, including investigation on the general circulation, on possible climate changes, on anthropogenic impacts on the seas, on regional eco-hydrodynamic processes, ...

Acknowledgements

We gratefully acknowledge the organizers and sponsors of 'The Colour of the Ocean Data' symposium. Many thanks to all data providers over the Mediterranean and the Black Sea. The FNRS provided valuable supercomputing assistance. The first author currently benefits from a EC Marie Curie Fellowship at the Southampton Oceanography Center.

References

Beckers J.-M., M. Rixen, P. Brasseur, J.-M. Brankart, A. Elmoussaoui, M. Crepon, C. Herbaut, F. Martel, F. Van den Berghe, L. Mortier, A. Lascaratos, P. Drakopoulos, P. Korres, N. Pinardi, E. Masetti, S. Castellari, P. Carini, J. Tintore, A. Alvarez, S. Monserrat, D. Parrilla, R. Vautard, and S. Speich. 2002. Model intercomparison in the Mediterranean. The MedMEx simulations of the seasonal cycle. Journal of Marine Systems 33-34:215-251.

Brankart J.-M. and P. Brasseur. 1996. Optimal analysis of in situ data in the Western Mediterranean using statistics and cross-validation. Journal of Atmospheric and Oceanic Technology 13(2):477-491.

Brankart J.-M. and P. Brasseur. 1998. The general circulation in the Mediterranean Sea: a climatological approach. Journal of Marine Systems 18:41-70.

Brankart J.-M. and N. Pinardi. 2001. Abrupt cooling of the Mediterranean levantine intermediate water at the beginning of the 1980s: observational evidence and model simulation. Journal of Physical Oceanography 31:2307-2320.

Brasseur P. 1991. A Variational Inverse Method for the reconstruction of general circulation fields in the Northern Bering Sea. Journal of Geophysical Research 96(C3):4891-4907.

Brasseur P., J.-M. Beckers, J.-M. Brankart and R. Schoenauen. 1996. Seasonal temperature and salinity fields in the Mediterranean Sea: Climatological analyses of an historical dataset. Deep-Sea Research 43(2):159-192.

Brasseur P. and J. Haus. 1991. Application of a 3-D variational inverse model to the analysis of ecohydrodynamic data in the Northern Bering and Southern Chuckchi seas. Journal of Marine Systems 1:383-401.

Bretherton F. P., R.E. Davis and C.B. Fandry. 1976). A technique for objective analysis and design of oceanographic experiment applied to MODE-73. Deep-Sea Research 23:559-582.

Karafistan A., J.-M. Martin, M. Rixen and J. Beckers. 2002. Space and time distributions of phosphates in the Mediterranean sea. Deep-Sea Research 49(1):67-82.

Maillard C., E. Balopoulos and The MEDAR Group. 2004. Recent advances in oceanographic data management of the Mediterranean and Black Seas: the MEDAR/MEDATLAS 2002 database. p. xxx-xxx. In: Vanden Berghe E., M. Brown, M.J. Costello, C. Heip, S. Levitus, P. Pissierssens (eds). Proceedings of 'Colour of Ocean Data' Symposium, Brussels, 25–27 November 2002. Paris, UNESCO/IOC, VLIZ, 2004. IOC Workshop Report 188 – VLIZ Special Publication 16.

Maillard C., M. Fichaut, G. Maudire, C.R. Coatanoan, B. Boudjelall, N. Eddalia, J.-M. Beckers, M. Rixen, G. Kortchev, G. Zodiatis, H. Dooley, S. El-Agami, I. Oliounine, E. Balopoulos, A. Iona, A. Lykiardopoulos, P. Karavregekis, I. Gertman, Y. Tsehtik, B. Manca, A. Giorgetti, A. Mosetti, G. Manzella, N. Pinardi, M. Zavatarelli, S. Lakkis, A. Drago, A. Orbi, J. Larissi, S. Zizah, N. Mikhailov, E. Vyazilov, A. Kuznetov, N. Puzova, M.-J. Garcia, M. Ozyalvac, F. Berkay, A. Suvorov, A. Khaliulin, R.

Gelfled, C. Sammari, V. Diaconu, K.Z.S. Bilashvili and D. Vlado. 2002. MEDAR/MEDATLAS 1998-2001: A Mediterranean and Black Sea oceanographic database and network. Bollettino di Geofisica Teorica ed Applicata. (submitted).

Rixen M., J.-M. Beckers, J.-M. Brankart and P. Brasseur. 2001. A numerically efficient data analysis method with error map generation. Ocean Modelling 2(1-2):45-60.