Detection of wrong analysis with Diva

Charles ctroupin@ulg.ac.be

December 7, 2011

Abstract

Analysis are performed on a common data set with different parameter values (correlation length L and signal-to-noise ratio λ). The goal is to provide examples of wrong analysis that can occur when the parameters are not correct, so that their identification is made easier for other applications.

Contents

1 The Diva method

Considering a series of N data anomalies d_i at locations (x_i, y_i), Diva minimises the cost function (here in Cartesian coordinates) over the domain Ω:

$$\begin{aligned} J[\varphi] &= \sum_{i=1}^{N} \mu_i \left[d_i - \varphi(x_i, y_i)\right]^2 \\ &+ \int_{\Omega} \left(\boldsymbol{\nabla}\boldsymbol{\nabla}\varphi : \boldsymbol{\nabla}\boldsymbol{\nabla}\varphi + \alpha_1 \boldsymbol{\nabla}\varphi \cdot \boldsymbol{\nabla}\varphi + \alpha_0 \varphi^2\right) \mathrm{d}\Omega, \end{aligned} \tag{1}$$

where

- α_0 fixes the length scale L for which the first and the last term of the integral in (1) have a similar importance:
$$\alpha_0 L^4 = 1. \tag{2}$$

- α_1 fixes the influence of gradients:
$$\alpha_1 L^2 = 2\xi, \tag{3}$$
where $\xi = 1$ in this implementation of Diva.

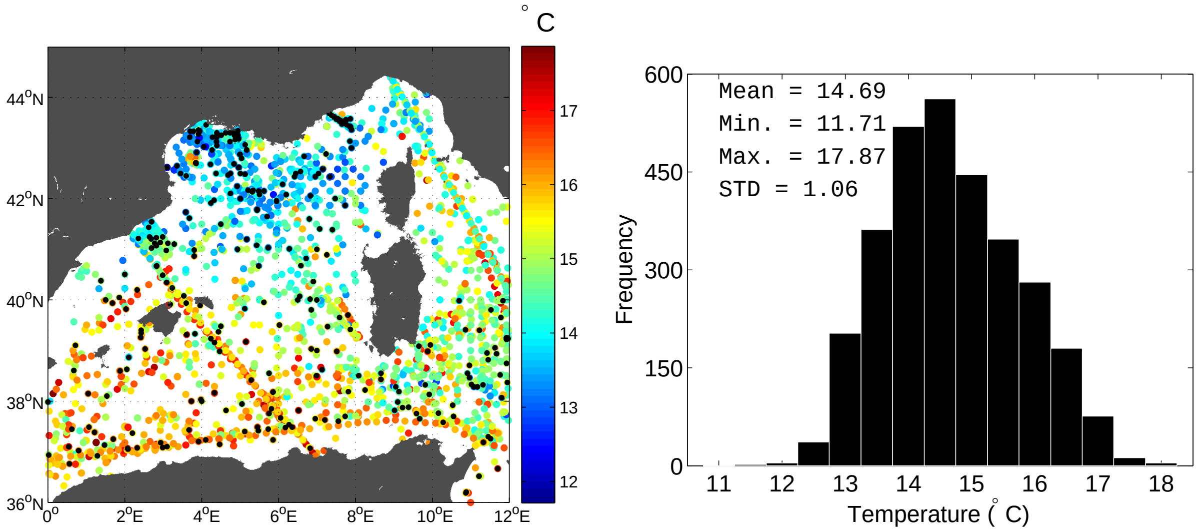
- μ_i fix the weights on the individual observations and is related to the signal-to-noise ratio:.
$$\mu_i L^2 = 4\pi\lambda. \tag{4}$$

$\boldsymbol{\nabla}$ denotes horizontal gradients, and the symbol : stands for double summation[1]. More details concerning the method can be found in the bibliography.

[1] $\boldsymbol{\nabla}\boldsymbol{\nabla}\varphi : \boldsymbol{\nabla}\boldsymbol{\nabla}\varphi = \sum_i \sum_j (\partial^2\varphi/\partial x_i \partial x_j)\,(\partial^2\varphi/\partial x_i \partial x_j)$, the generalisation of the scalar product of two vectors.

2 Data

The data set is made up of temperature measurements between 0 and 5 m depth, during the month of April, in the period 1985-2005. It contains 3372 data points, distributed as shown in Fig. 1.



