SANGOMA: Stochastic Assimilation for the Next Generation Ocean Model Applications EU FP7 SPACE-2011-1 project 283580

Deliverable 5.2:
Ecosystem data report
Due date: 01/11/2014
Delivery date: 4/11/2015
Delivery type: Report, public

Peter Jan Van Leeuwen Sanita Vetra-Carvalho
University of Reading, UK

Pierre Brasseur Jean-Michel Brankart Adeline Bichet
CNRS-LEGI, FRANCE

Lars Nerger Paul Kirchgessner
Alfred-Wegener-Institute, GERMANY

Arnold Heemink Nils van Velzen
Martin Verlaan M. Umer Altaf
Delft University of Technology, NETHERLANDS

Jean-Marie Beckers Alexander Barth
University of Liège, BELGIUM

Pierre De Mey
CNRS-LEGOS, FRANCE

Laurent Bertino Alberto Carrassi
NERSC, NORWAY

Contents

Introduction

An important part of Sangoma project is to investigate appropriate methods and algorithms for assessing observation systems. A simple and powerful method for assessing the impact of observing systems is based on the computations of degrees of freedom for signal (DFS) or trace method [Rodgers, 2000, Rabier et al., 2002], which will be implemented here and used to evaluate the assimilation impact of satellite date of new ecosystem data. However, since some new data assimilation methods can produce non-Gaussian priors (such as particle filter based methods) and most new observational data is non-Gaussian (e.g. from satellite, radar etc) we need to use more advanced methods to asses the potential impact this data can have in our assimilation system. These include relative entropy between prior and posterior distributions, mutual information and DFS computed with the anamorphosed variables.

Hence, this report is divided into two parts, where in first part (Chapter 2) we present a summary of measures of observation impact including methods that can be applied to non-Gaussian distributions (prior and/or posterior) paying attention to distinguish between linear and non-linear cases. We have included four methods: *sensitivity*, *trace method* or *DFS*, *mutual information*, and *relative entropy*. In the second part (Chapter 3) we present results from experiments done within Sangoma to compare the impact of assimilating ocean colour data from three satellites (Sea-viewing Wide Field-of-view Scanner instrument (SeaWIFS) aboard the OrbView-2 (a.k.a. SeaStar) satellite, the Moderate Resolution Imaging Spectroradiometer instrument aboard the Aqua satellite (AquaMODIS), and the Medium Resolution Imaging Spectrometer (MERIS)) given an assimilation system that uses the Ensemble Kalman filter [Evensen, 2003]. We use an ensemble of simulations that covers the North Atlantic and is designed to assess the effects of the stochastic parametrisation on the chlorophyll representation [Garnier et al., IN PRESS]. Here, we assess the pertinence of each of these three satellite products by quantifying their individual potential to reduce forecast error. Such an evaluation should allow a better understanding of the way different observational networks affect the assimilation, thereby leading to a better optimisation in the future design of the observational networks that are to be assimilated. For this we use the trace method (or DFS) applied to anamorphosed variables, a novel approach that is compatible with the Sangoma toolbox (WP2) and that is used in the Deliverable 5.6. Using the trace method allows us to exploit existing, costly to perform, ensembles of simulations, and evaluate the performance of different observational array at detecting prior errors without performing data assimilation, which keeps computational cost relatively low. Dealing with the effectiveness of the correction in observation-space, the trace

method quantifies the impact that an observational array would have, if it was to be assimilated in the forecast ensemble. While in Deliverable 5.6 we use the trace method to quantify the potential impact of assimilating physical ocean data (i.e. sea surface height, salinity and temperature profiles) from different satellite products and in situ measurements, this report uses the trace method to quantify the potential impact of assimilating ocean colour data (from which chlorophyll concentration is extracted) from different satellite products. Note that the trace method method will be further developed in the EC-funded Atlantos project.

Measures of observation impact

This chapter provides a description of:

- Sensitivity to observations;
- Degrees of Freedom of Signal (DFS);
- Relative Entropy RE;
- Mutual Information MI;

methods and will discuss some examples on how they are implemented and what they tell us about the impact of observations. We note that *sensitivity to observations*, *Relative Entropy* and *Mutual Information* are included in Sangoma diagnostic tool package and can be downloaded from Sangoma repository:

http://sourceforge.net/p/sangoma/code/HEAD/tree/tools/trunk/Matlab/diagnostics/.

2.1 Sensitivity to observations

2.1.1 Description

Gaussian case

In Gaussian data assimilation where the analysis, $\mathbf{x}^a$, is a linear function of the observations and prior estimate,

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \tag{2.1}$$

where $\mathbf{K}$ is the Kalman gain, a function of $\mathbf{P}$ and $\mathbf{R}$, $\mathbf{H}$ is the linear observation operator, $\mathbf{y}$ are observations at time of analysis.

The sensitivity of the analysis to the observations has an obvious interpretation in terms of observation impact [Cardinali et al., 2004]. It is defined as

$$\mathbf{S} = \frac{\partial \mathbf{H}\mathbf{x}^a}{\partial \mathbf{y}}. \tag{2.2}$$

This is a $m \times m$ matrix where m is the size of the observation space.

