Université de Liège

Université de Liège

Spatial analysis of irregularely distributed observations

ESA Summer school 2010

Jean-Marie Beckers and GHER group Aida Alvera, Alexander Barth, Luc Vandenbulcke, Charles Troupin, Damien Sirjacobs, Mohamed Ouberdous

MARE INTERFACULTARY CENTER FOR MARINE RESEARCH - LIEGE UNIVERSITY

http://modb.oce.ulg.ac.be/GHER

MARE INTERFACULTARY CENTER FOR MARINE RESEARCH - LIEGE UNIVERSITY

Université de Liège
MARE-GHER Sart-Tilman B5
4000 Liège, Belgique

Outline

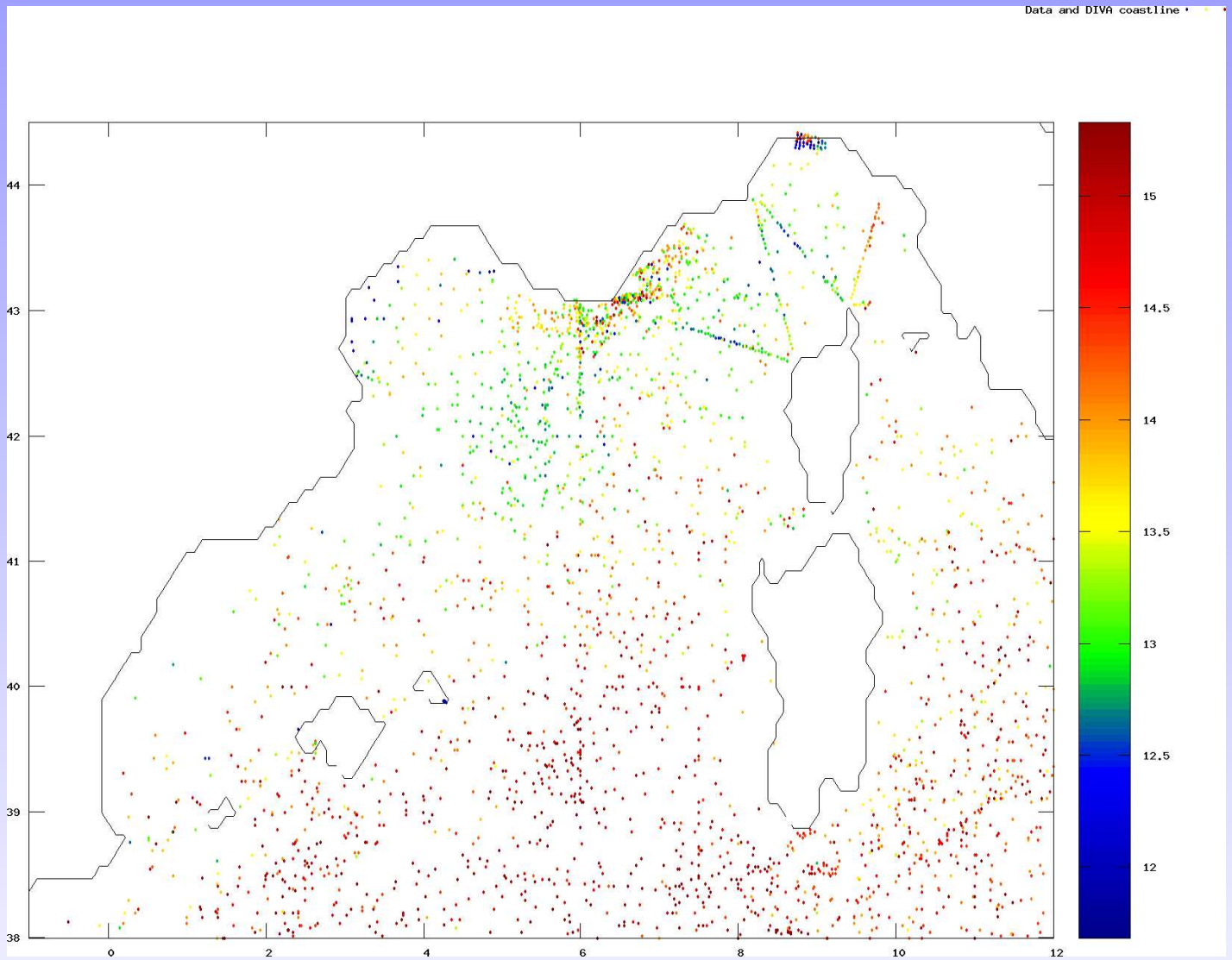
- *Field estimation theory*
- *DIVA*
- *Critical points*
- *Examples*
- *Summary*

Université de Liège

- ***Field estimation theory***
- *DIVA*
- *Critical points*
- *Examples*
- *Summary*

Université de Liège

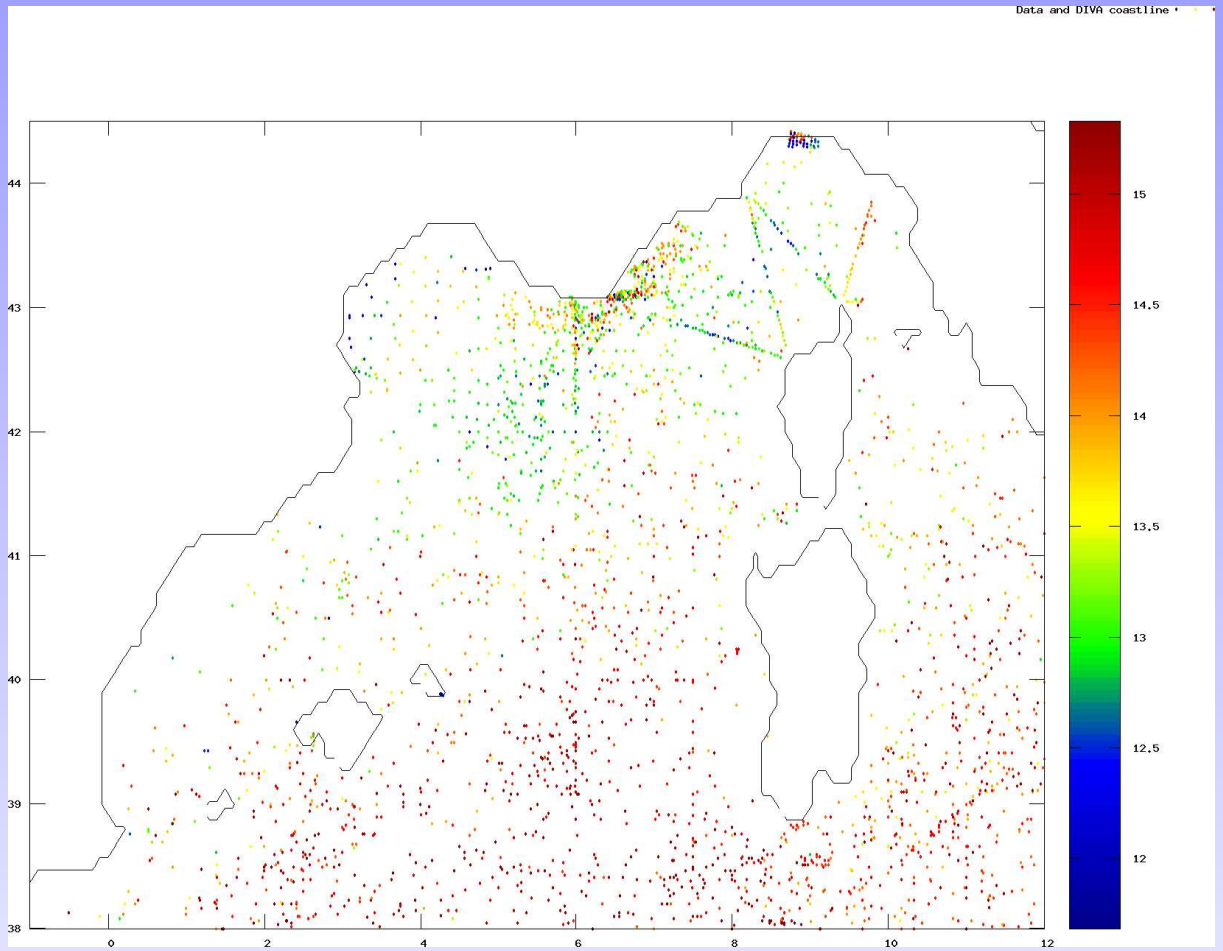
Common problem



Appears when trying to produce maps, calculate volume averages, prepare initial conditions for models, quality control of data ...