(a)

(b)

Figure 1: (a) Spatial distribution of the temperature measurements. Black dots indicate the measurements set apart for the validation and (b) histogram of the measurements.

2.1 Validation data set

Out of the 3372 data points, 10%, randomly selected, were set apart in order to perform a validation after the analysis. (Fig. 1).

2.2 Statistics

The temperature measurements range between 11.71°C and 17.87°C (Fig. 1b).

3 Analysis parameters

We will limit ourselves to two parameters:

- The correlation length L, which measures distance over which a data point influence its neighbours.
- The signal-to-noise ratio λ, which measures the confidence we have in the measurements

The output grid covers the region shown in Fig. 1 with a resolution of 0.1° × 0.1° (121 × 91 grid points).
A set of parameter files (param.par) is generated from lists of L and λ values (Tab. 1) using the bash script prepare_paramfiles. 14 values of L and 13 values of λ were tested, yielding a total number of 182 cases.

Table 1: Lists of parameters tested for the analysis.

L	0.05	0.1	0.2	0.3	0.5	0.75	1	2	3	5	10	20	30	50
λ	0.01	0.05	0.1	0.3	0.6	1	3	6	10	30	60	100	300	

```
# Correlation Length lc in km or degree??? according to param icoordchange
1
# icoordchange (=0 if position of data in km ; =1 if position of data in degree)
2
# ispec (output files required, comments to come)
0
# ireg
2
# xori (origin of output regular grid, min values of X)
0
# yori (origin of output regular grid, min values of Y)
36
# dx (step of output grid)
0.1
# dy (step of output grid)
0.1
# nx max x of output grid
121
# ny max y of output grid
91
# valex (exclusion value)
-99
# snr signal to noise ratio
1
# varbak variance of the background field 2.5
0
```

Example file 3.1: Base parameter file.

4 Results

For each analysis performed, we will show:

(a) The analysed field obtained with the given of parameters (L, λ).

(b) The differences between the validation data set (back dots in Fig. 1) and the reconstructed values at these points. They will be referred to as misfits.

(c) The histogram for the analysed field, to be compared with the histogram of Fig. 1b.

Out of the 182 cases, we select four combinations of (L, λ) leading to analysis with problems easy to spot. The results are presented together in Fig. 2. The range for the color scales are the same for the corresponding figures.

4.1 Signal-to-noise ratio is too low $\boxed{L = 1^\circ \textbf{ and } \lambda = 0.01}$

(a) Analysis: the field is very smooth, meaning that the regularisation constraint dominate the data influence. The field only exhibits a meridional gradient of temperature.

(b) Misfits: nothing particular to observe.

(c) Histogram: the distribution is tighter than with the original measurements (Fig. 1b). The mean value is slightly higher, this difference is probably due to the spatial distribution of the data. The standard deviation decreased of almost 40%.

Such an analysis is acceptable when very few data are available, or if the considered period is long.

4.2 Correlation length is too small $\boxed{L = 0.1^\circ \textbf{ and } \lambda = 3}$

(d) Analysis: the field is noisy and cruise tracks can be observed (compare with data positions in Fig. 1).

(e) Misfits: again, nothing particular, except that the RMS is decreased with respect to the previous case.

(f) Histogram: as expected, the range of values and the variance is higher than in the previous case, while the mean value is similar. Most of the field values are not between 14°C and 16°C.

In this example the signal-to-noise ratio was set to 3, but analysis with lower values for λ yielded results with similar features. Such an analysis would be acceptable when dealing with measurements taken over a short period (a few hours or days, for instance satellite data). Note that it is reasonable to select a value of L not smaller than the mean distance between data.

4.3 Correlation length is too large $\boxed{L = 50^\circ \textbf{ and } \lambda = 3}$

(g) Analysis: the field is very smooth and similar to the analysis with a low λ.

(h) Misfits: nothing particular to observe.

(i) Histogram: again, the distribution is tighter and similar to Fig. 2c.