Differentiating eq. 2.1 with respect to $\mathbf{y}$ we see that

$$\mathbf{S}^G = \mathbf{H}\mathbf{K}, \tag{2.3}$$

where superscript $(.)^G$ refers to Gaussian assumption. The Kalman gain can be written in many different forms including $\mathbf{K} = \mathbf{P}^{a,G}\mathbf{H}^T\mathbf{R}^{-1}$ where $\mathbf{P}^{a,G}$ is the analysis error covariance matrix. Therefore, the sensitivity is inversely proportional to $\mathbf{R}$ and proportional to $\mathbf{P}^{a,G}$. Hence, the analysis has greatest sensitivity to independent observations with the smallest error variance which provide information about the region of state spaces with the largest prior error [Fowler and van Leeuwen, 2012]. In the range between 0 and 1, $\mathbf{K}$ increases with the amount of observation used to alter the background vector, that is to say, when the covariance matrix of observational error $\mathbf{R}$ is relatively smaller than the covariance matrix of the forecast error $\mathbf{P}^f$.

Non-Gaussian case

When prior is not Gaussian, the simple linear relationship between the mean of the posterior and the observations in eq. 2.1 breaks down. Also, in most non-Gaussian cases the posterior mode will not correspond to the mean of the posterior as the posterior could be bi- or multi-modal. This makes the mode more difficult to uniquely define and the mode will have infinite sensitivity to observations when the mode transfers from one peak to another [Fowler and van Leeuwen, 2012].

As an example assume that the prior is a Gaussian mixture with two modes, Fowler and van Leeuwen [2012] shows that in that case the sensitivity of the analysis may be computed as

$$\mathbf{S} = \frac{1}{k+1} + \frac{kw(1-w)(\mu_1-\mu_2)^2 \exp^{-a_1-a_2}}{(1+k)^2\sigma^2\left[w\exp^{-a_1}+(1-w)\exp^{-a_2}\right]^2}, \tag{2.4}$$

where

1. w, the prior weight given to the first Gaussian, leaving the weight given to the second Gaussian as $1-w$;
2. μ_1, the mean of the first Gaussian;
3. μ_2, the mean of the second Gaussian;
4. σ^2, the variance of both Gaussian components;
5. $a_i = \left[(\mu_y-\mu_i)^2/2(1+k)\sigma^2\right]$;
6. $k = \frac{\sigma^2_{obs}}{\sigma^2}$, a scalar.

Thus in non-Gaussian case $\mathbf{S}$ is a function of the observation value due to the exponent a_i. For more detailed discussion and interpretation see Fowler and van Leeuwen [2012].

From figure 2.1 we can see that when full prior is used to assimilate the observation the analysis may be both more or less sensitive to the observation than when the prior is approximated by a Gaussian. The degree to which the sensitivity is affected depends on the value of the observation.

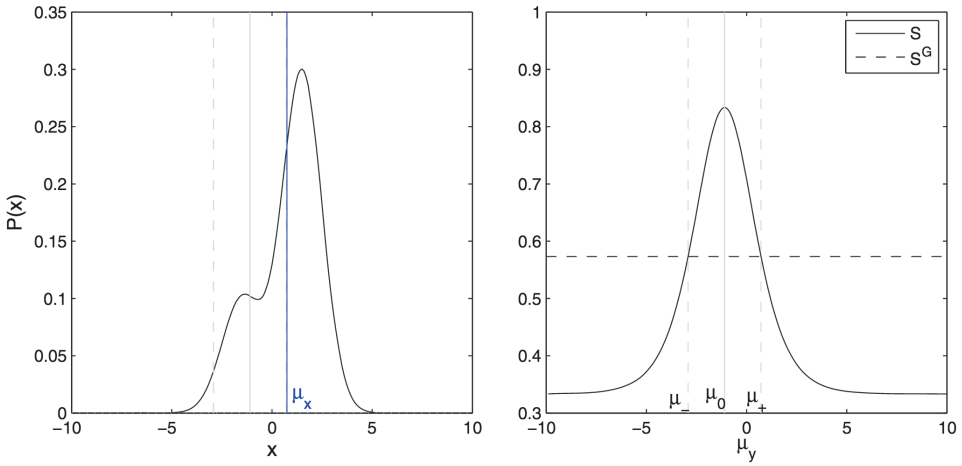

Figure 2.1: Left: the prior distribution. The vertical blue line shows the prior mean, μ_x. Right: **S** is the solid line and the Gaussian approximation is given by the dashed line. For more detail see Fowler and van Leeuwen [2012].

2.1.2 Available code as part of Sangoma tools

Sensitivity function is implemented in Sangoma tools and you can find the code in Matlab and Fortran on Sangoma Sourceforge site. Deliverable DL2.5 contains descriptions of all tools implemented in Sangoma project including *Sensitivity*.

2.2 Trace method or Degrees of Freedom for Signal (DFS)

The trace of **S** in eqn. 2.3 gives the degrees of freedom for signal and we will use this method in Chapter 3 to evaluate the impact of new ecosystem data. Within this report we refer to this method as the trace method.

2.2.1 Description

The trace method consists of quantifying the sensitivity of an ensemble to observations. In other words, how a given assimilation system uses the observations to "pull" the forecast signal from the background. This sensitivity is represented by the $\mathbf{S} = \mathbf{HK}$ matrix (see eqn. 2.3 in section above), which compares the forecast error covariance matrix $\mathbf{P}^f$ (given by the stochastic ensemble) with the observational error covariance matrix **R**.