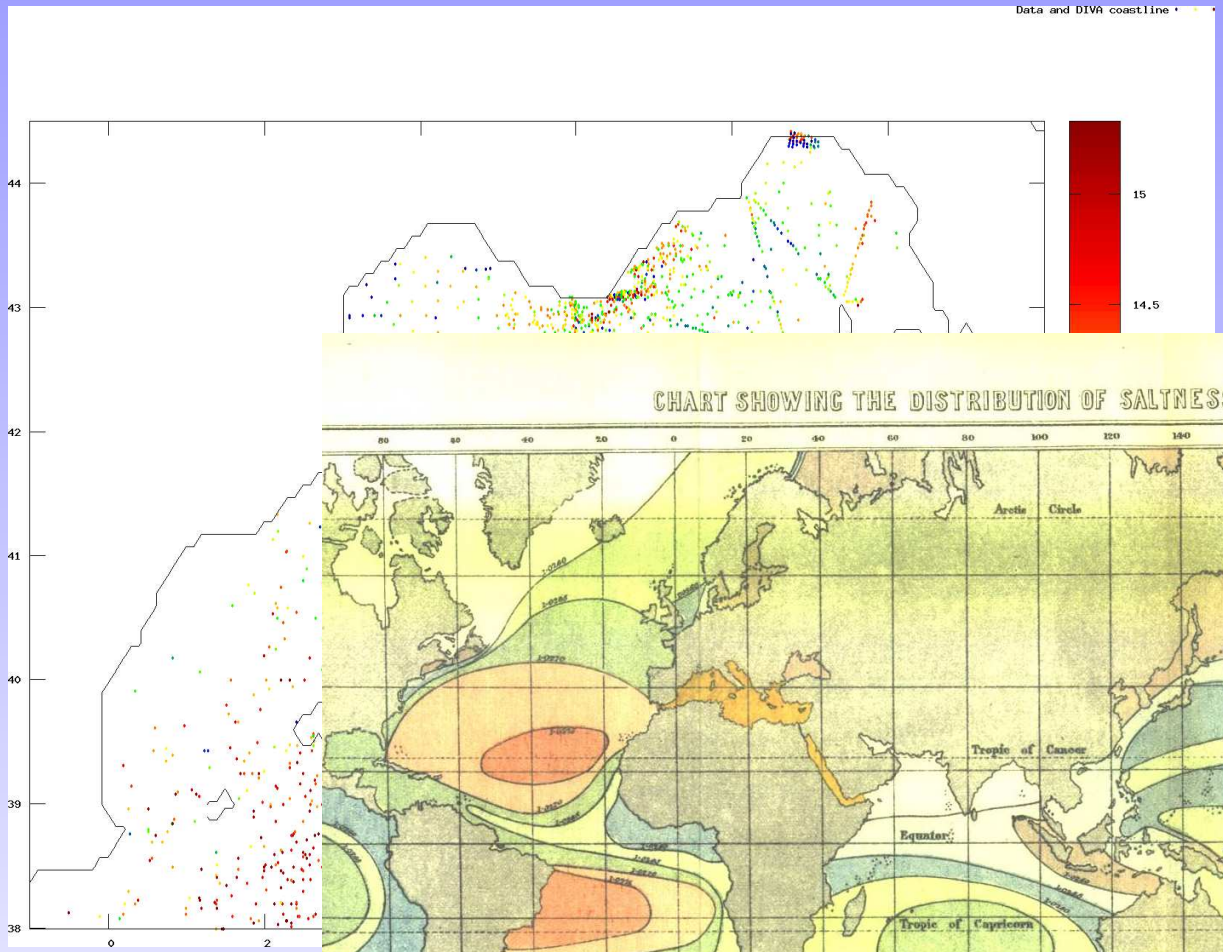
Université de Liège

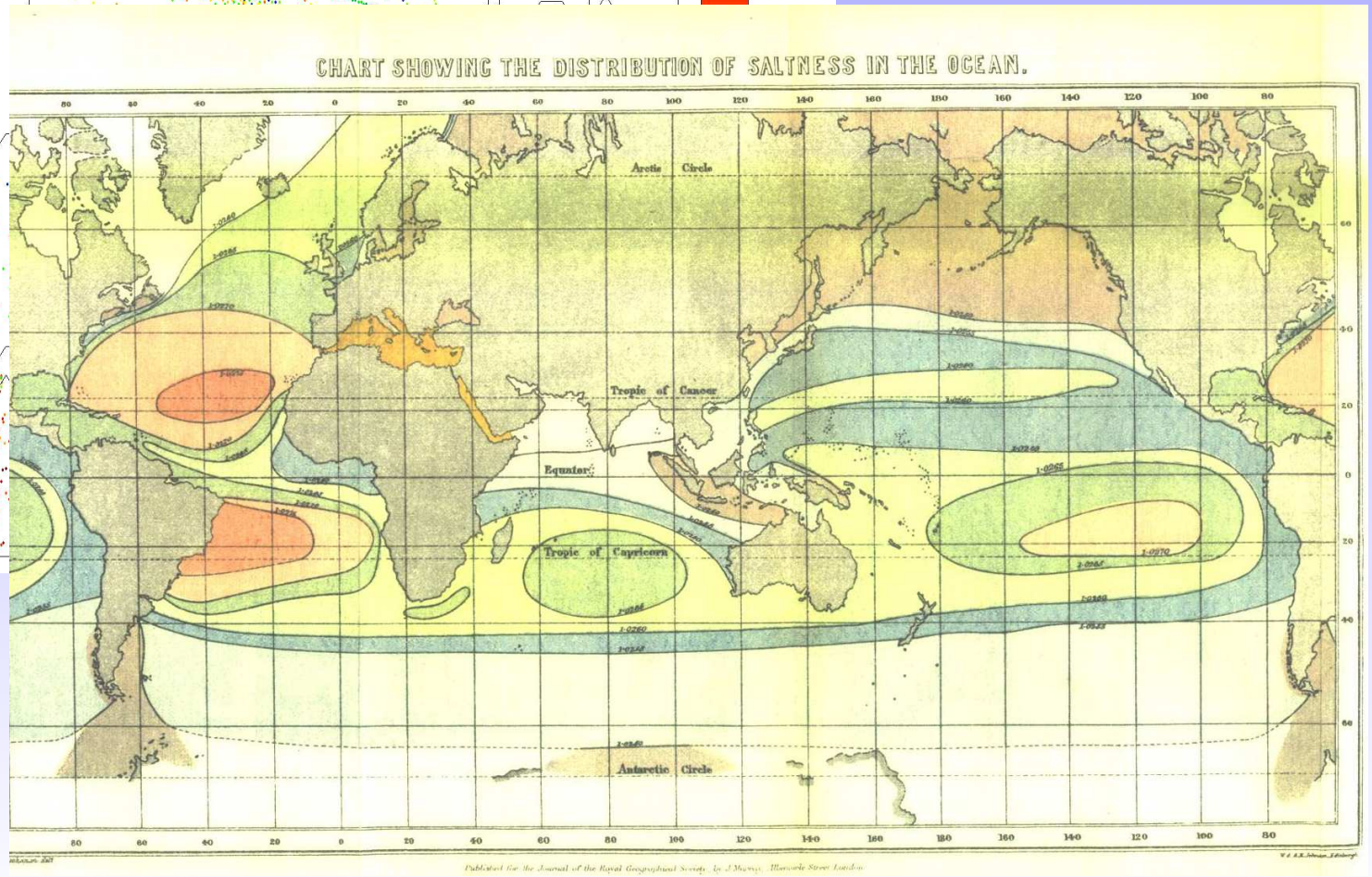
Solutions



Université de Liège

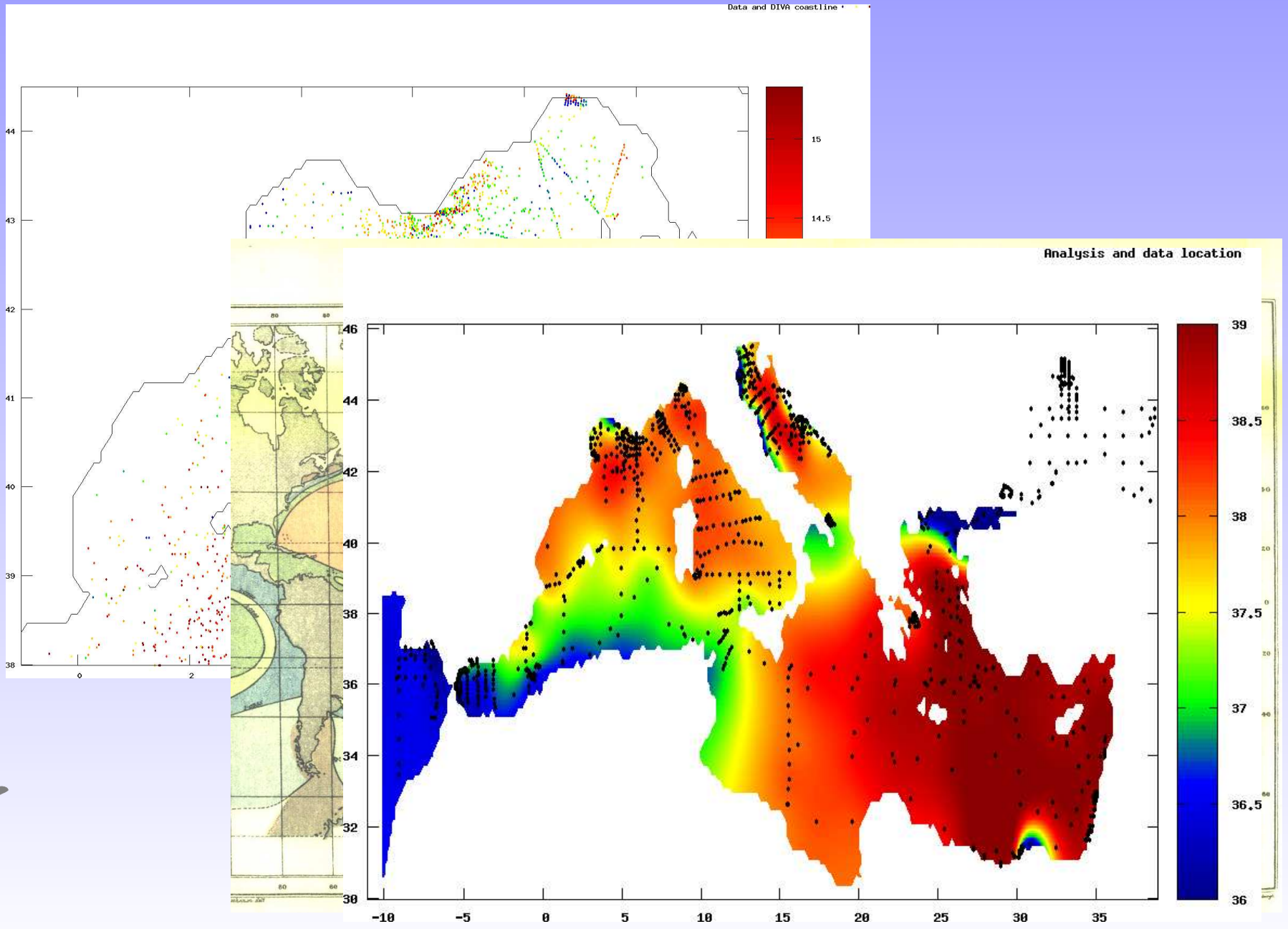
Solutions





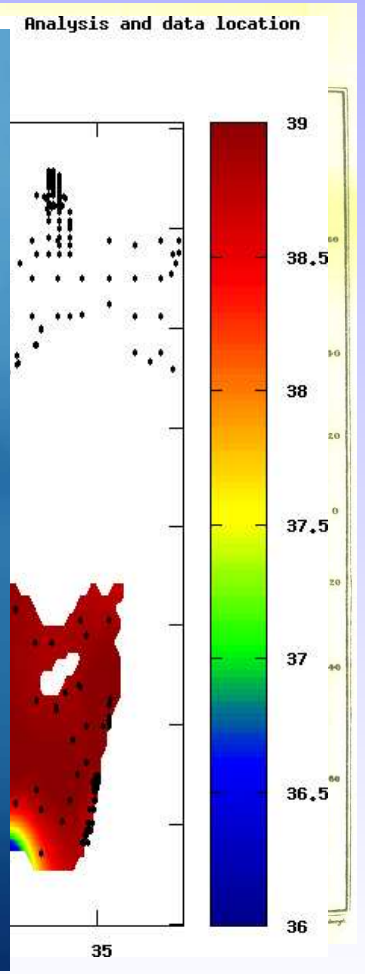
Université de Liège

Solutions



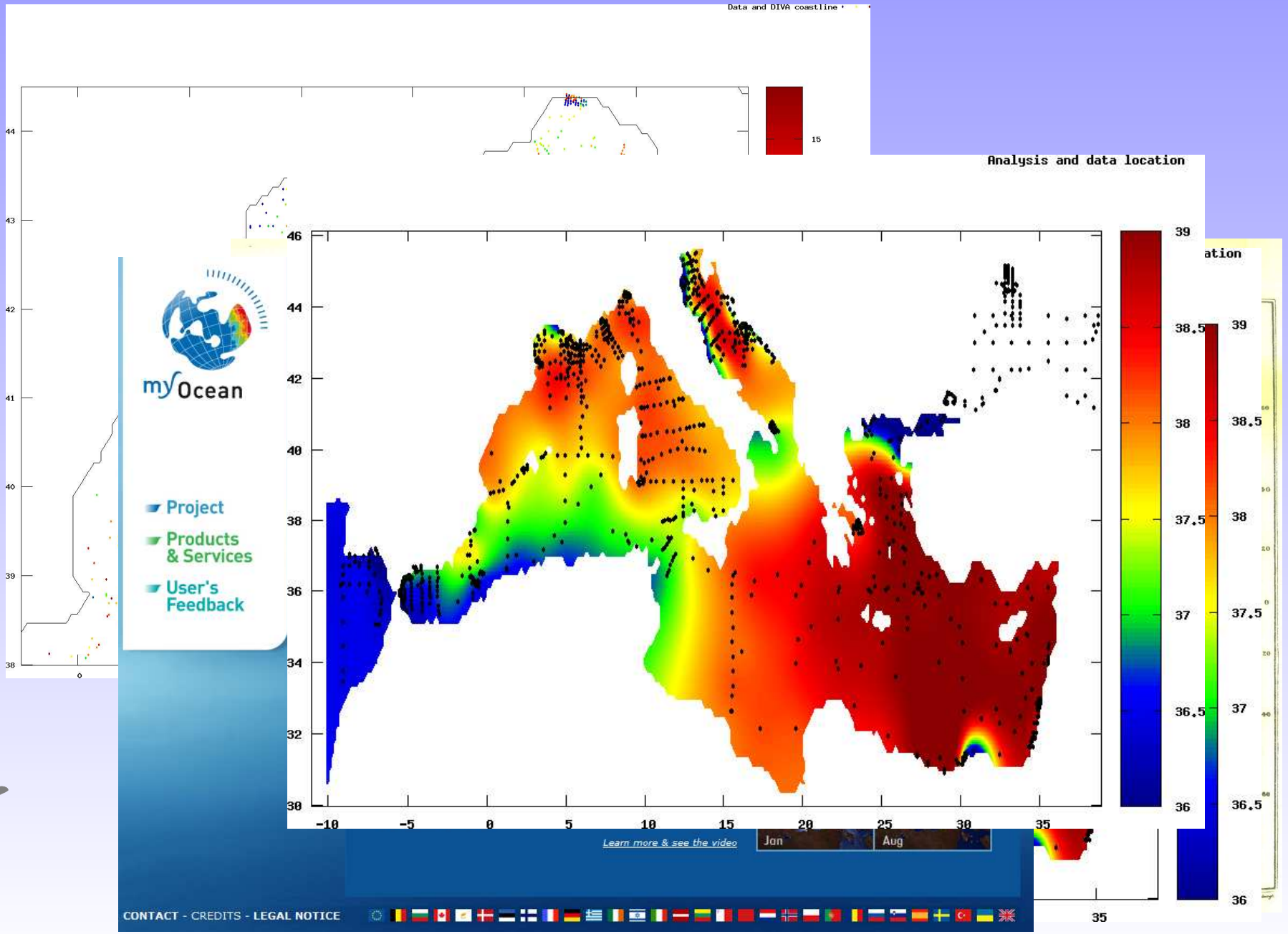
Université de Liège

Solutions



Université de Liège

Solutions



Estimation

- Observer 1: 14°
- Observer 2: 16°

Your best guess ?

Estimation

- Observer 1: 14°
- Observer 2: 16°

Your best guess ?
15°

Estimation

- Observer 1: 14°
- Observer 2: 16°

Your best guess ?

15°

But what if observer 1 uses digital thermometer and observer 2 his finger ?

Estimation

- Observer 1: 14°
- Observer 2: 16°

Your best guess ?

15°

But what if observer 1 uses digital thermometer and observer 2 his finger ?

Best guess probably near 14°.

Estimation

- Observer 1: 14°
- Observer 2: 16°

Your best guess ?

15°

But what if observer 1 uses digital thermometer and observer 2 his finger ?

Best guess probably near 14°.

Exploit knowledge of errors !

Optimal estimate

$$T_1 = T^t + \epsilon_1, \quad \langle \epsilon_1 \rangle = 0, \quad T_2 = T^t + \epsilon_2, \quad \langle \epsilon_2 \rangle = 0 \qquad (1)$$

statistical average, denoted by $\langle \quad \rangle$ with unbiased estimates $\langle \epsilon_* \rangle = 0$
Linear estimate

$$T = w_1 T_1 \; + \; w_2 T_2 \; = \; (w_1 + w_2) T^t \; + \; (w_1 \epsilon_1 + w_2 \epsilon_2) \qquad (2)$$

$$\langle T \rangle = (w_1 + w_2) T^t, \qquad (3)$$

we obtain an unbiased estimate of the true state if we take $w_1 + w_2 = 1$. This leaves one parameter free to chose: w_2

Exploit knowledge on errors to find optimal value of w_2

Choice of weighting ?

$$T^a = (1 - w_2)T_1 + w_2 T_2 \;=\; T_1 + w_2(T_2 - T_1) \tag{4}$$

while in reality there is an error

$$T^a - T^t = (1 - w_2)\epsilon_1 + w_2\epsilon_2, \tag{5}$$

This error is zero on average but its variance is not zero:

$$\left\langle (T^a - T^t)^2 \right\rangle = (1 - w_2)^2 \left\langle \epsilon_1^2 \right\rangle + w_2^2 \left\langle \epsilon_2^2 \right\rangle + 2(1 - w_2) w_2 \left\langle \epsilon_1 \epsilon_2 \right\rangle \tag{6}$$

The actual errors ϵ_1 and ϵ_2 are not known, but the error variance $\left\langle \epsilon_1^2 \right\rangle$ are. Often we can reasonably suppose that the errors ϵ_1 and ϵ_2 are uncorrelated $\left\langle \epsilon_1 \epsilon_2 \right\rangle = 0$. The error variance $\left\langle \epsilon^2 \right\rangle$ of the analysis is

$$\left\langle \epsilon^2 \right\rangle = (1 - w_2)^2 \left\langle \epsilon_1^2 \right\rangle + w_2^2 \left\langle \epsilon_2^2 \right\rangle . \tag{7}$$

So what ?

Minimisation

$$\left\langle \epsilon^2 \right\rangle = (1 - w_2)^2 \left\langle \epsilon_1^2 \right\rangle + w_2^2 \left\langle \epsilon_2^2 \right\rangle . \tag{8}$$

Naturally, the best estimate for T is the one with the lowest expected error variance and we will use w_2, which minimizes the right-hand side:

$$w_2 = \frac{\left\langle \epsilon_1^2 \right\rangle}{\left\langle \epsilon_1^2 \right\rangle + \left\langle \epsilon_2^2 \right\rangle} \tag{9}$$

$$T^a = \frac{\left\langle \epsilon_1^2 \right\rangle \left\langle \epsilon_2^2 \right\rangle}{\left\langle \epsilon_1^2 \right\rangle + \left\langle \epsilon_2^2 \right\rangle} \left(\frac{T_1}{\left\langle \epsilon_1^2 \right\rangle} + \frac{T_2}{\left\langle \epsilon_2^2 \right\rangle} \right) . \tag{10}$$

Best estimate

With (9) we obtain the minimal error variance

$$\left\langle \epsilon^2 \right\rangle = \frac{\left\langle \epsilon_1^2 \right\rangle \left\langle \epsilon_2^2 \right\rangle}{\left\langle \epsilon_1^2 \right\rangle + \left\langle \epsilon_2^2 \right\rangle} = \left(1 - \frac{\left\langle \epsilon_1^2 \right\rangle}{\left\langle \epsilon_1^2 \right\rangle + \left\langle \epsilon_2^2 \right\rangle} \right) \left\langle \epsilon_1^2 \right\rangle , \qquad (11)$$

while the estimate of the temperature itself reads

$$T^a = T_1 \; + \; \left(\frac{\left\langle \epsilon_1^2 \right\rangle}{\left\langle \epsilon_1^2 \right\rangle + \left\langle \epsilon_2^2 \right\rangle} \right) (T_2 - T_1) \, . \qquad (12)$$

Error variance on the combination of T_1 and T_2 is smaller than both $\left\langle \epsilon_1^2 \right\rangle$ and $\left\langle \epsilon_2^2 \right\rangle$.