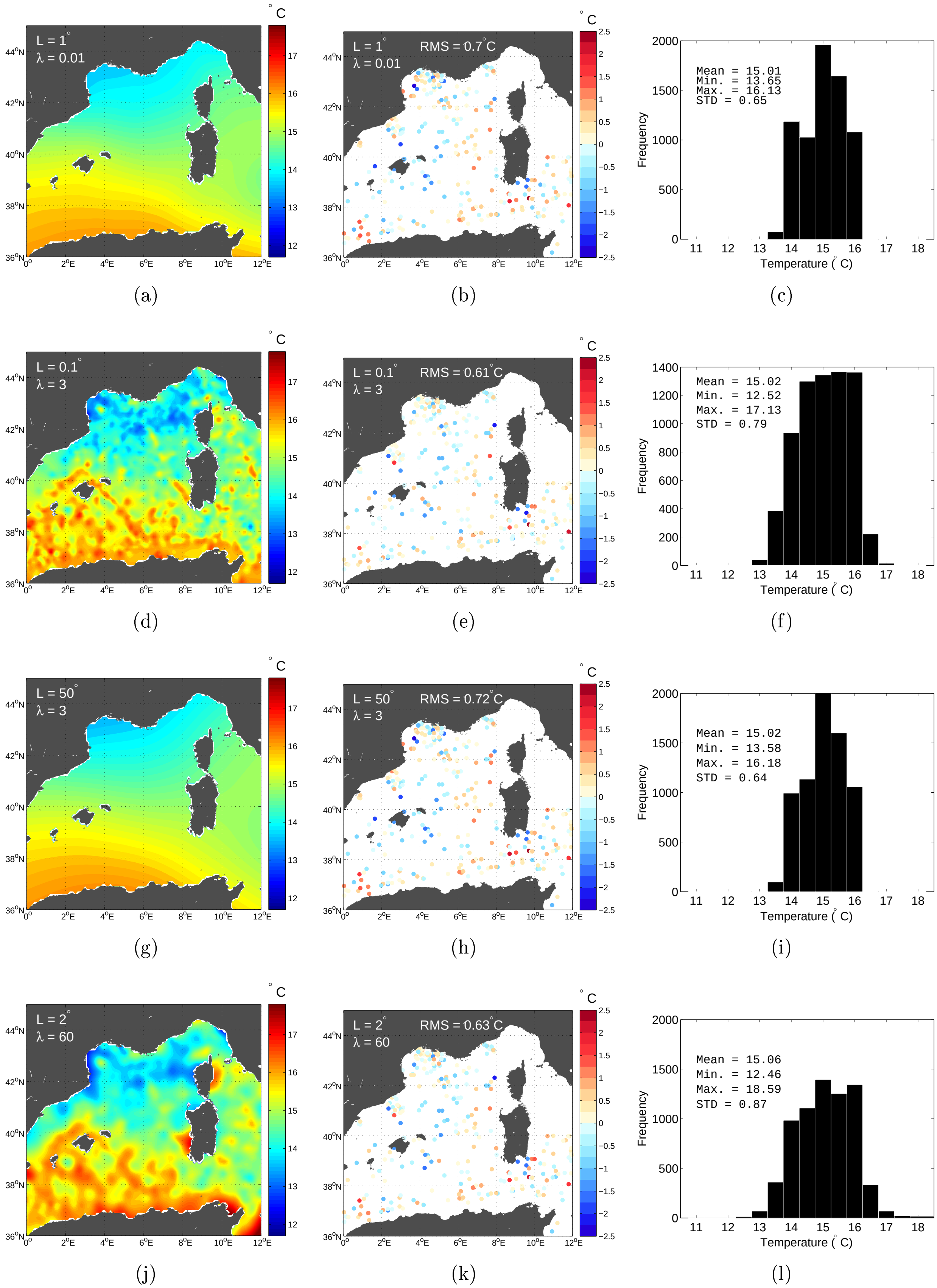
4.4 Signal-to-noise ratio is too high $\boxed{L = 2^\circ \textbf{ and } \lambda = 60}$

(j) Analysis: similar to the case with a small L, except that here the cruise tracks are less visible.

(k) Misfits: the RMS is lower than in previous cases.

(l) Histogram: the distribution is closer to the original one, with a wide range of values and a higher variance. We can also note a distribution different from the low-L case (Fig. 2f);

Note that this case may be difficult to distinguish from the low-L case. For particular combinations of parameters, the analysis are very similar.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 2: First column: analysis, second column, misfit, third column, histogram.

5 Estimation of the parameters with Diva tools

In most cases, the tools associated with Diva (divafit, divagcf, divacv, divacvrand) provide suitable values for the analysis parameters. However, the quality of the results given by these tools depends on:

- The number of data available.
- The spatial coverage of the data.
- The dependence of the measurements.