The gain in information brought by the observations is quantified via the computation of $tr(\mathbf{HK})$, which is the sum of the singular values of the **HK** matrix. Hence, $d_s = tr(\mathbf{HK})$ describes the number of useful, independent quantities in the observations (= degrees of freedom for signal) that are used to reduce the uncertainty of $\mathbf{x}^f$, by quantifying how many degrees of freedom the observations are able to detect in $\mathbf{P}^f$. In other words, it evaluates which observational network detects most degree of freedom in the background vector $\mathbf{x}^f$. d_s can vary between 0 (observations have no influence on the analysis, because $\mathbf{R} \gg \mathbf{P}^f$), and

m, the number of independent observations ($\mathbf{R} \ll \mathbf{P}^f$, so $\mathbf{HK} = \mathbf{I}$).

2.3 Relative Entropy (RE)

2.3.1 Description

Relative entropy measures the gain in information of the posterior relative to the prior and is given by

$$\text{RE} = \int p(\mathbf{x}|\mathbf{y}) \ln \frac{p(x|\mathbf{y})}{p(\mathbf{x})} \mathrm{d}\mathbf{x}. \tag{2.5}$$

Relative entropy can be taught of as a measure of the 'distance' between $p(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{x})$. However, it is not true distance because it is not symmetric [Cover and Thomas, 1991].

Gaussian case

When both the prior and posterior are Gaussian, relative entropy is given by [Bishop, 2006]

$$\text{RE} = \frac{1}{2}(\mathbf{x}^a - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x}^a - \mathbf{x}^b) + \frac{1}{2} \ln \left[\mathbf{B}\mathbf{P}_a^{-1}\right] + \frac{1}{2} tr \left[\mathbf{B}^{-1}\mathbf{P}_a\right] - \frac{1}{2} n. \tag{2.6}$$

The first term in known as the signal term, which measures the change in the mean if the distribution. The rest is known as the dispersion term, which measures the change in the covariance and can be written in terms of the eigenvalues of the sensitivity matrix whilst the signal term depends on the value of observations and the prior mean.

The dependance of relative entropy on both the mean and variance of the posterior makes it an attractive measure, as it gives a more complete description of the observation impact. It can also be shown that relative entropy is invariant under a general non-linear change of variables [Kleeman, 2011].

2.3.2 Available code as part of Sangoma tools

Relative entropy function is implemented in Sangoma tools and you can find the code in Matlab and Fortran on Sangoma Sourceforge site. Further, the deliverable DL2.5 contains descriptions of all tools implemented in Sangoma project including *Relative entropy*.

2.4 Mutual Information (MI)

2.4.1 Description

Mutual information measures the reduction in entropy when observation is made, that is the difference between entropy in the prior and the posterior. In the information theory, entropy is a measure of the uncertainty associated with a random variable. For a probability distribution $p(x)$, entropy can be defined as

$$\int p(x) \ln p(x) \mathrm{d}x.$$

The entropy of posterior pdf, $p(x|z)$ is defined as

$$\int\int p(x,y)\ln p(x|y)\mathrm{d}x\mathrm{d}y.$$

Mutual information is given by the prior entropy minus the conditional entropy

$$\mathrm{MI} = -\int p(x)\ln]\,[p(x)]\,\mathrm{d}x + \int\int p(x,y)\ln\left[p(x|y)\right]\mathrm{d}x\mathrm{d}y, \tag{2.7}$$

where despite the dependence of the posterior error variance on the observations, the conditional entropy is independent of the value of the observations. Therefore, mutual information, unlike the sensitive of the posterior mean to the observations, is independent of the value of the observations, see Fowler and van Leeuwen [2012] for more detailed description with figures.

Gaussian case

When $p(x)$ is Gaussian, the entropy associated with x depends only on its covariance matrix $\mathbf{C}_x$. The entropy in this case is given by $(1/2)\ln\left[(2\pi e)^n|\mathbf{C}_x|\right]$, where n is the size of the vector x and $|*|$ denotes the determinant [Rodgers, 2000]. Mutual information, therefore, for a Gaussian prior and posterior is given by

$$\mathrm{MI} = \frac{1}{2}\ln\left[\mathbf{B}\mathbf{P}_a^{-1}\right], \tag{2.8}$$

and therefore, it is a measure of the difference in the determinant of the prior and posterior covariance matrices.

Mutual information can be written in terms of the eigenvalues of the sensitivity matrix, $\mathbf{S}$, as follows

$$\mathrm{MI} = -\frac{1}{2}\sum_{i=1}^{r}\ln|1-\lambda_i|, \tag{2.9}$$

whereλ_i is the ith eigenvalue of $\mathbf{S}$ (ordered in descending magnitude) and $r \leq \min(n,p)$ is the rank of $\mathbf{S}$. It is a scalar interpretation of the observation impact, and therefore the impact of individual observations may not be easily quantified. However, mutual information can be shown to be additive with successive observations, see Fowler and van Leeuwen [2012].