VAR approach

Same solution by

$$\min_T J = \frac{(T - T_1)^2}{2\left\langle \epsilon_1^2 \right\rangle} + \frac{(T - T_2)^2}{2\left\langle \epsilon_2^2 \right\rangle}. \tag{13}$$

Optimal interpolation

Analysis $\mathbf{x}^a$ as a linear combination of the forecast $\mathbf{x}^f$ and the observations $\mathbf{y}$:

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^f\right) \tag{14}$$

H observation operator and innovation vector

$$\boldsymbol{d} = \mathbf{y} - \mathbf{H}\mathbf{x}^f \tag{15}$$

Objective: prescribe an optimal matrix **K**: Kalman gain matrix

Knowledge on error distribution

$$\boldsymbol{\epsilon} = \mathbf{x} - \mathbf{x}^t \tag{16}$$

$$\boldsymbol{\epsilon}^o = \mathbf{y} - \mathbf{y}^t. \tag{17}$$

Error-covariance matrix

$$\mathbf{R} = \left\langle \boldsymbol{\epsilon}^o \boldsymbol{\epsilon}^{o\mathsf{T}} \right\rangle \tag{18}$$

is semi-positive defined since $z^{\mathsf{T}}\mathbf{R}z = \left\langle \left(z^{\mathsf{T}}\boldsymbol{\epsilon}^o\right)^2 \right\rangle$

The analysis step (14) reads

$$\mathbf{x}^t + \boldsymbol{\epsilon}^a = \mathbf{x}^t + \boldsymbol{\epsilon}^f + \mathbf{K}\left(\boldsymbol{\epsilon}^o - \mathbf{H}\boldsymbol{\epsilon}^f\right) + \mathbf{K}\underbrace{\left(\mathbf{y}^t - \mathbf{H}\mathbf{x}^t\right)}_{=0} \tag{19}$$

Kalman gain

$$\boldsymbol{\epsilon}^a = \boldsymbol{\epsilon}^f + \mathbf{K}\left(\boldsymbol{\epsilon}^o - \mathbf{H}\boldsymbol{\epsilon}^f\right). \tag{20}$$

Construct the error covariance $\left\langle \boldsymbol{\epsilon}^a \boldsymbol{\epsilon}^{a\mathsf{T}} \right\rangle$ of the analysis by multiplying (20) by its transposed and take the statistical average or expectation

$$\begin{aligned}\left\langle \boldsymbol{\epsilon}^a \boldsymbol{\epsilon}^{a\mathsf{T}} \right\rangle &= \left\langle \boldsymbol{\epsilon}^f \boldsymbol{\epsilon}^{f\mathsf{T}} \right\rangle + \mathbf{K}\left\langle (\boldsymbol{\epsilon}^o - \mathbf{H}\boldsymbol{\epsilon}^f)\boldsymbol{\epsilon}^{f\mathsf{T}} \right\rangle + \left\langle \boldsymbol{\epsilon}^f (\boldsymbol{\epsilon}^{o\mathsf{T}} - \boldsymbol{\epsilon}^{f\mathsf{T}}\mathbf{H}^{\mathsf{T}}) \right\rangle \mathbf{K}^{\mathsf{T}} \\ &+ \mathbf{K}\left\langle (\boldsymbol{\epsilon}^o - \mathbf{H}\boldsymbol{\epsilon}^f)(\boldsymbol{\epsilon}^{o\mathsf{T}} - \boldsymbol{\epsilon}^{f\mathsf{T}}\mathbf{H}^{\mathsf{T}}) \right\rangle \mathbf{K}^{\mathsf{T}}. \end{aligned} \tag{21}$$

Define covariance matrices

$$\mathbf{P} = \left\langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}} \right\rangle \tag{22}$$

and assume that observational errors and model errors are not correlated, $\left\langle \boldsymbol{\epsilon}^o \boldsymbol{\epsilon}^{\mathsf{T}} \right\rangle = 0$.

Kalman gain

The error-covariance matrix after analysis can then be written as

$$\begin{aligned}\mathbf{P}^a &= \mathbf{P}^f - \mathbf{K}\mathbf{H}\mathbf{P}^f - \mathbf{P}^f\mathbf{H}^\mathsf{T}\mathbf{K}^\mathsf{T} + \mathbf{K}\left(\mathbf{R} + \mathbf{H}\mathbf{P}^f\mathbf{H}^\mathsf{T}\right)\mathbf{K}^\mathsf{T} \\ &= \mathbf{P}^f - \mathbf{P}^f\mathbf{H}^\mathsf{T}\mathbf{A}^{-1}\mathbf{H}\mathbf{P}^f + \left(\mathbf{P}^f\mathbf{H}^\mathsf{T} - \mathbf{K}\mathbf{A}\right)\mathbf{A}^{-1}\left(\mathbf{H}\mathbf{P}^f - \mathbf{A}\mathbf{K}^\mathsf{T}\right) \quad (23)\end{aligned}$$

where we define matrix

$$\mathbf{A} = \mathbf{H}\mathbf{P}^f\mathbf{H}^\mathsf{T} + \mathbf{R} \quad (24)$$

which is symmetric and we suppose that it can be inverted. Global error estimate:

$$\epsilon^a = \left\langle \boldsymbol{\epsilon}^{a\,\mathsf{T}}\boldsymbol{\epsilon}^a \right\rangle = \mathrm{trace}\left(\mathbf{P}^a\right). \quad (25)$$

Search for an optimal **K** which minimizes this trace or for which

$$\epsilon^a(\mathbf{K} + \mathbf{L}) - \epsilon^a(\mathbf{K}) = 0 \quad (26)$$

for any small departure matrix **L**

Kalman gain

$$\mathrm{trace}\left(-\mathbf{L}\left(\mathbf{H}\mathbf{P}^{f}-\mathbf{A}\mathbf{K}^{\mathsf{T}}\right)-\left(\mathbf{P}^{f}\mathbf{H}^{\mathsf{T}}-\mathbf{K}\mathbf{A}\right)\mathbf{L}^{\mathsf{T}}\right)=0, \qquad (27)$$

where we neglected quadratic terms in **L**. The two terms are the transposed version of each other and since the trace of the matrix and its transposed are identical, we must request

$$\mathrm{trace}\left(\left(\mathbf{P}^{f}\mathbf{H}^{\mathsf{T}}-\mathbf{K}\mathbf{A}\right)\mathbf{L}^{\mathsf{T}}\right)=0.$$

Since **L** is arbitrary, the optimal solution with minimum error is obtained when

$$\mathbf{K}=\mathbf{P}^{f}\mathbf{H}^{\mathsf{T}}\mathbf{A}^{-1}=\mathbf{P}^{f}\mathbf{H}^{\mathsf{T}}\left(\mathbf{H}\mathbf{P}^{f}\mathbf{H}^{\mathsf{T}}+\mathbf{R}\right)^{-1} \qquad (28)$$

Kalman gain

The error covariance of the analysis is obtained by injecting (28) into (23)

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{P}^f = \left(\mathbf{I} - \mathbf{P}^f\mathbf{H}^\mathsf{T}\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^\mathsf{T} + \mathbf{R}\right)^{-1}\mathbf{H}\right)\mathbf{P}^f \qquad (29)$$

which is the analogue of (11). Note that both the Kalman gain matrix and the error covariance after the analysis do not depend on the *value* of the observations or the forecasted state vector but only on their statistical error covariances. The only field that depends on the actual values is of course the state vector itself:

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{P}^f\mathbf{H}^\mathsf{T}\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^\mathsf{T} + \mathbf{R}\right)^{-1}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^f\right). \qquad (30)$$

The use of (28) in (14) to combine the forecast and observation with prescribed error covariance $\mathbf{P}^f$ and $\mathbf{R}$ is known as optimal interpolation (OI)

3D-Var

Find the state vector that minimizes the error measure J given by

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^f)^\mathsf{T} \mathbf{P}^{f\,-1}(\mathbf{x} - \mathbf{x}^f) \; + \; \frac{1}{2}(\mathbf{H}\mathbf{x} - \mathbf{y})^\mathsf{T} \mathbf{R}^{-1}(\mathbf{H}\mathbf{x} - \mathbf{y}) \quad (31)$$

Yields the same optimal state.

3D-Var

Find the state vector that minimizes the error measure J given by

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^f)^{\mathsf{T}} \mathbf{P}^{f\,-1}(\mathbf{x} - \mathbf{x}^f) \; + \; \frac{1}{2}(\mathbf{H}\mathbf{x} - \mathbf{y})^{\mathsf{T}} \mathbf{R}^{-1}(\mathbf{H}\mathbf{x} - \mathbf{y}) \quad (32)$$

Yields the same optimal state.

VAR minimises a sum of a quadratic term of the statevariables and a quadratic term of misfits (residuals)

3D-Var

Find the state vector that minimizes the error measure J given by

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^f)^\mathsf{T}\mathbf{P}^{f\,-1}(\mathbf{x} - \mathbf{x}^f) \; + \; \frac{1}{2}(\mathbf{H}\mathbf{x} - \mathbf{y})^\mathsf{T}\mathbf{R}^{-1}(\mathbf{H}\mathbf{x} - \mathbf{y}) \quad (33)$$

Yields the same optimal state.

VAR minimises a sum of a quadratic term of the statevariables and a quadratic term of misfits (residuals)

Kalman filter and 3DVAR yield same results when used with same covariances in linear case.

Spatial interpolation

- "Model forecast": Background field.
- $\mathbf{HP}^f\mathbf{H}^\mathsf{T}$ is then the covariance of the background field between data points: each element i, j of $\mathbf{B}$ provides the covariance between points in location i and j. Covariance between a given point and all data points is stored in column vector $\mathbf{c}$ and the local variance at the analysis point is noted σ^2.
- Analysis ϕ of anomaly $\mathbf{y}$ with respect to background leads to spatial analysis at any desired location of covariance between any two points is known.

$$\phi = \mathbf{c}^\mathsf{T}(\mathbf{B}+\mathbf{R})^{-1}\mathbf{y} \tag{34}$$

with a local error variance of the analysis

$$\epsilon_a^2 = \sigma^2 - \mathbf{c}^\mathsf{T}(\mathbf{B}+\mathbf{R})^{-1}\mathbf{c} \tag{35}$$

Note that inversion of matrix is needed (cost increases as the cube of number of data points).

Background covariance

Problem, how to specify background covariances (between all data points and between data points and the desired analysis location).

- c_i= covariance between location of the analysis and data location of point i = $C(x, x_i)$
- B_{ij}=covariance between location of data point i and location of point j= $C(x_i, x_j)$

Approaches

- Normally obtained via statistics on data. Seldom possible (noticable exception: satellite images).
- Standard OI: via functions $B_{ij} = f(r/L)$ where r is the distance between points i and j, but still function f needs to be determined. L is the so-called correlation length. Here statistics on all data couples as a function of distance. Example: $f = \sigma^2 \exp(-r^2/L^2)$.
- Via functionals (see Kernel of DIVA later)

Signal to noise ratio

$$\mathbf{B} = \sigma^2 \tilde{\mathbf{B}} \tag{36}$$

$$\mathbf{R} = \epsilon^2 \tilde{\mathbf{R}} \tag{37}$$

$$\mathbf{c} = \sigma^2 \tilde{\mathbf{c}} \tag{38}$$

with non-dimensional correlation matrixes $\tilde{\mathbf{B}}$, $\tilde{\mathbf{R}}$, $\tilde{\mathbf{c}}$

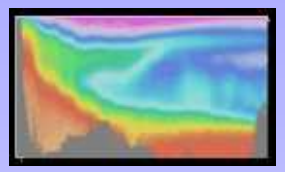
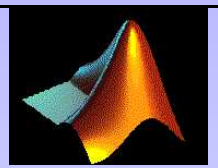
$$\phi = \tilde{\mathbf{c}}^{\mathsf{T}} \left(\tilde{\mathbf{B}} + \frac{1}{\lambda} \tilde{\mathbf{R}} \right)^{-1} \mathbf{y} \tag{39}$$

with signal-to noise ratio

$$\lambda = \frac{\sigma^2}{\epsilon^2} \tag{40}$$

Also the error field is only depending on the ratio.

Statistical spatial analysis

Tools	Formats	Method
	ODV spreadsheet WOCE WOA ...	Dist. weighting, VIM
	netCDF (toolbox), CSV ascii, ...	Polynomial interpolation ...
	CSV ascii, ...	Kriging, OI, ...
	ODV spreadsheet	Variational

Mostly graphics oriented, without "oceanographic" knowledge.

Université de Liège

- *Field estimation theory*
- ***DIVA***
- *Critical points*
- *Examples*
- *Summary*

DIVA basics

Variational Inverse Method, (Brasseur $et\ al.$, 1996). Knowing data d_j at location (x_j, y_j), search the field φ which minimizes

$$J[\varphi] = \sum_{j=1}^{Nd} \mu_j [d_j - \varphi(x_j, y_j)]^2 + \|\varphi - \varphi_b\|^2 \qquad (41)$$

$$\|\varphi\| = \int_D (\alpha_2 \nabla\nabla\varphi : \nabla\nabla\varphi + \alpha_1 \nabla\varphi \cdot \nabla\varphi + \alpha_0 \varphi^2) \mathrm{d}D \qquad (42)$$

The background field φ_b is typically the data average value.

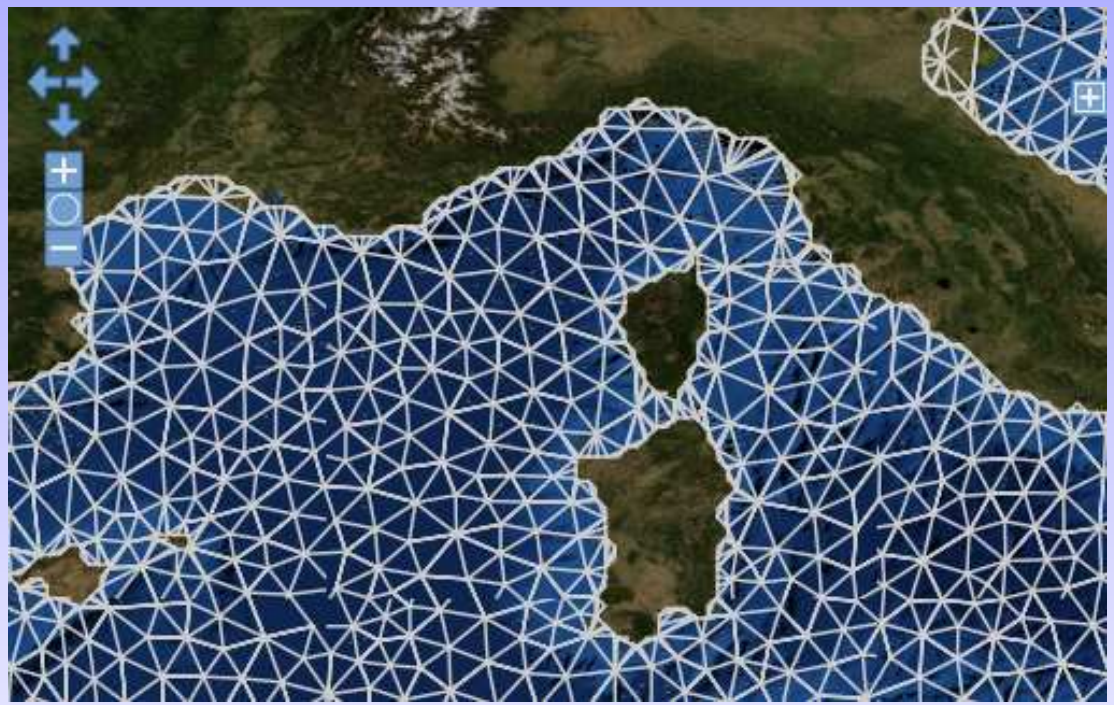
- α_0 penalizes the field itself (anomalies),
- α_1 penalizes gradients $\nabla\varphi$ (no trends) ,
- α_2 penalizes variability (regularization of second derivatives $\nabla\nabla\varphi$).
- α_* can be related to a length scale L of the analysis.
- μ_j penalizes data-analysis misfits (objective).

Université de Liège

DIVA basics

$$\mu = \frac{\sigma^2}{\epsilon^2}\frac{4\pi}{L^2} \tag{43}$$

where the σ^2/ϵ^2 is known as a signal to noise ratio S/N.