Hence the values of L and λ provided by the corresponding tools do not always constitute the best option to work with, since they can generate analysed fields with with problems similar to those described in Section 4.

5.1 Correlation length

It is evaluated by fitting the theoretical kernel of the second term of (1) to the correlation between data. Applied to our dataset, the fit (Fig. 3) gives the value $L = 1.15°$ ($\approx$ 128 km).

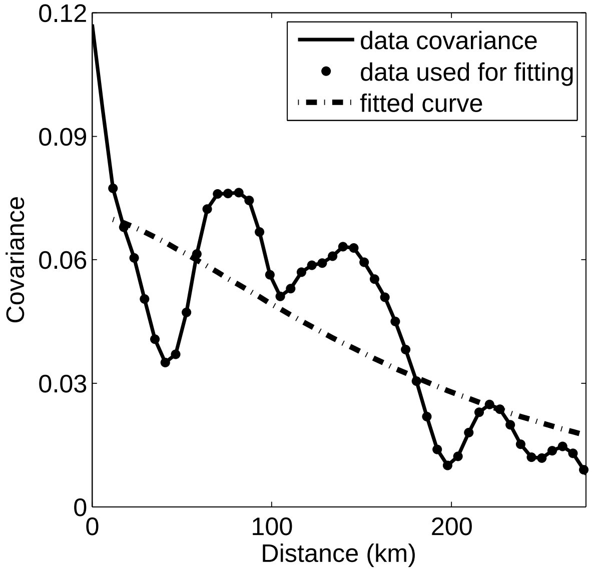


Figure 3: Fit of the data correlation to the theoretical kernel (dashed line).

5.2 Signal-to-noise ratio

The signal-to-noise ratio λ is estimated with ordinary or generalised cross-validation methods. The idea is to test various values for λ and compute an estimation of the global analysis error though the generalised cross-validator Θ. The values of λ that minimises Θ will be then provided.
Three variations of the method are available:

1. `divagcv` performs a generalized cross-validation (GCV).
2. `divacv` performs an ordinary cross-validation (CV).
3. `divacvrand` performs a repeated ordinary cross-validation on a sub-sample of the measurements chosen randomly (CVRAND). Here we repeated 10 times the cross-validation on 300 randomly selected data points.

The results obtained with three different methods are presented in Fig. 4 (note the semi-logarithmic scale) and are summarised in Tab. 2. CV and CVRAND yield close values, while GCV gives a larger estimate for λ.
More generally, we have to take into account that:

1. The three methods may provide significantly different values for λ.
2. For a given method, there is always a range of λ values for which the value of Θ does not vary much.

Table 2: Generalized cross validator (Θ) for a set of λ values.

Method	Θ	λ
GCV	0.666	32.424
CV	0.697	9.187
CVRAND	0.676	11.165

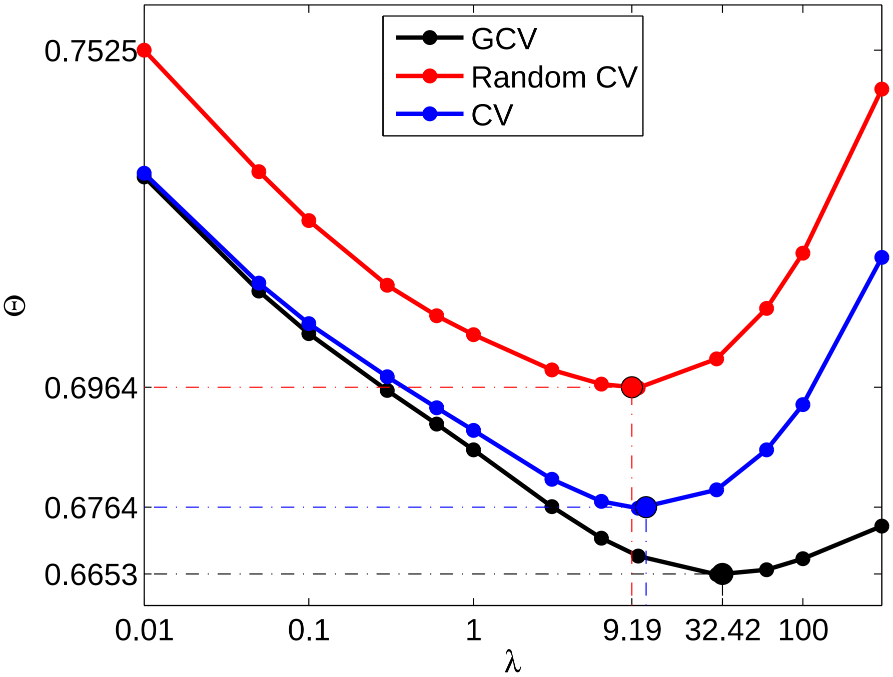


Figure 4: Generalized cross validator (Θ) for a set of λ values.