In case of non-Gaussian prior and posterior the expected value of relative entropy can be shown to be equal to mutual information. This can be shown by writing mutual information in its equivalent form

$$\mathrm{MI} = \int\int p(\mathbf{x},\mathbf{y})\ln\left[\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right], \tag{2.10}$$

where now mutual information is interpreted as how 'close' two variables are to being independent, that is the error approximating $p(\mathbf{x},\mathbf{y})$ by $p(\mathbf{x})p(\mathbf{y})$ [Cover and Thomas, 1991]. In this form MI can be seen to be

$$\int p(\mathbf{y})\mathrm{RE}\,\mathrm{d}y.$$

Figure 2.2 shows the behaviour of the different measures for the simple 2-component Gaussian mixture prior pdf, as function of the value of the observation. For each measure it compares the full nonlinear expression of the measure with its Gaussian approximation, as a ratio of the two. So a value of 1 means that the full expression and the Gaussian approximation are the same, and computing the simpler Gaussian approximation will do fine. However, for example the sensitivity shows values ranging from 0.58 to 1.5 showing that the Gaussian approximation can lead to a factor 2 over- or underestimation of the impact of that observation.

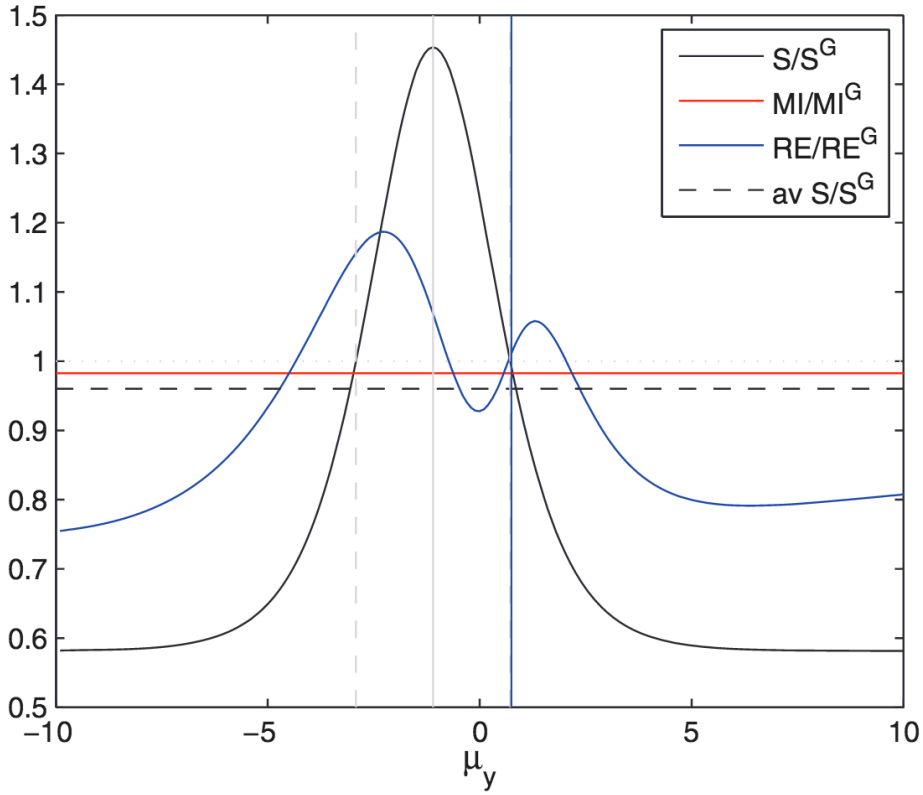

Figure 2.2: **S** (black), MI (red), RE (blue) all normalised by their Gaussian approximations. The black dashed line shows $\int p(y)Sdy$ normalised by its Gaussian approximations. For more detail see Fowler and van Leeuwen [2012].

2.4.2 Available code as part of Sangoma tools

Mutual information function is implemented in Sangoma tools and you can find the code in Matlab and Fortran on Sangoma Sourceforge site. Further, the deliverable DL2.5 contains descriptions of all tools implemented in Sangoma project including *mutual information*.

Impact of new ecosystem data

This Tasks aims to evaluate the impact of assimilating new ecosystem data. As an alternative to Observing System Simulations Experiments (OSSE), we propose to use the $tr(\mathbf{HK})$ method, a novel approach that is compatible with the Sangoma toolbox (WP2) and that is used in the Sangoma Task 5.6. Using the $tr(\mathbf{HK})$ method allows to exploit existing, costly to perform ensembles of simulations, and evaluate the performance of different observational array at detecting prior errors without performing data assimilation, which keeps computational cost relatively low. Dealing with the effectiveness of the correction in observation-space, $tr(\mathbf{HK})$ quantifies the impact that an observational array would have, if it was to be assimilated in the forecast ensemble. While Task 5.6 uses $tr(\mathbf{HK})$ to quantify the potential impact of assimilating physical ocean data (i.e. sea surface height, salinity and temperature profiles) from different satellite products and in situ measurements, this Tasks uses $tr(\mathbf{HK})$ to quantify the potential impact of assimilating ocean color data (from which chlorophyll concentration is extracted) from different satellite products. Note that the $tr(\mathbf{HK})$ method will be further developed in the EC-funded Atlantos project.

Here, we compare the impact of assimilating ocean colour data from three satellites (Sea-viewing Wide Field-of-view Scanner instrument (SeaWIFS) aboard the OrbView-2 (a.k.a. SeaStar) satellite, the Moderate Resolution Imaging Spectroradiometer instrument aboard the Aqua satellite (AquaMODIS), and the Medium Resolution Imaging Spectrometer (MERIS)) given an assimilation system that uses the Ensemble Kalman filter [Evensen, 2003]. We use an ensemble of simulations that covers the North Atlantic and is designed to assess the effects of the stochastic parametrisation on the chlorophyll representation [Garnier et al., IN PRESS]. Here, we assess the pertinence of each of these three satellite products by quantifying their individual potential to reduce forecast error. Such an evaluation should allow a better understanding of the way different observational networks affect the assimilation, thereby leading to a better optimisation in the future design of the observational networks that are to be assimilated.