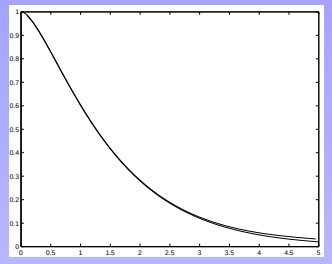
Solution by finite element method including DIVA mesh generator using topographic data. Note decoupling of subbasins.

Université de Liège

Université de Liège

DIVA as OI

DIVA is identical to the well known Optimal Interpolation



- if so-called reproducing kernel of the norm=covariance function of OI,
- if the noise is random, spatially uncorrelated and the signal/noise ratio parameter is identical.

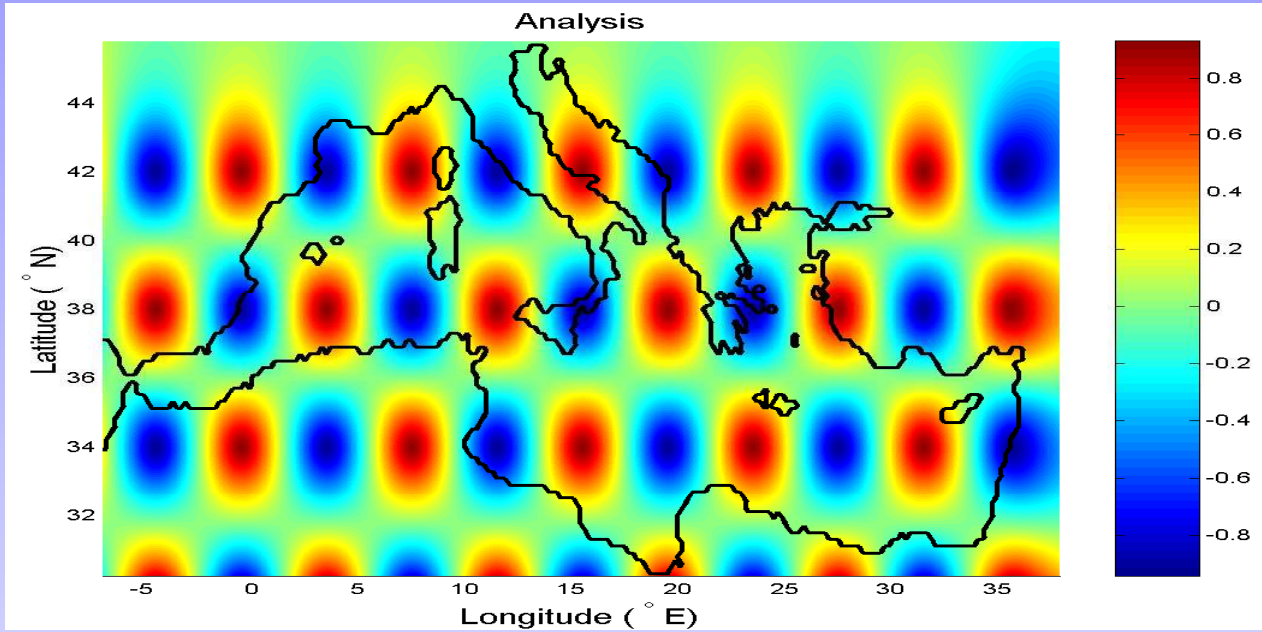
In this case, the OI solution=DIVA solution.

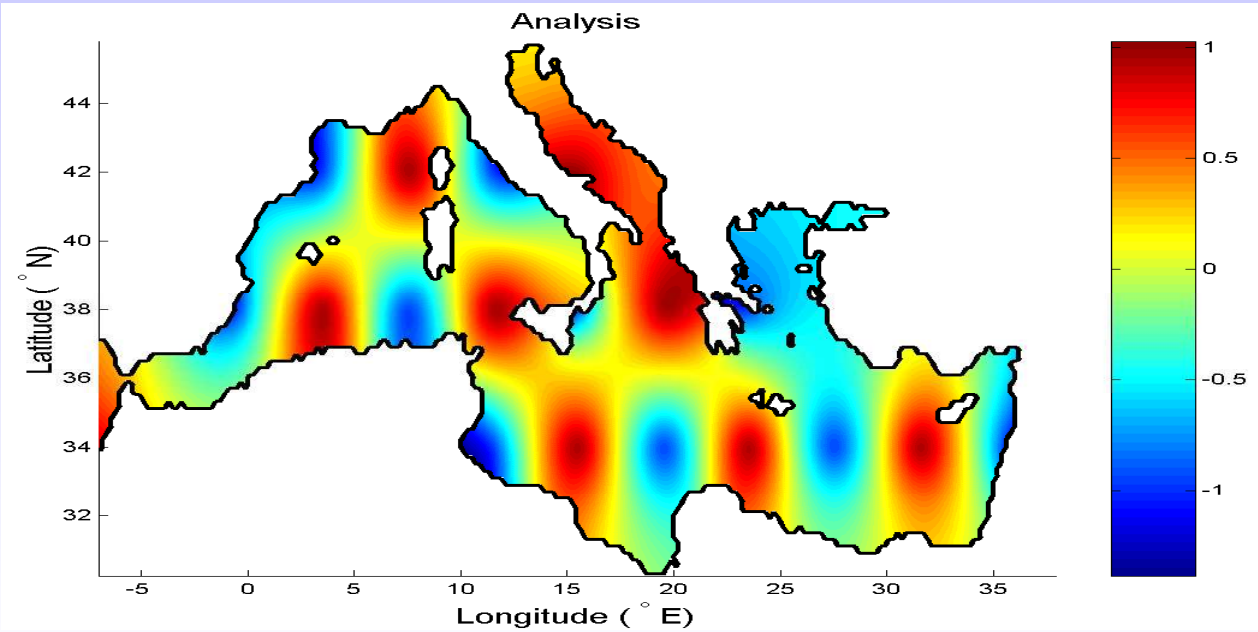
- Advantages of DIVA: regularization, fast finite element solution, boundary effects taken into account.
- Difficulties: generalizations to 3D, error estimates and multivariate versions are "hybrid".

Major direct advantage of DIVA: matrix to invert is related to the finite element mesh, NOT the number of data. Useful for large data sets (Rixen *et al.*, 2000). Equivalence allows to calculate error fields with DIVA even if formulation does not rely on error minimisation.

Université de Liège

Illustration of covariance functions





Additions to basic tool

Advection constraint: Augmented cost function to deal with preferred correlation directions, eg, via advection with velocity $\boldsymbol{u}$ and diffusion $\mathcal{A}$

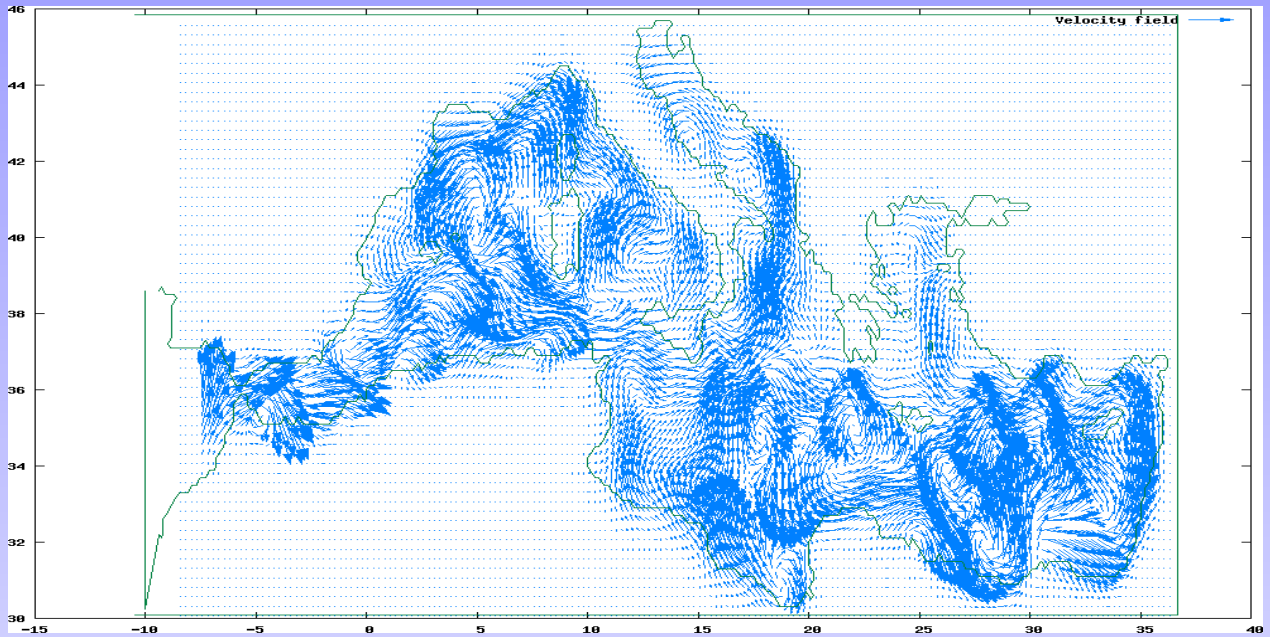
$$J_a = J(\varphi) + \theta \int_D \left[\boldsymbol{u} \cdot \nabla\varphi - \mathcal{A}\nabla\cdot\nabla\varphi\right]^2 \mathrm{d}D \qquad (44)$$

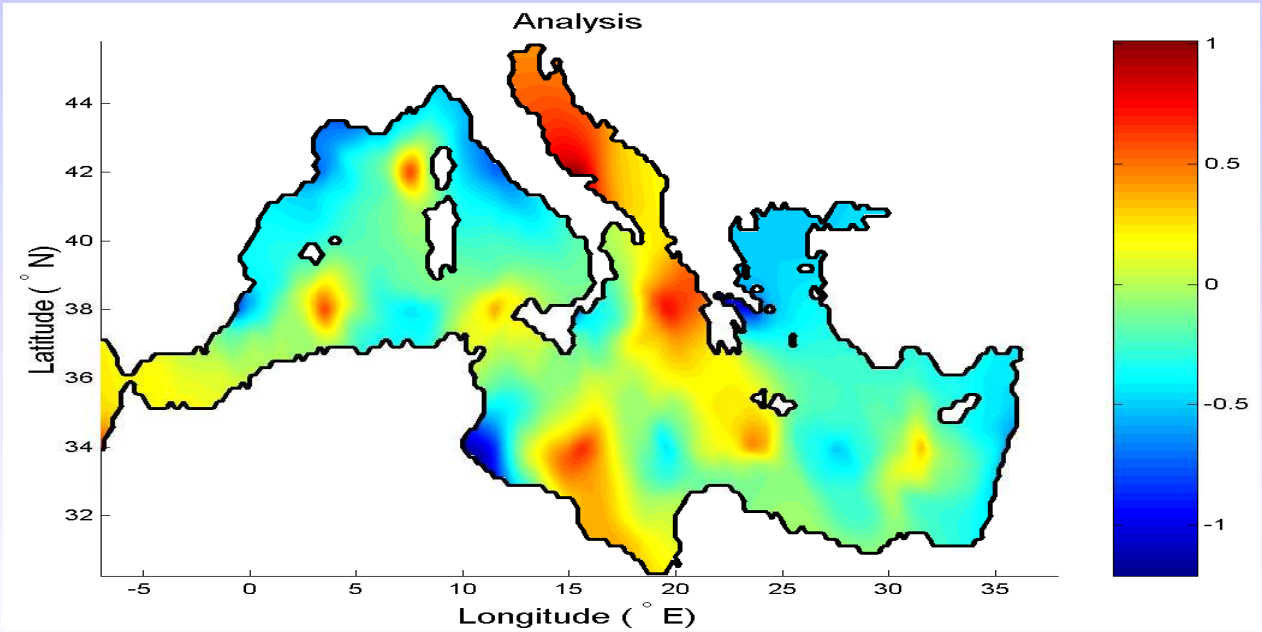
Other features

- Error fields taking data distribution into account.
- Toolbox approach allowing to design own versions.
- 3D and 4D modes by looping, hydrostatic constraint in 3D mode.
- Cross validation tools to infer statistical parameters and error estimates.
- Climatology production version with heterogeneous data distributions (detrending) .
- Outlier detection.
- ...