The analysis obtained with parameter values similar to those provided by the Diva tools is shown in Fig. 5. The case is similar to Fig. 2f, corresponding to an analysis with $L = 2°$ and $\lambda = 60$. Remark that the fields obtained using 10 or 30 for λ are similar.

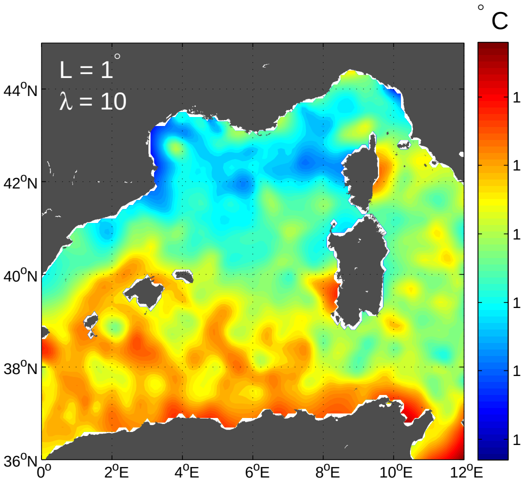


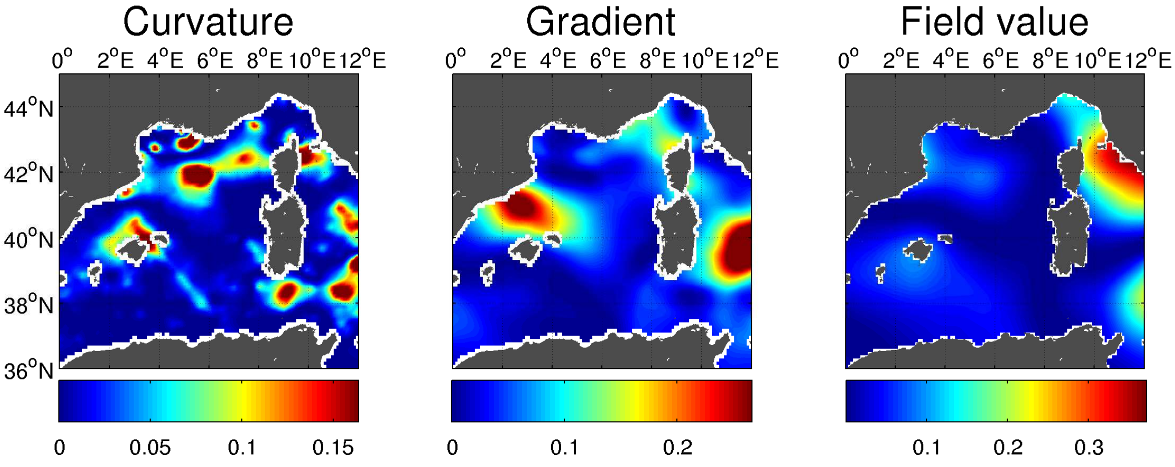
Figure 5: Analysed field obtained with $L = 1°$ and $\lambda = 10$.