Because the $tr(\mathbf{HK})$ method assumes a Gaussian distribution, and the distribution of chlorophyll across the prior ensemble is not Gaussian, we apply an anamorphic transformation [Brankart et al., 2012] to the prior ensemble before performing the analysis. The anamorphic transformation consists in transforming an non-Gaussian distribution into a Gaussian distribution using a change in non-linear variables to remap the non-Gaussian quantiles into Gaussian quantiles [Brankart et al., 2012].

3.1 Implementation the trace method

Computing $tr(\mathbf{HK})$ requires the knowledge of $\mathbf{P}^f$, $\mathbf{R}$, and $\mathbf{H}$. In this sense, the implementation of this method is very efficient with most Sangoma methods. In theory, $\mathbf{P}^f$ has a very large number of entries (= number of state variables). For computational purposes, it is possible to select a reduced number of entries for $\mathbf{P}^f$ (reduced ranked approach). We derive $\mathbf{P}^f$ from the 60-member ensemble described below and in Garnier et al. [IN PRESS]. In this example, $\mathbf{P}^f$ has therefore 60 entries. $\mathbf{H}$ and $\mathbf{R}$ are given by the observations. To compute $tr(\mathbf{HK})$, we decompose $\mathbf{P}^f$ such as:

$$\mathbf{P}^f = \mathbf{S}^f \mathbf{S}^{f^T}, \tag{3.1}$$

$$\text{with } \mathbf{S}^f{}_{(i)} = (\mathbf{x}^f{}_{(i)} - \mathbf{x}^f)(m-1)^{-1/2} \tag{3.2}$$

where i corresponds to the individual ensemble members, and m to the total number of ensemble members. We can then rewrite $\mathbf{K}$ such as:

$$\mathbf{K} = \mathbf{S}^f[\mathbf{I} + \Gamma]^{-1}(\mathbf{H}\mathbf{S}^f)^T\mathbf{R}^{-1} \tag{3.3}$$

$$\text{with } \Gamma = (\mathbf{H}\mathbf{S}^f)^T\mathbf{R}^{-1}\mathbf{H}\mathbf{S}^f = \mathbf{U}\Lambda\mathbf{U}^T. \tag{3.4}$$

Therefore, $\mathbf{HK}$ can be expressed as:

$$\mathbf{HK} = \mathbf{H}\mathbf{S}^f[\mathbf{I} + \Gamma]^{-1}(\mathbf{H}\mathbf{S}^f)^T\mathbf{R}^{-1} \tag{3.5}$$

and $tr(\mathbf{HK})$ as:

$$tr(\mathbf{HK}) = tr[(\mathbf{I} + \Lambda)^{-1}\Lambda] = \Sigma(\lambda_k)(1+\lambda_k)^{-1}, \tag{3.6}$$

with λ_k = singular values of Γ.

3.2 Data

3.2.1 Prior ensemble

We use the ensemble of simulations described in Garnier et al. [IN PRESS]. It is performed with the North Atlantic DRAKKAR configuration of NEMO version 3.4 at the eddy-permitting horizontal resolution of 1/4° (NATL025), and covers the time period from January to December 2005. The ocean circulation model is coupled with the biogeochemical model PISCES-v2, which is suitable to represent the biological behaviour in geographical provinces with various biogeochemical regimes. It includes 24 biogeochemical variables, amongst which we distinguish two different types of phytoplankton, each of them containing its own chlorophyll concentration. The total chlorophyll concentration is therefore the addition of these two species contribution. In order to include the years covered by the SeaWIFS data, this NATL025 configuration is forced by ERA Interim atmospheric forcing fields instead of the Drakkar Forcing Sets. To build the prior ensemble, we use two types of stochastic parametrisation in the equation state

[Brankart, 2013]: One accounts for uncertainties in the biogeochemical parameters (in this case we choose seven parameters), the other for uncertainties in unresolved spatial scales [Garnier et al., IN PRESS]. Then, a 60-member ensemble of perturbed simulations is run for 12 months (January to December 2005), that is designed to assess the effects of the stochastic parametrisation on the chlorophyll representation.

To illustrate the spatial representation of the prior ensemble, Figure 3.1 shows, for May 15, 2005, the simulated surface chlorophyll as (a) the ensemble mean and (b) the relative ensemble standard deviation (absolute standard deviation normalised by the ensemble mean to avoid correlations with the magnitude of chlorophyll concentrations). According to Figure 3.1b, the ensemble dispersion is maximal where small scale effects are predominant (i.e. Gulf Stream region, equatorial current and around the coastal zones), and small where chlorophyll concentrations are high (high latitudes). As expected, a small dispersion is generated in March and October, when the biological activity is less intense (not shown). To further illustrate the distribution of this ensemble, Figure 3.2 shows, for May 15, 2005, the composites of simulated surface chlorophyll concentrations (considering simultaneously the information coming from all ensemble members) representing at each grid point the minimum, the 25% percentile, the median, the 75% percentile, and the maximum of the ensemble. Unlike Figure 3.1, Figure 3.2 shows a significant dispersion generated at high latitudes. In addition, Figure 3.2 shows that the difference between the maximum and the third quartile is higher than the difference between the first quartile and the minimum, which results mainly from the zeros lower bound. Hence, where concentrations are small, the simulation can hardly make it decease, while extreme values always correspond to high chlorophyll concentrations. This means that the shape of the stochastic perturbations does not modify the prior lognormal chlorophyll distribution, and that it is important to consider a probability density that characterises the ensemble better than only the ensemble mean and covariance.