Université de Liège

Covariances with advection

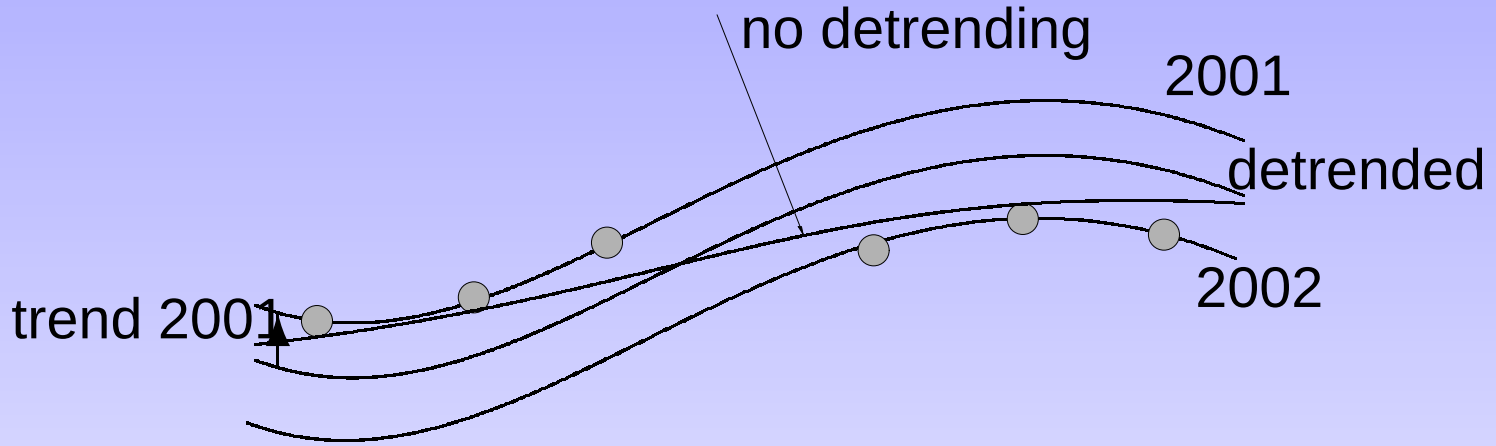




Université de Liège

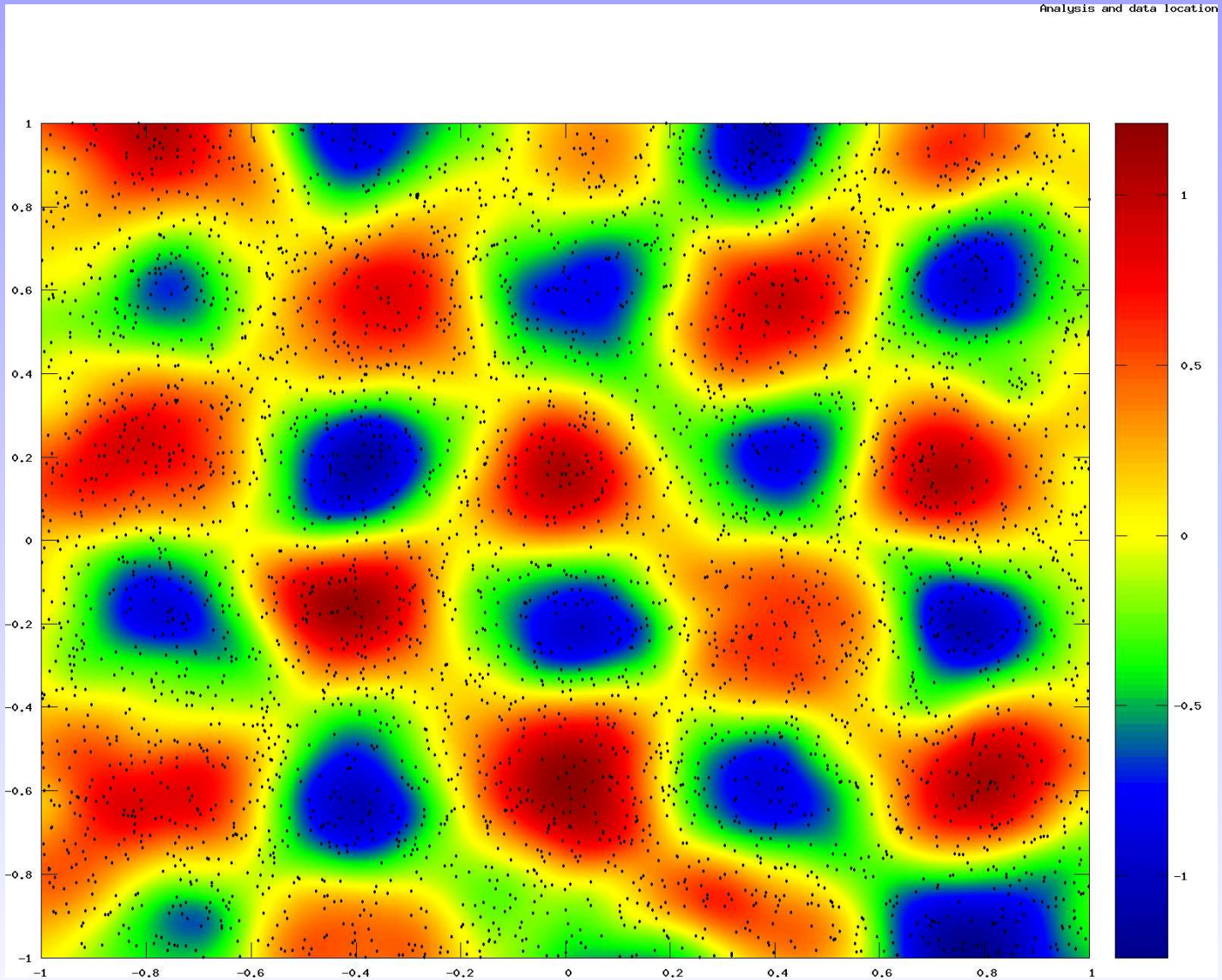
Detrending

Heterogeneous data distribution:



First analysis show a biais for each year's data with respect to first analysis. Subtract the biais estimate and redo the analysis, accumulating the biais. After convergence, detrended analysis+biais of the year.

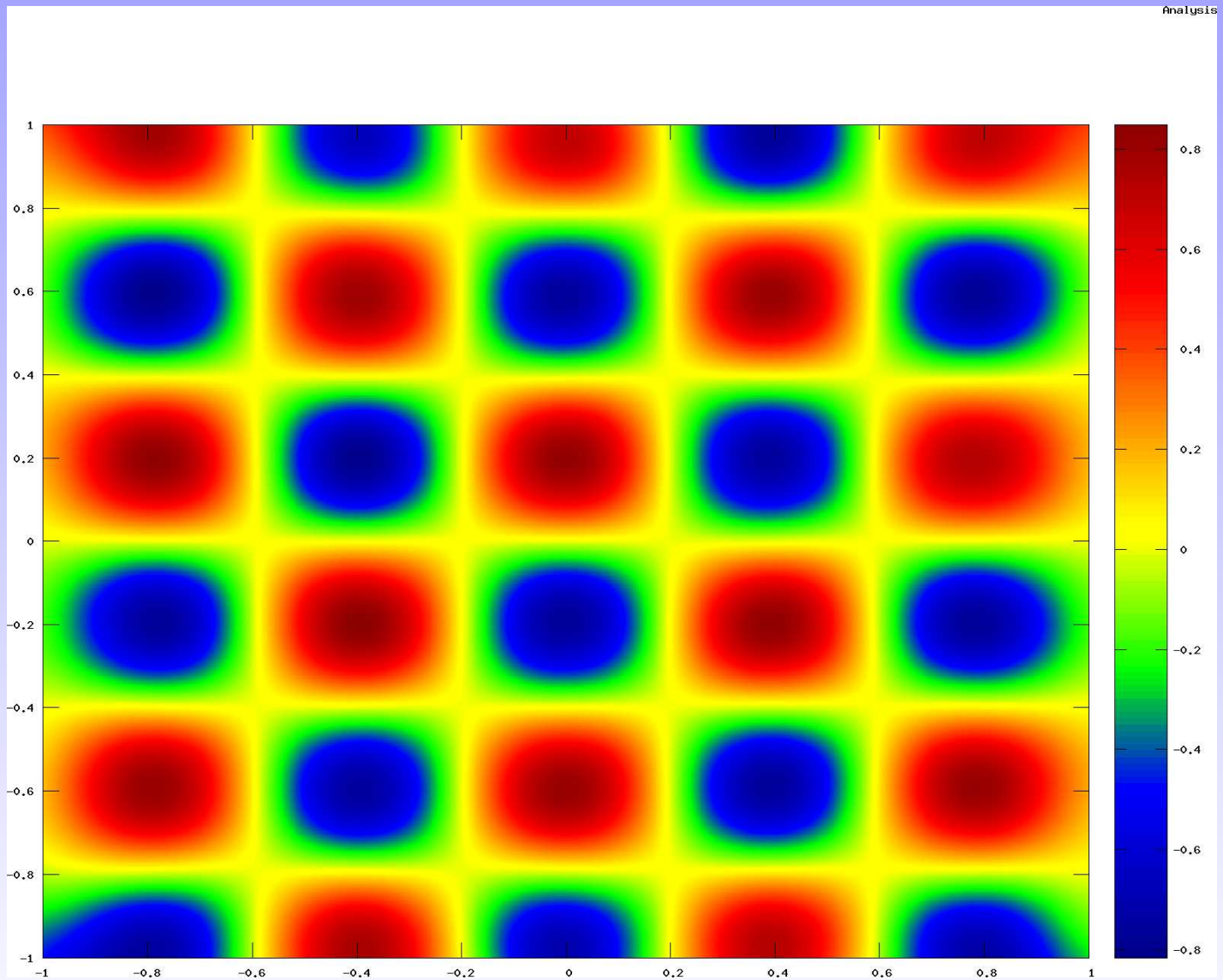
Example without detrending



DIVA analysis of sin-cosine spatial structure with superimposed decadal, seasonal and daily cycles and noise.

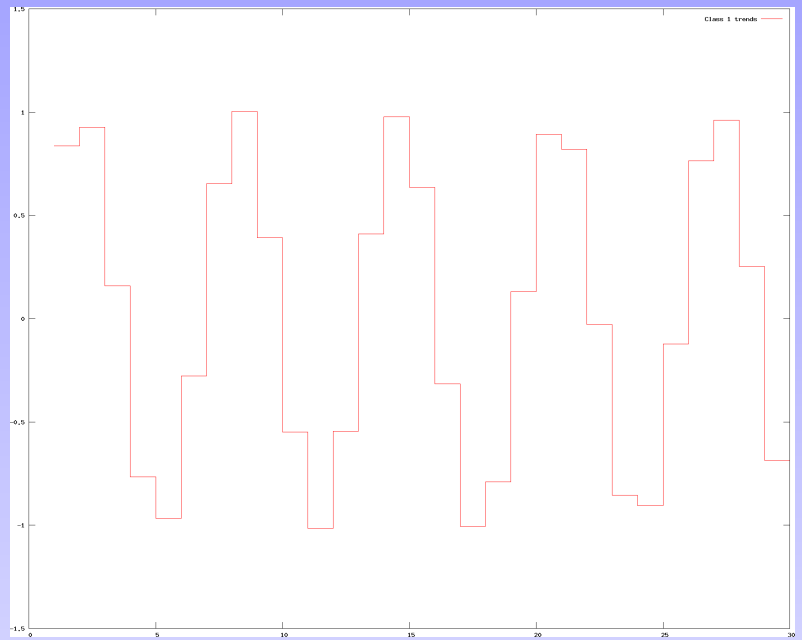
Université de Liège

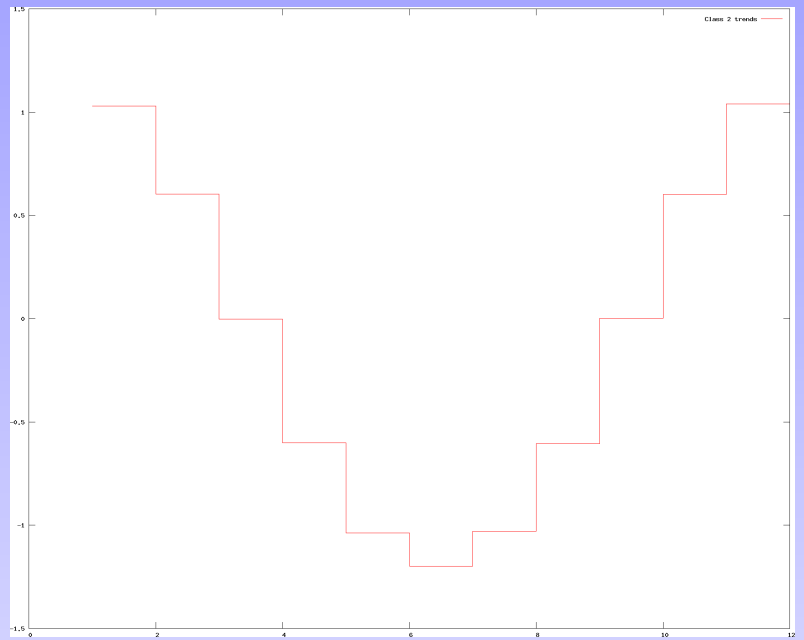
Example with detrending

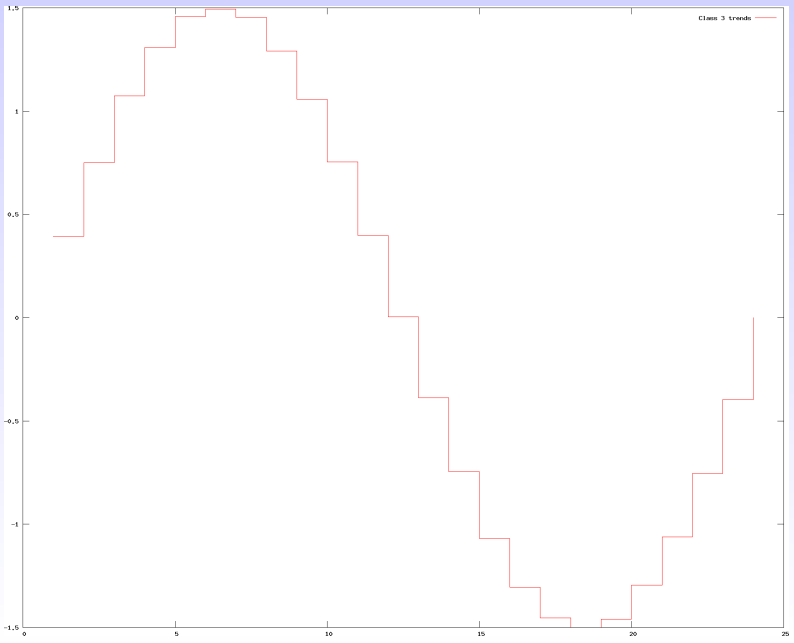


DIVA analysis of sin-cosine spatial structure with superimposed decadal, seasonal and daily cycles and noise.