6 Calculation of the Diva norm

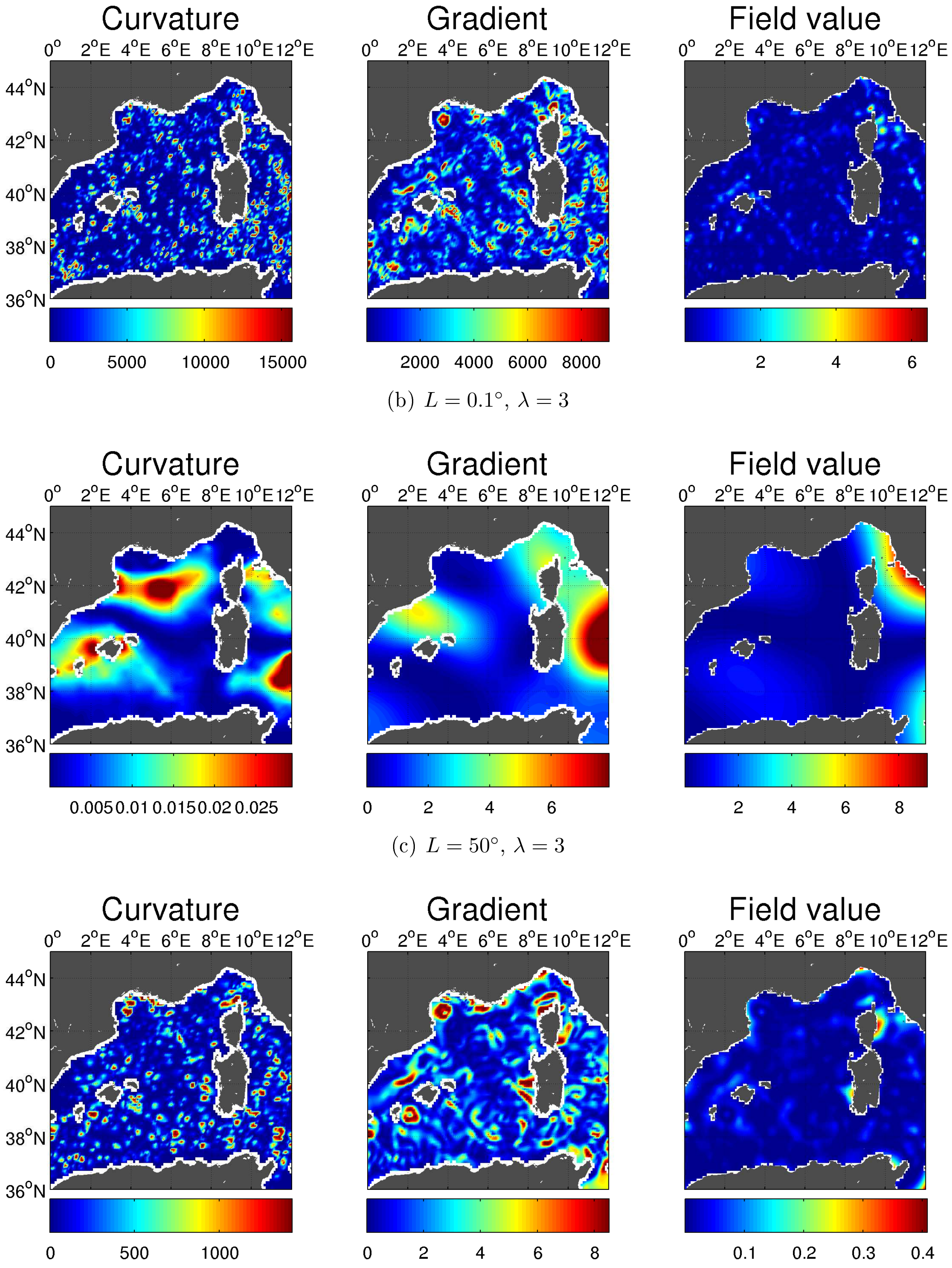
The integral term of (1) is made up of three contributions, referred to as:

1. The curvature: $\boldsymbol{\nabla}\boldsymbol{\nabla}\varphi : \boldsymbol{\nabla}\boldsymbol{\nabla}\varphi$,
2. the gradients: $\alpha_1 \boldsymbol{\nabla}\varphi \cdot \boldsymbol{\nabla}\varphi$ and
3. the field value: $\alpha_0 \varphi^2$.

We will now represent these three fields for different combinations (L, λ) in order to visualize the influence of these parameters.



(a) $L = 1°$, $\lambda = 0.01$



(b) $L = 0.1°$, $\lambda = 3$

(c) $L = 50°$, $\lambda = 3$

(d) $L = 2°$, $\lambda = 60$

Figure 6: Curvature, gradient and field value for different combinations of parameters.

7 Solutions

We have seen in the previous sections that, even if the Diva tools for optimising the parameters are used, the resulting parameters are not always the best choice. Here we provide several procedures that will help to minimize the problems.

- In the `driver`, set the minimal/maximal values acceptable for L and λ. This method will help improving the results, but cannot avoid strong changes of the parameter values from one layer to the other.

- Set the mean distance between the data as the lower limit for the correlation length. This is done easily by using the command:
  `divafit -l`

- Vertically filter the parameters values (if working in 3-D analysis). Again this is done by setting the correct values to `isoptimise` in the file `driver`.