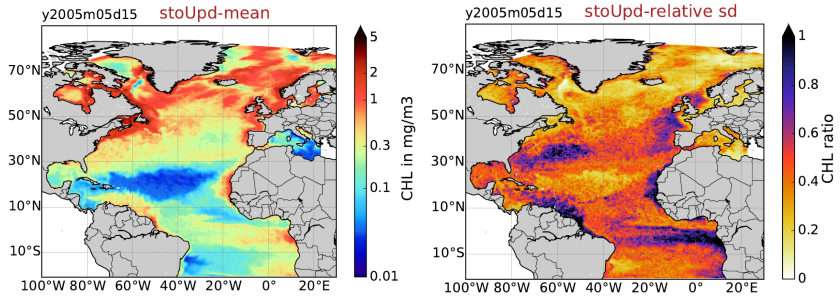

Figure 3.1: Surface chlorophyll as simulated in the prior ensemble, and shown as (a) the ensemble mean and (b) the relative standard deviation, for May 15, 2005.

3.2.2 Observations

We use ocean colour data from the three following satellites data: SeaWIFS, AquaMODIS, and MERIS. Ocean colour data from SeaWIFS is obtained from the NASA's Goddard Space Flight Centre, where near-surface chlorophyll-a concentrations are derived from the OC-4 operational algorithm. We use two products

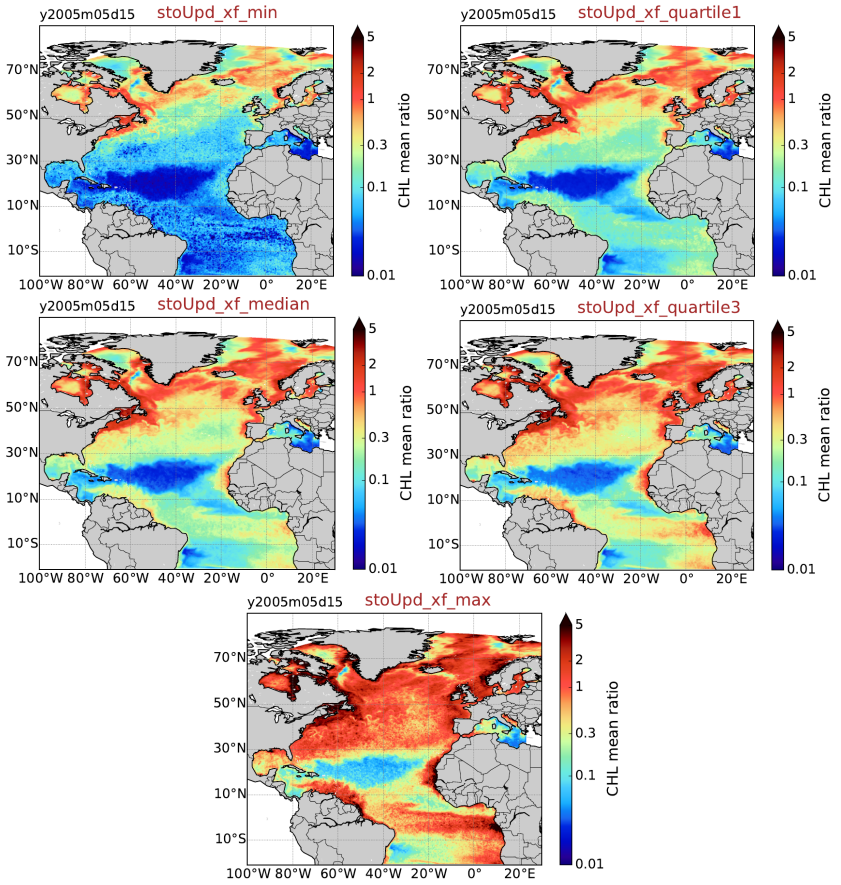

Figure 3.2: Surface chlorophyll as simulated in the prior ensemble, and shown as (a) the minimum, (b) the 25% percentile, (c) the median, (d) the 75% percentile, and (e) the maximum of the ensemble, for May 15, 2005.

from SeaWIFS: The untouched, original low resolution data (SeaWIFS$_{LR}$), and the filled-in, high resolution data (SeaWIFS$_{HR}$) that corresponds to SeaWIFS$_{LR}$ filled-in with data from AquaMODIS and MERIS.