Trends can also be retrieved

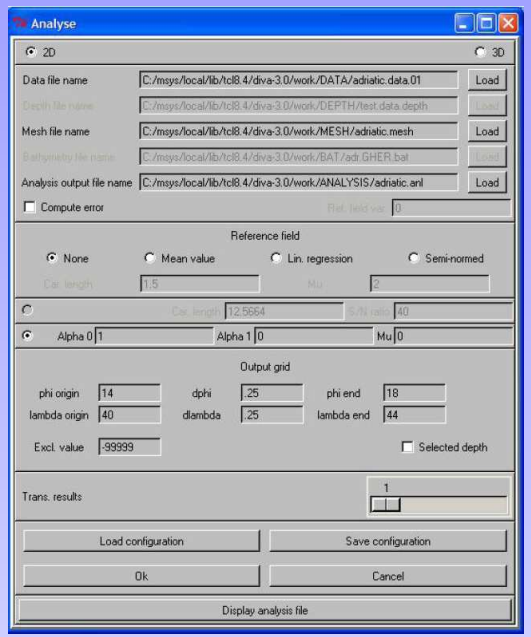


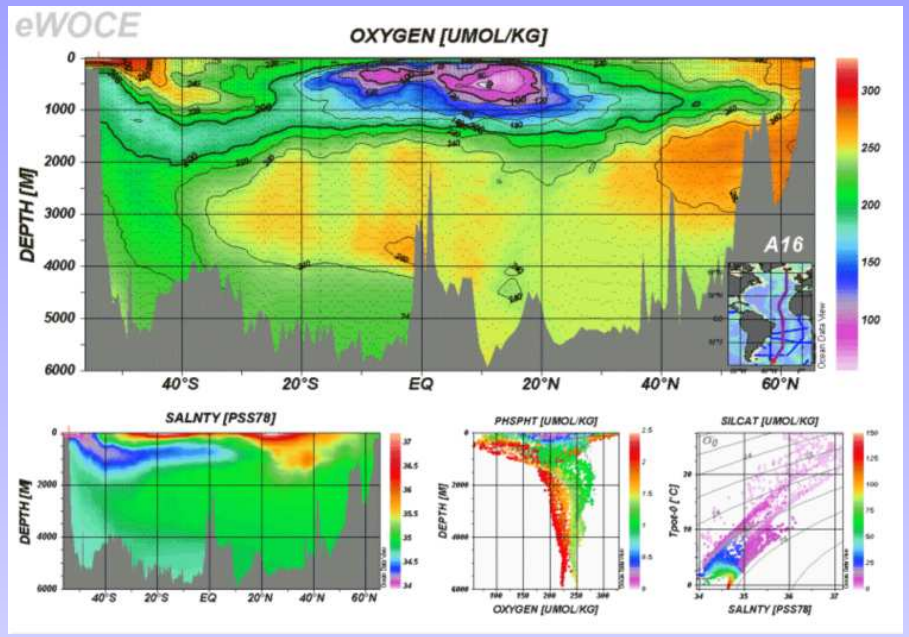


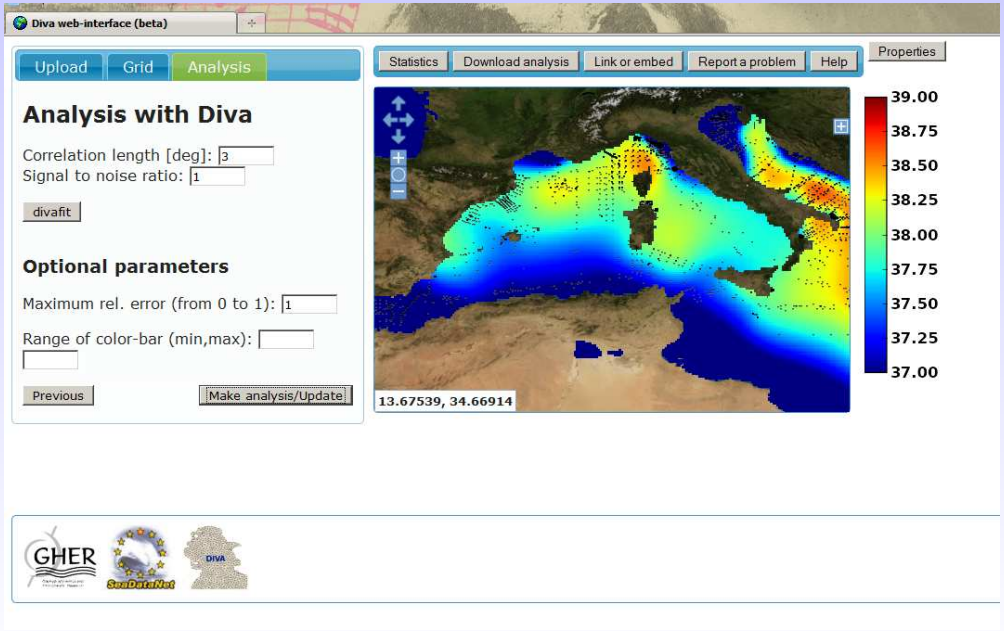


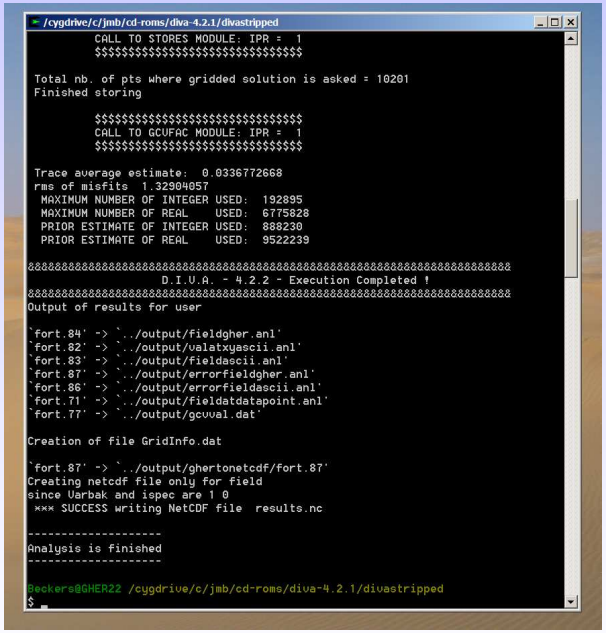
Université de Liège

How to use DIVA?

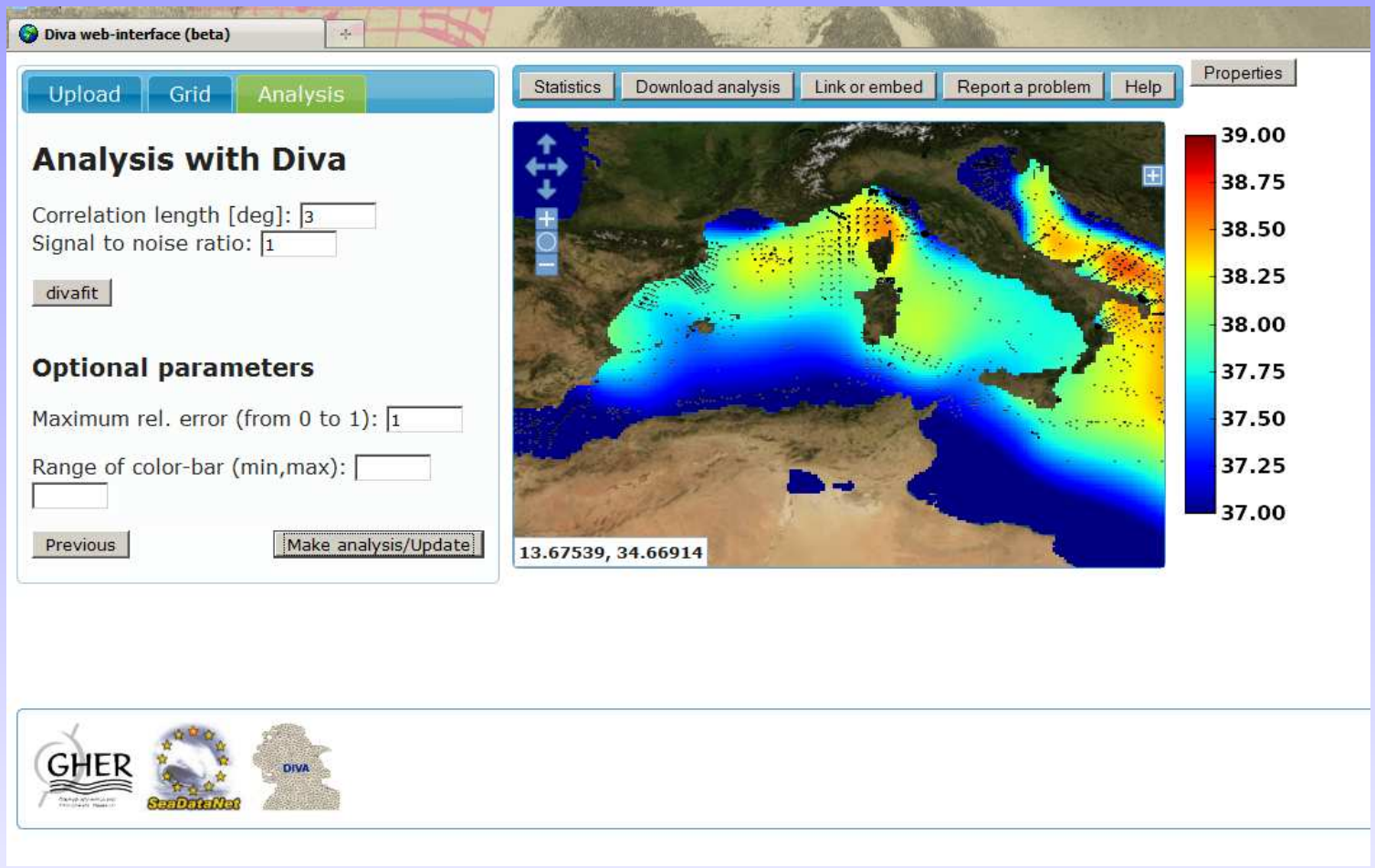








DIVA on WEB



`http://gher-diva.phys.ulg.ac.be`

For occasional use or quick data exploration

Université de Liège

Université de Liège

DIVA on WEB

- A first prototype of DIVA on WEB was developped (not aimed as a tool for climatology production but as an easy way to access the tools for occasional uses).
- User uploads its 2D data in simple ascii format (test version now with ONE ODV4 spreadsheet support).
- Based on OpenLayer, OGC-compliant using Phyton. Maps are generated internally by a request to a server located in Liege. The request itself is prepared by a Web interface. Results are shown in OpenLayers.
- Requests could come from other servers. (ICES and VLIZ expressed interest).
- Additional outputs as netCDF files, Matlab (or Octave) files and KML files (Google Earth).

For the moment only restrictions on CPU time and data quantity.

Comparison

Method	$\min(\epsilon^2)$	3D	Multivar	Ops/image	$\epsilon(\boldsymbol{r})$	a priori	C.V.	anisotropy
Cressman		$\star$	$\star$	$N_d N_a$		$w(r/L)$	(L)	$(\star)$
O.I.	$\star$	$\star$	$\star$	$N_d^3 + N_d N_a$	$\star$	$c(r/L)$	$L, \sigma^2/\mu^2$	$(\star)$
DIVA	$\star$	$(\star)$	$(\star)$	$N_e^{5/2}$	$\star$	$K(r/L)$	$L, \sigma^2/\mu^2$	$\star$
DINEOF	$(\star)$	$\star$	$\star$	$N_a^{5/4}$	$(\star)$	stat.	N	$\star$

N_d : number of data points

N_a : number of grid points for analysis

N_e : number of finite elements

N : number of EOFs

L : correlation length

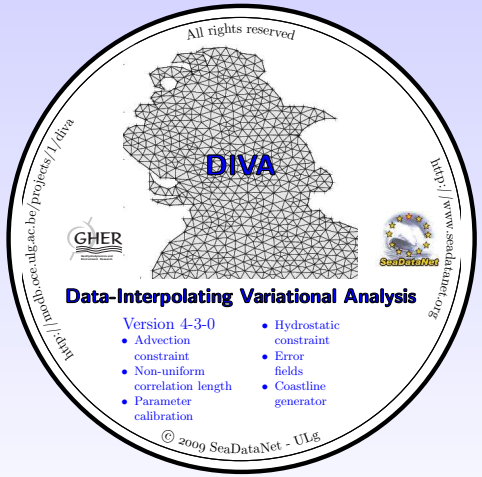
σ^2/ϵ^2 : signal to noise ratio

$\star$: available feature

$(\star)$: available with some adaptations

Reference tool for SeaDataNet:

- Large data sets expected
- Error fields requested
- GUI, ODV-interface, WEB-interface (beginners) or scripts (expert mode)
- 3D (= stacking of 2D)
- G.C.V. for calibration of L and S/N.
- Land and underwater islands/obstacles taken into account (covariance changed).



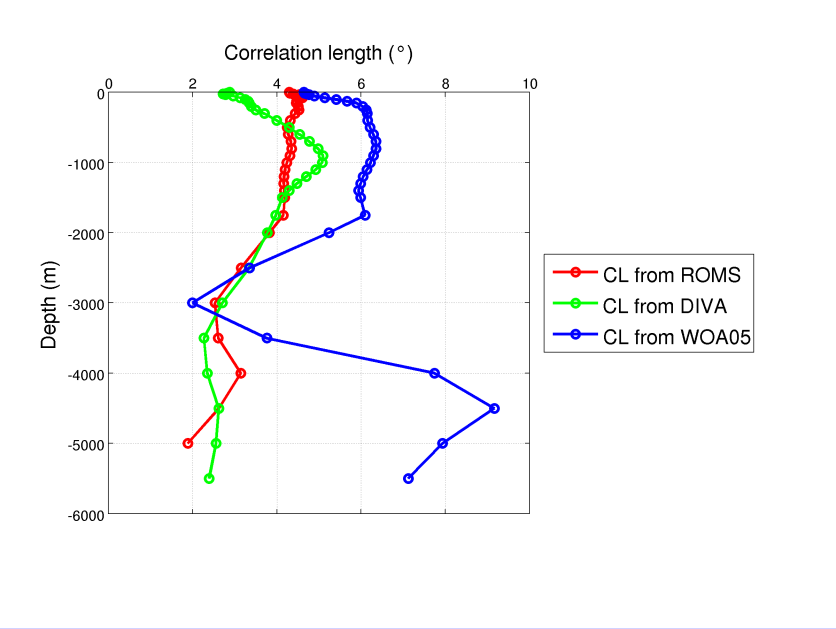
`http://modb.oce.ulg.ac.be/projects/1/diva`

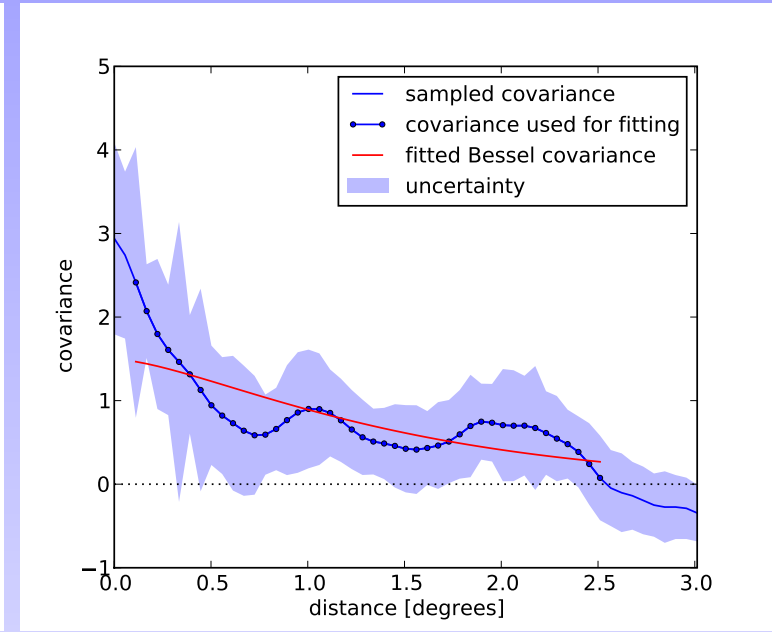
Université de Liège

- *Field estimation theory*
- *DIVA*
- ***Critical points***
- *Examples*
- *Summary*

Université de Liège

Parameter calibration

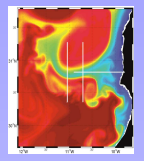




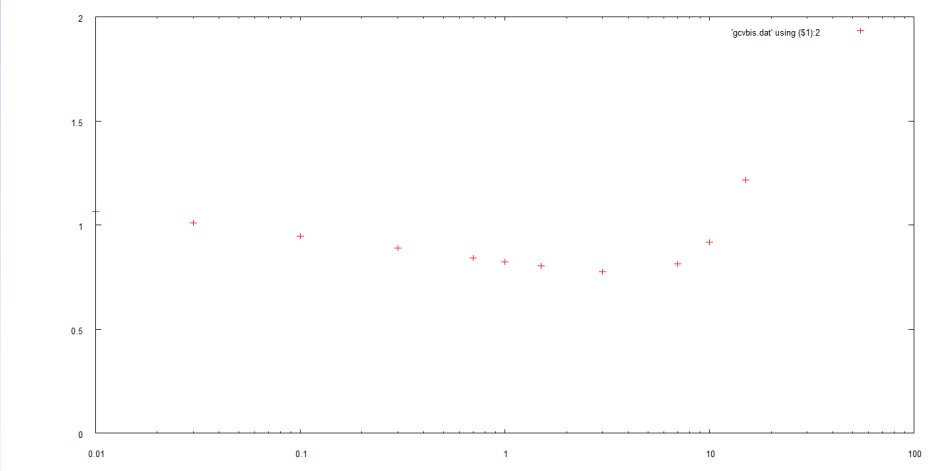
Spatial coherence of parameters: here correlation length obtained with covariance fitting (Troupin *et al.*, 2010).

Signal to noise ratio

The most elusive parameter.

- Noise is not only instrumental error: 
- Very hard problem to decide on value with dependent data (cross-validation approaches fail).

A series of estimation tools are provided with DIVA, but here the experience of oceanographers is critical. A posteriori analysis of residuals allows to verify coherence. With reasonable amount of data, parameter not critical for analysis but for error estimates.