Conclusions

We have performed various analysis with Diva on a common data set, in order to evidence the problems that can crop up when the analysis parameters have unsuitable values. The illustration of these problems should serve as a guide to point at incorrect analysed fields, especially in 4-D analysis.
Though the bad choice of parameters can be mitigated with certain options in Diva, the main conclusion is that a visual checking of the products is more than necessary to confirm their quality.

References

Barth, A., Alvera-Azcárate, A., Troupin, C., Ouberdous, M. & Beckers, J.-M. (2010). A web interface for griding arbitrarily distributed in situ data based on Data-Interpolating Variational Analysis (DIVA). *Adv. Geosci.*, **28**: 29–37. doi:10.5194/adgeo-28-29-2010. URL `www.adv-geosci.net/28/29/2010/`

Brankart, J.-M. & Brasseur., P. (1996). Optimal analysis of in situ data in the Western Mediterranean using statistics and cross-validation. *Journal of Atmospheric and Oceanic Technology*, **13**: 477–491. doi:10.1175/1520-0426(1996)013<0477:OAOISD>2.0.CO;2.

Brankart, J.-M. & Brasseur, P. (1998). The general circulation in the Mediterranean Sea: a climatological approach. *Journal of Marine Systems*, **18**: 41–70. doi:10.1016/S0924-7963(98)00005-0.

Brasseur, P. (1994). *Reconstruction de champs d'observations océanographiques par le Modèle Variationnel Inverse: Méthodologie et Applications*. Ph.D. thesis, University of Liège.

Brasseur, P., Beckers, J.-M., Brankart, J.-M. & Schoenauen, R. (1996). Seasonal temperature and salinity fields in the Mediterranean Sea: Climatological analyses of a historical data set. *Deep-Sea Research I*, **43**: 159–192. doi:10.1016/0967-0637(96)00012-X.

Brasseur, P. 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. doi:10.1029/90JC02387.

Denis-Karafistan, A., Martin, J.-M., Minas, H., Brasseur, P., Nihoul, J. & Denis, C. (1998). Space and seasonal distributions of nitrates in the mediterranean sea derived from a variational inverse model. *Deep-Sea Research I*, **45**: 387–408. doi:10.1016/S0967-0637(97)00089-7.

Karafistan, A., Martin, J.-M., Rixen, M. & Beckers, J.-M. (2002). Space and time distributions of phosphates in the Mediterranean Sea. *Deep-Sea Research I*, **49**: 67–82. doi:10.1016/S0967-0637(01)00042-5.

Rixen, M., Beckers, J.-M. & Allen, J. (2001). Diagnosis of vertical velocities with the QG Omega equation: a relocation method to obtain pseudo-synoptic data sets. *Deep-Sea Research I*, **48**: 1347–1373. doi:10.1016/S0967-0637(00)00085-6.

Rixen, M., Beckers, J.-M., Brankart, J.-M. & Brasseur, P. (2000). A numerically efficient data analysis method with error map generation. *Ocean Modelling*, **2**: 45–60. doi:10.1016/S1463-5003(00)00009-3.

Rixen, M., Beckers, J.-M., Levitus, S., Antonov, J., Boyer, T., Maillard, C., Fichaut, M., Balopoulos, E., Iona, S., Dooley, H., Garcia, M.-J., Manca, B., Giorgetti, A., Manzella, G., Mikhailov, N., Pinardi, N., Zavatarelli, M. & the Medar Consortium (2005a). The Western Mediterranean Deep Water: a proxy for global climate change. *Geophysical Research Letters*, **32**: L12608. doi:10.1029/2005GL022702.

Rixen, M., Beckers, J.-M., Maillard, C. & the MEDAR Group (2005b). A hydrographic and bio-chemical climatology of the Mediterranean and the Black Sea: a technical note on the use of coastal data. *Bollettino di Geofisica Teorica e Applicata*, **46**: 319–327.

Troupin, C., Machín, F., Ouberdous, M., Sirjacobs, D., Barth, A. & Beckers, J.-M. (2010). High-resolution climatology of the north-east atlantic using data-interpolating variational analysis (Diva). *Journal of Geophysical Research*, **115**: C08005. doi:10.1029/2009JC005512.
URL `http://www.agu.org/pubs/crossref/2010/2009JC005512.shtml`