To compute $tr(\mathbf{HK})$, we use an observational error (measurement and representativeness) of 30%, that accounts for a 1° spatial correlation using a diagonal matrix that contains the diagonal entries of **R** inflated by a factor greater than one, which in practice consists in giving less weight to each observation since they have correlated errors [Abdelnur Ruggiero et al., Submitted]. The observational window accounts for the 5 days preceding the date of interest. To illustrate the observational space, Figure 3.3 shows the surface chlorophyll as observed in (a) SeaWIFS$_{HR}$, (b) SeaWIFS$_{LR}$, (c) AquaMODIS, and (d) MERIS. According to Figure 3.3, SeaWIFS$_{HR}$ includes a limited amount of missing data, mostly located around the western tropical Atlantic. On the other hand, SeaWIFS$_{LR}$, AquaMODIS, and MERIS show larger areas of missing data, located mostly over the 0-30° N band north of the tropical Atlantic, as well as over the eastern tropical Pacific for SeaWIFS$_{LR}$.

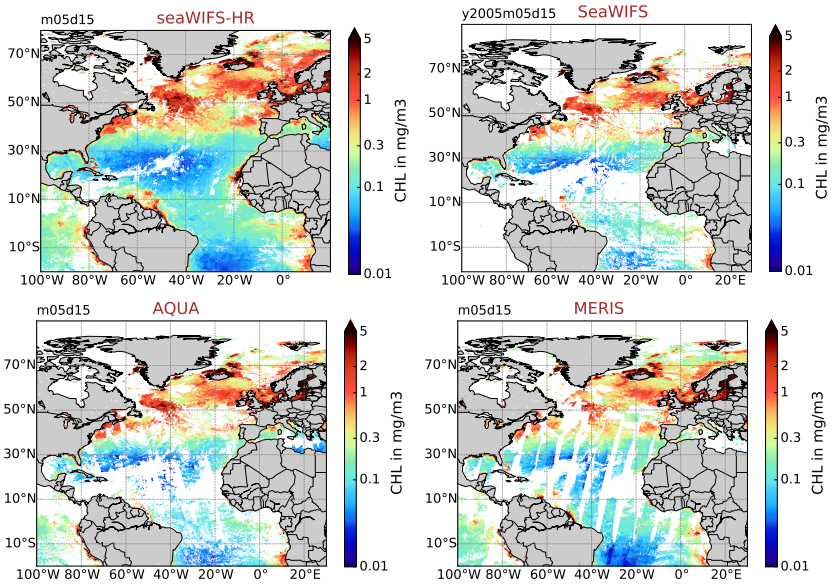

Figure 3.3: Surface chlorophyll as observed in (a) SeaWIFS$_{HR}$, (b) SeaWIFS$_{LR}$, (c) AquaMODIS, and (d) MERIS, using a 5-day observational window from May 10 to May 15 2005.

3.3 Results

Because the Garnier et al. [IN PRESS] ensemble has not been tested for its compatibility with other observations than SeaWIFS$_{HR}$, this Task compares, qualitatively, the sensitivity of $tr(\mathbf{HK})$ to different observational networks rather than providing quantitative conclusions.

Figure 3.4 shows $tr(\mathbf{HK})$ computed from the prior ensemble (May 15, 2005) and ocean colour data taken from (a) SeaWIFS$_{HR}$, (b) SeaWIFS$_{LR}$, (c) AquaMODIS, and (d) MERIS, using a 5-day observational window from May 10 to May 15, 2005. According to Figure 3.4, assimilating any of the four observations brings information around the Gulf Stream the tropical Atlantic areas. This is directly related to the maps of standard deviations seen in Figure 3.1b and Figure 3.2, which show the largest dispersion of the ensemble over these areas. Quantitatively, Figure 3.4 shows that assimilating SeaWIFS$_{HR}$ is the most efficient at reducing the forecast error, in particular over the eastern part of the Gulf Stream area, and assimilating MERIS is the least efficient. In addition, MERIS show regularly spaced areas where no information is brought by the observations. These differences are directly related to the location of the observations, as seen in Figure 3.3, where SeaWIFS$_{HR}$ shows the least amount of missing data over the western part of the Gulf Stream, and MERIS shows regularly spaced areas of missing data all across the domain.

On the other hand, it seems that the assimilation is not affected by the lack of observation seen over the 0-30° N band in SeaWIFS$_{LR}$ and AquaMODIS, and over the eastern tropical Pacific in SeaWIFS$_{LR}$ (Compare Figure 3.3 with Figure 3.4). According to Figure 3.1b and Figure 3.2, this is most likely due to the fact that the forecast error is very low in these areas: In these areas, the forecast

does not require a strong modification and the lack of observation has limited consequences for the assimilation.

We conclude that given the ensemble and assimilation system chosen in our case, assimilating any of the four satellite data would reduce the forecast error over the Gulf Stream and the tropical Atlantic areas, because this is where the forecast error is the largest. The strength of the modification brought to the forecast then depends on the location of the observations, and as expected from Figure 3.3, our results show that assimilating SeaWIFS$_{HR}$ in our case brings the most information to the forecast, as it corresponds to the observation with the least amount of missing data.