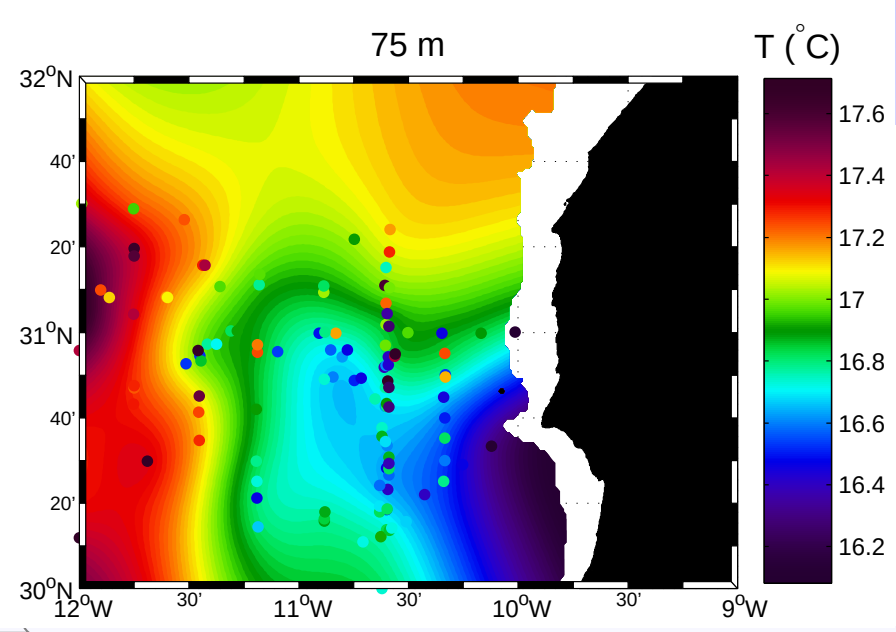


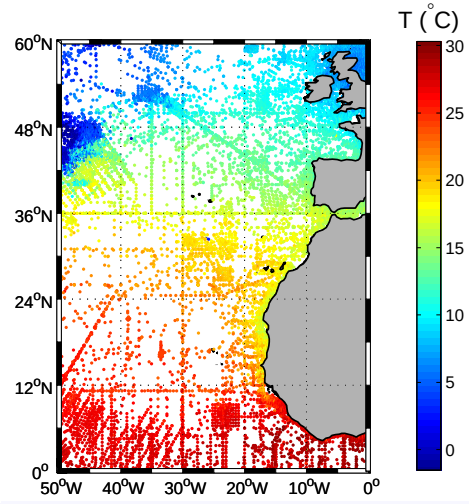
Université de Liège

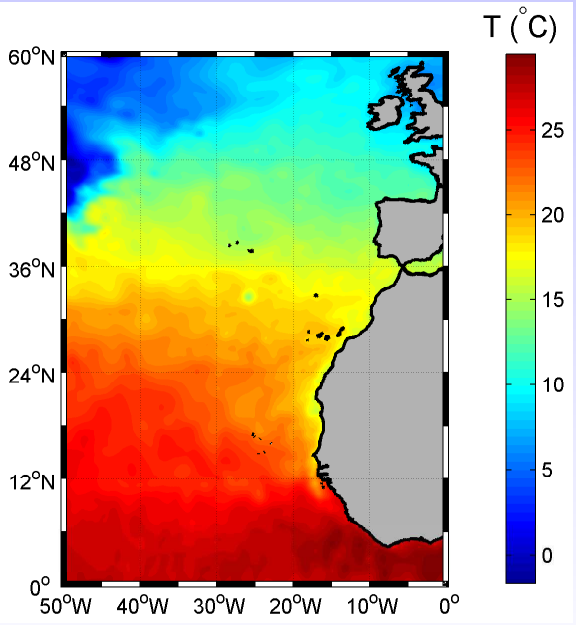
Data availability

THE problem.

- "Enough data" different for cruise snapshots or climatologies
- Recurrent discussion: if and how to mask analyses. Scientific approach (provide field and error estimate) vs large public approach (danger of misinterpretation or misuse).
- Validation and QC of products (independent data, other climatologies).





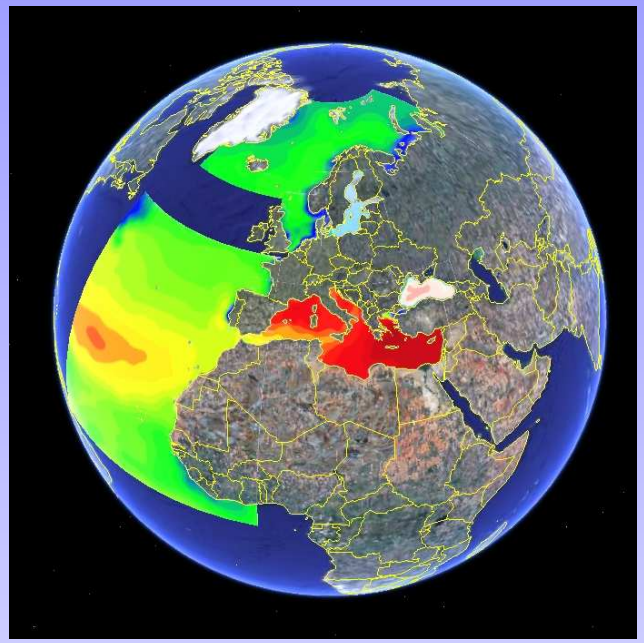


Université de Liège

- *Field estimation theory*
- *DIVA*
- *Critical points*
- ***Examples***
- *Summary*

Université de Liège

Regional products within SeaDataNet



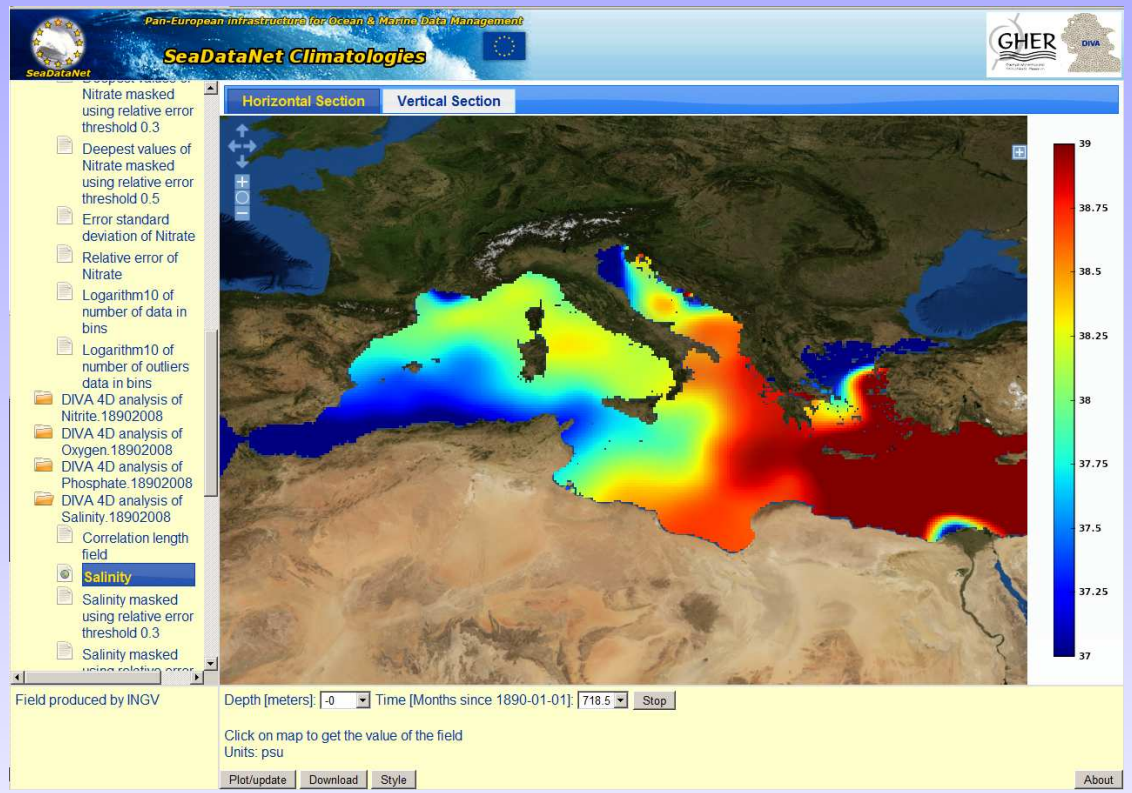
Installation of latest DIVA version and additional climatology-production tools.

Université de Liège

Université de Liège

Climatology on WEB

Regional products uniformly and automatically prepared as 4D netCDF files. Hence possibility for common interface:

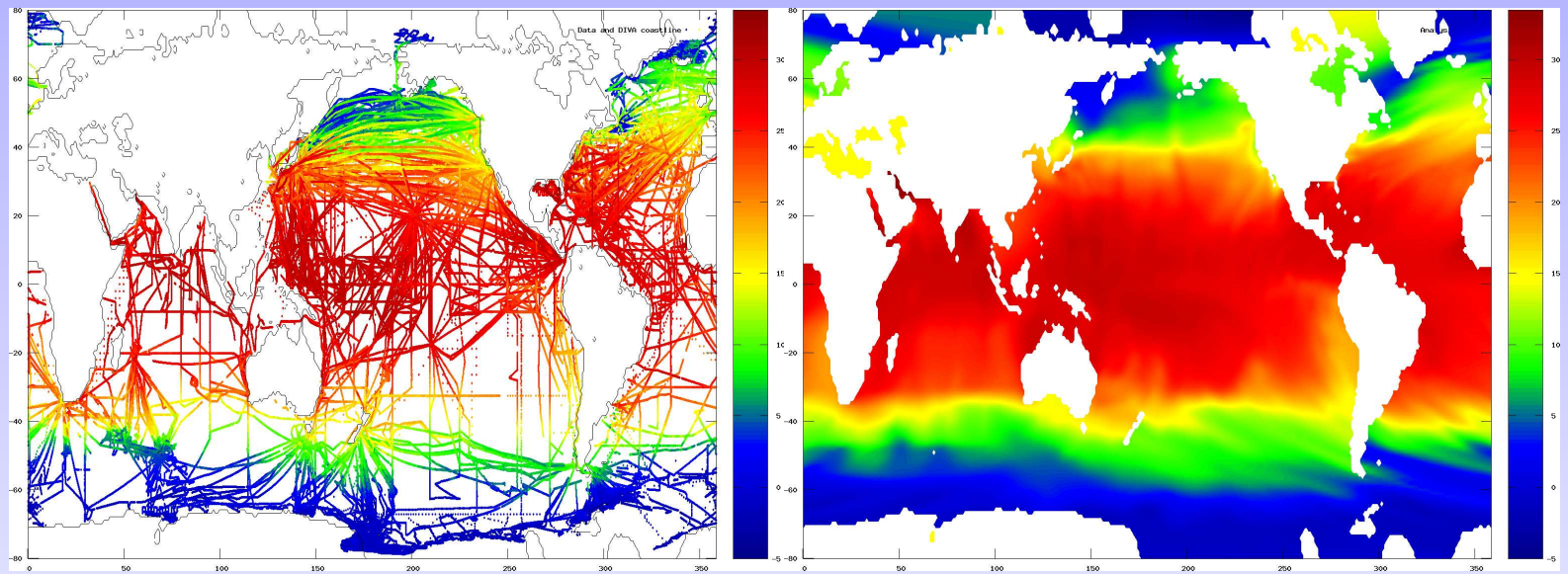


`http://gher-diva.phys.ulg.ac.be`

Includes vertical transect tool.

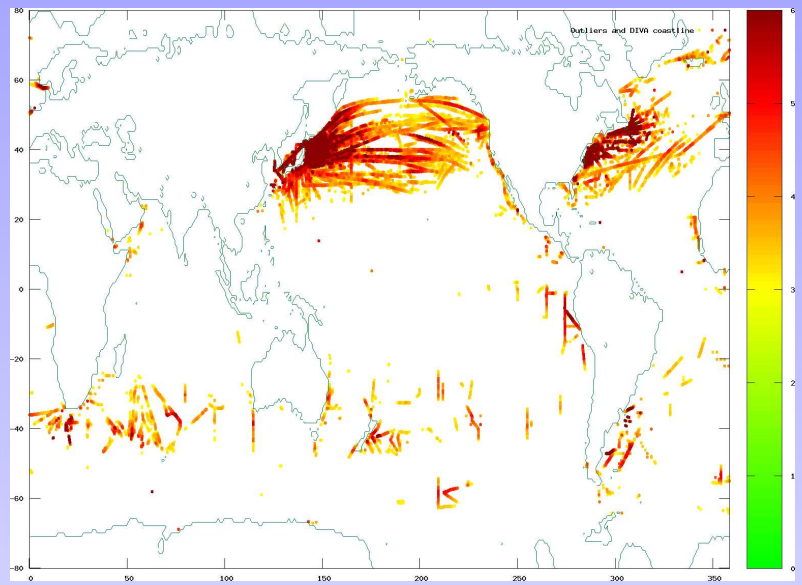
Huge problems

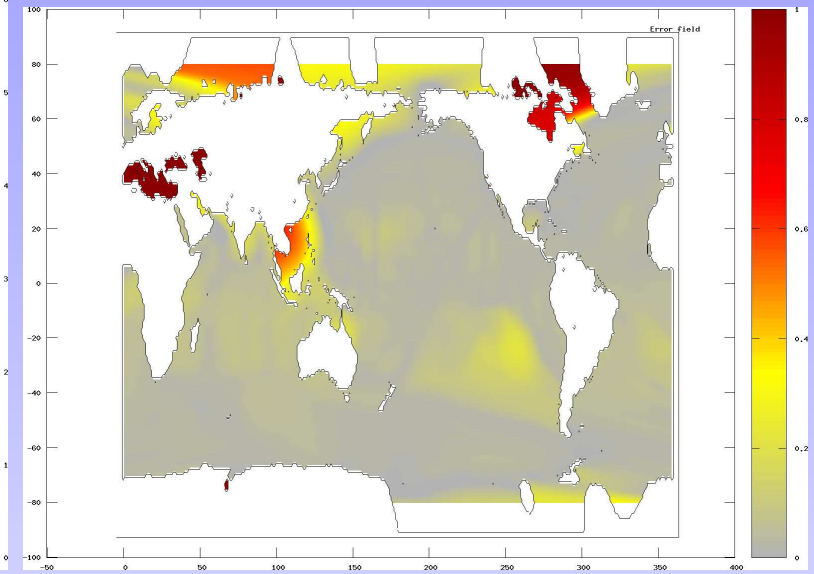
LDEO data base with 4.5 10^6 measurements (Takahashi *et al.*, 2009). Running on a laptop within a few minutes. Shown here, temperature fields.