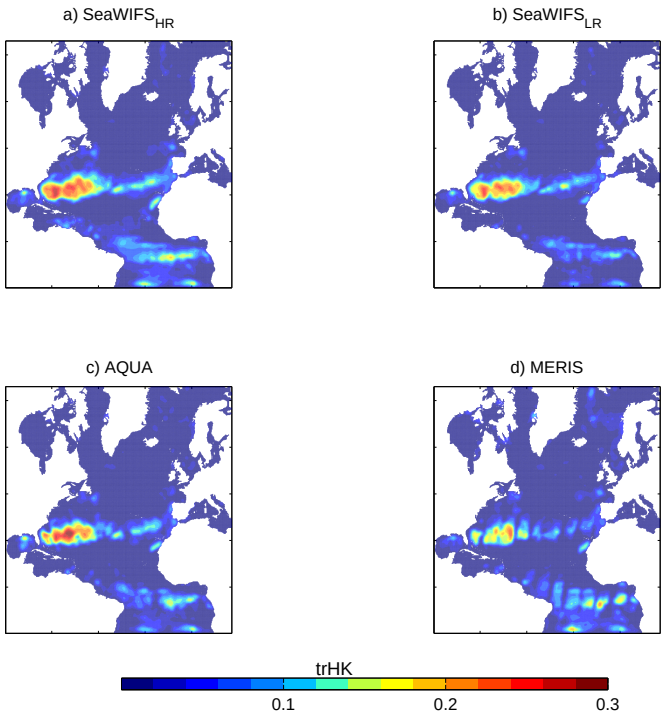

Figure 3.4: $tr(\mathbf{HK})$ computed from the prior ensemble (May 15, 2005) and ocean color data taken from (a) SeaWIFS$_{HR}$, (b) SeaWIFS$_{LR}$, (c) AquaMODIS, and (d) MERIS, using a 5-day observational window from May 10 to May 15, 2005.

Conclusion

In this report we have presented a summary of four methods that can be used to asses the observation impact for linear and non-linear prior and posterior distributions, namely *sensitivity*, *trace method*, *mutual information* and *relative entropy* and presented application of the trace method to non-Gaussian data to evaluate an observation network.

In the methods we have distinguished between linear and non-linear cases where applicable and where method has been implemented in Sangoma toolbox we have included the links to the description of the tools and to the code of the methods.

Through the application of the trace method we show that assimilating ocean colour data from satellites is potentially very efficient at reducing the forecast error. These results show that the quantitative gain of information brought by the satellite products depends primarily on the location and size of the forecast error (the gain increases with the forecast error), in the cases discussed in this report. Further, assimilating ocean colour data from any of the four satellites tested brings information mostly over the Gulf Stream and tropical Atlantic regions, that is to say where the forecast error is the largest. Additionally, the efficiency of the assimilation depends on the location of the observations, and collecting observations in areas where the forecast error is small only has a limited impact on the assimilation: e.g. over the 0-30° N band in our case. Assimilating SeaWIFS$_{HR}$ is the most efficient at reducing the forecast uncertainty because it includes the least amount of missing data. Finally, it is important to keep in mind that unlike assimilating other data such as sea surface height, assimilating ocean colour data is only efficient in Spring and Summer (March to August), when the phytoplankton blooms.

We conclude that the trace method used here to evaluate an observational network has a relatively low computational cost, is simple to implement and interpret, and can be used on data that does not follow a Gaussian distribution (e.g. Chlorophyll). Our results are very encouraging, and this method will be used within the framework of EC-funded project Atlantos to optimise the deployment of an observational network in the Atlantic Ocean. In particular, $tr(\mathbf{HK})$ will be used to evaluate the potential impact of assimilating Chlorophyll and Nitrate concentration measured by the bioARGO floats (that measure the concentration profiles of Chlorophyll and Nitrate across the upper 400 meters of the ocean, at variable frequency).

Bibliography

G. Abdelnur Ruggiero, Cosme, J-M. Brankart, J. Le Sommer, and C. Ubelmann. An efficient way to account for error correlations in the assimilation of observations from the future swot high-resolution altimeter mission. Submitted.

C. Bishop. *Patter recognition and machine learning*. Springer, New York, 2006.

J.-M. Brankart. Impact of uncertainties in the horizontal density gradient upon low resolution global ocean modelling. *Ocean Modelling*, 66:64–76, 2013.

J.-M. Brankart, C.-E. Testut, D. Bèal, M. Doron, C. Fontana, M. Meinvielle, P. Brasseur, and J. Verron. Towards an improved description of ocean uncertainties: effect of local anamorphic transformations on spatial correlations. *Ocean Science*, 8:121–142, 2012.

C. Cardinali, S. Pezzulli, and E. Andersson. Influence-matrix diagnostics of a data assimilation system. *Q. J. R. Meteorol. Soc.*, 130:2767–2786, 2004.

T. M. Cover and J.A. Thomas. *Elements of information theory*. Wiley series in Telecommunications. John Wiley and Sons, New York, 1991.

G. Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003.

A. Fowler and P.J. van Leeuwen. Measures of observation impact in non-gaussian data assimilation. *Tellus*, 64:1–16, 2012.

F. Garnier, J.-M. Brankart, P. Brasseur, and E. Cosme. Stochastic parameterizations of biogeochemical uncertainties in a 1/4° nemo/pisces model for probabilistic comparisons with ocean colour data. *Journal of Marine Systems*, IN PRESS.

R. Kleeman. Information theory and dynamical system predictability. *Entopy*, 13: 612–649, 2011.

F. Rabier, N. Fourrié, D. Chafa, and P. Prunet. Channel selection methods for infrared atmospheric sounding interferometer radiances. *Q. J. R. Meteorol. Soc.*, 128:1011–1027, 2002.

C.D. Rodgers. *Inverse methods for atmospheric sounding*. World Scientific Publishing, SIngapore, 2000.