Université de Liège

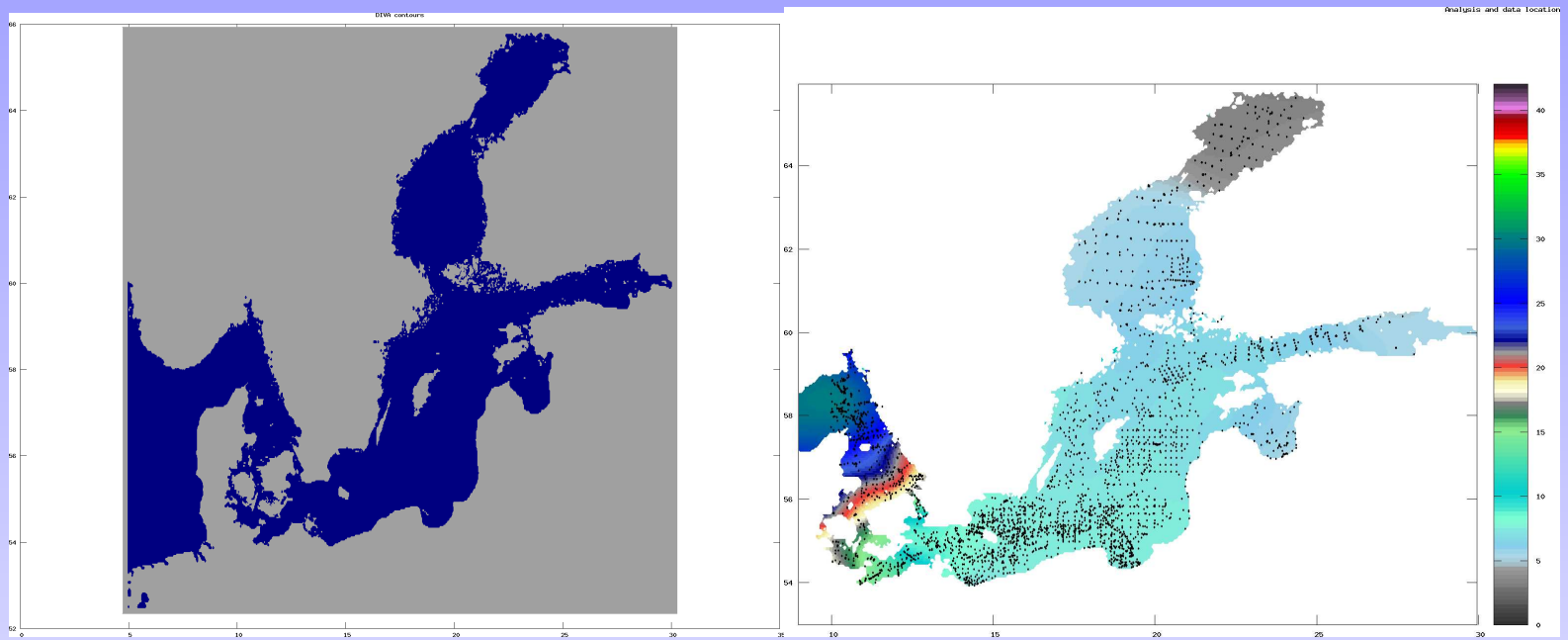
Huge problems, outliers and relative error field





Outliers detected via comparison of statistically expected residuals (value provided by the DIVA analysis) and actual residuals.

Heterogeneous case

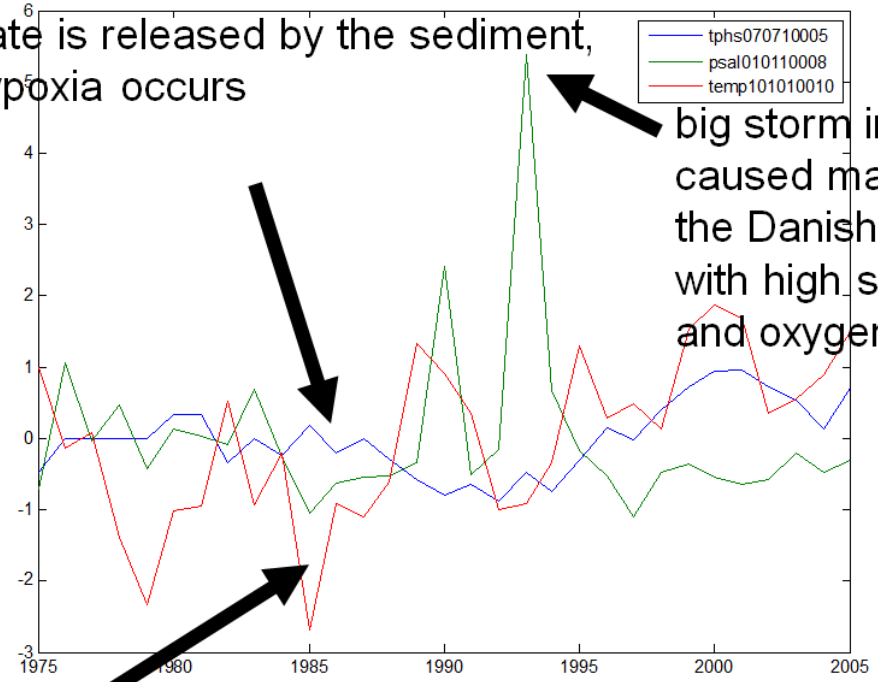


Baltic Salinity Climatology (Bassompierre *et al.*, 2010)

Université de Liège

Heterogeneous case, trends

X= nutrient & climatic trends from Kattegat to Bothnian Bay (SDN products)



phosphate is released by the sediment, when hypoxia occurs

big storm in January 1993, caused massive water through the Danish straits inflows with high salinity and oxygenated the Baltic

increased temperature increases risk of hypoxia

tphs0707005= total phosphate July -100 m
psal 0101108=salinity January level -40 m
temp101010010=temperature October -20 m

(Bassompierre *et al.*, 2010)

Diva-on-web

`http://gher-diva.phys.ulg.ac.be`

Take the example provided on the Web page

- Change S/N ratio
- Change L
- Use divafit
- Look at a posteriori values of S/N
- Change colorbars and ranges for plotting

Diva-on-web

`http://gher-diva.phys.ulg.ac.be`

Take a 2D data set of your research.

- If not in the ocean, provide pseudo-coordinates
- Prepare a file in the correct format
- Change S/N ratio
- Change L
- Use divafit
- Look at a posteriori values of S/N
- Change colorbars and ranges for plotting

- *Field estimation theory*
- *DIVA*
- *Critical points*
- *Examples*
- ***Summary***

DIVA at work

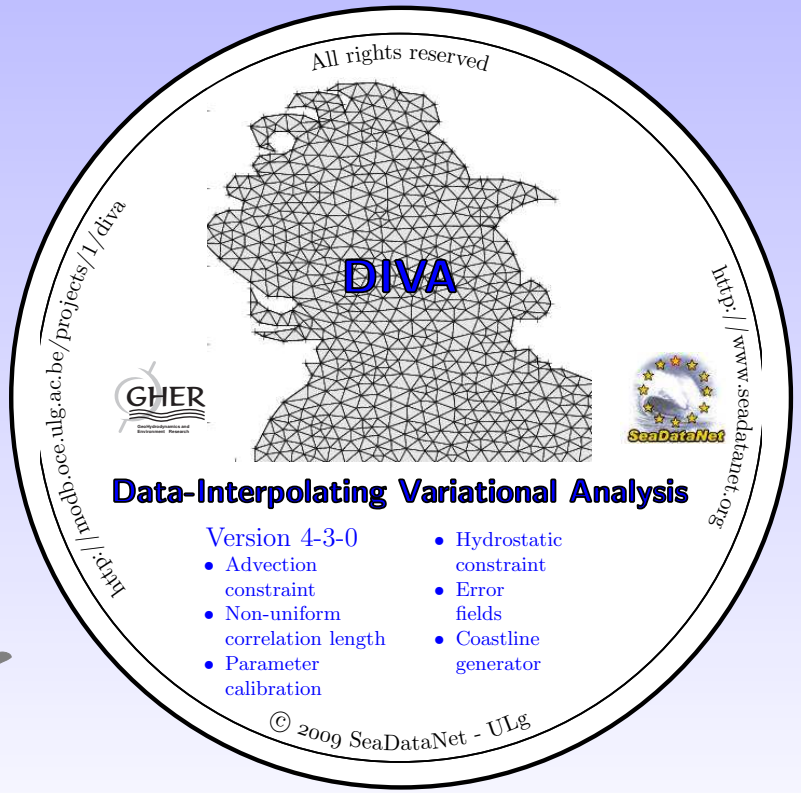
- SeaDataNet regional centers now well working with DIVA4D.
- Group that was most sceptical at Athens SDN annual meeting now the most happy user.
- DIVA development nicely user driven (annual workshop).
- ODV4 spreadsheet format gaining acceptance and easying up operations.
- DIVA incorporation into ODV very stable.
- Product preparation very effective at spotting outliers (visually or by automated a posteriori outlier detection) and testing data flow.

Optimal interpolation

- Kalman Filter as Optimal interpolation
- Optimal ONLY if second order statistics correctly known
- In practize covariance specifications by parametric functions or functionals
- Role of correlation length and signal/noise ratio
- Rarely error dependence is take into account (non diagonal **R**)

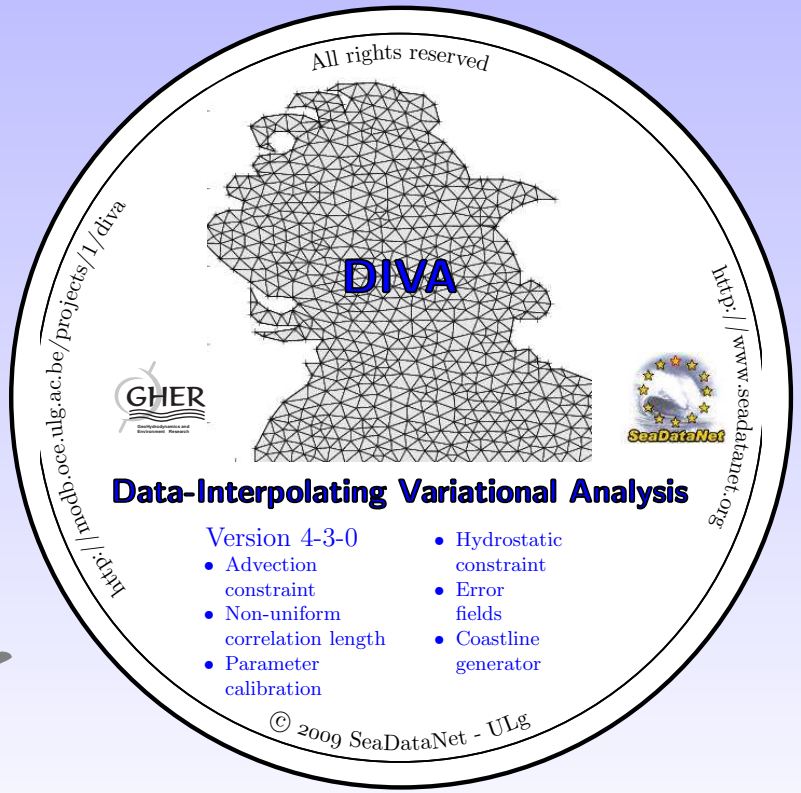
Acknowledgements

Special thanks to Helge Sagen, Marina Tonani, Sissy Iona, Cristina Tronconi, Jacob Carstensen, Marc Bassompierre, Serge Scory, Øivind Østensen, Pierre Brasseur, Jean-Michel Brankart and Michel Rixen for comments on DIVA evolution. Reiner Schlitzer designed and implemented the ODV coupling. SeaDataNet financed the continued development. The support from the Fonds pour la Formation la Recherche dans l'Industrie et dans l'Agriculture (FRIA) and National Fund for Scientific Research is also acknowledged.



Acknowledgements

Special thanks to Helge Sagen, Marina Tonani, Sissy Iona, Cristina Tronconi, Jacob Carstensen, Marc Bassompierre, Serge Scory, Øivind Østensen, Pierre Brasseur, Jean-Michel Brankart and Michel Rixen for comments on DIVA evolution. Reiner Schlitzer designed and implemented the ODV coupling. SeaDataNet financed the continued development. The support from the Fonds pour la Formation la Recherche dans l'Industrie et dans l'Agriculture (FRIA) and National Fund for Scientific Research is also acknowledged.



P. Brasseur, J.-M. Beckers, J.-M. Brankart, and R. Schoenauen. Seasonal temperature and salinity fields in the Mediterranean Sea: Climatological analyses of a historical data set. *Deep Sea Research*, 43:159–192, 1996.

Ch. Troupin, F. Machin, M. Ouberdous, D. Sirjacobs, A. Barth, J.-M. Beckers. High-resolution Climatology of the North-East Atlantic using Data-Interpolating Variational Analysis (Diva). *Journal of Geophysical Research*, accepted, 2010.

J.-M. Beckers and M. Rixen. EOF calculations from incomplete oceanographic data sets. *Journal of Atmospheric and Ocean Technologies*, 20:1839–1856, 2003.

M. Rixen, J.-M. Beckers, J.-M. Brankart, and P. Brasseur. A numerically efficient data analysis method with error map generation. *Ocean Modelling*, 2:45–60, 2000.

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

A. Alvera-Azcárate, A. Barth, M. Rixen, and J.-M. Beckers. Reconstruction of incomplete oceanographic data sets using Empirical Orthogonal Functions. Application to the Adriatic sea. *Ocean Modelling*, 9:325–346, 2005.

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

Diva command line

`http://modb.oce.ulg.ac.be/projects/1/diva`

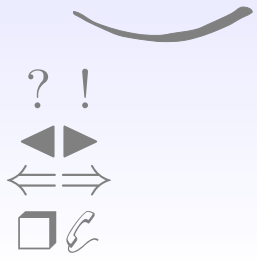
Take one of the examples of the distribution using `divaload`

`../examples/anexample`

- Use `divadress` for full analysis
- Change S/N ratio (by editing `./input/param.par`)
- Change L
- For plotting either load results (`./output/fieldgher.anl`) into matlab or the netCDF version into your preferred plotting system

Diva command line

Try other examples and with the help of the documentation, try to use different features. (advection, detrending)



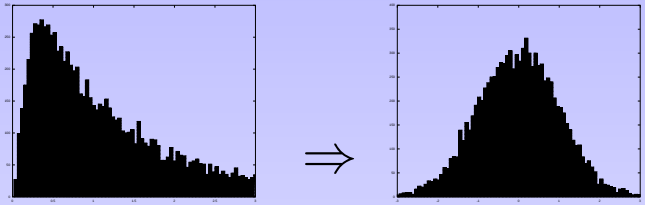
Computational aspects

- Migration to F95 with dynamic memory allocation.
- Open source tools for large matrix calculations and parallel machines. (But best parallelization remains on parameters, seasons ...)
- Feeding of very large data arrays (10^7).
- Web visualization (both single DIVA analysis and 4D netCDF files).

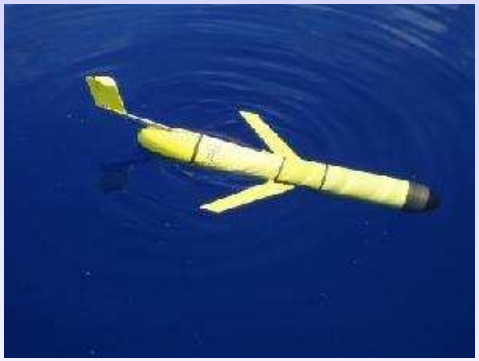
In all cases, backward compatibility, portability and independence on proprietary software

Functional aspects

- Common problem in SeaDataNet and other projects: lack of biochemical data, strange statistical distributions.
  - ⋆ Multivariate approach with non-collocated data.
  - ⋆ Data transformation tools including anamorphosis.
  - ⋆ Absence-presence data for probability analysis.



- Adapting input possibilities.



- Investigate solutions to calibration with dependent data.

Very profound changes?

- N-Dimensional generalization. Easy in finite differences, awful in FEM, with need to rewrite completely the code.
- More complicated physical laws (source terms, vertical advection...).
- Diva-on-web to be used from within CDI interface for analyses on the fly ? (Including possibility to process restricted data without actually delivering them?).
- Rethinking of folder structure for better support of multiple users in a single installation.

Some kind of conservatism here: robustness is priority (10^2 to 10^4 spatial analyses for a 4D product). Funding ?